Meta-Psychology, 2023, vol 7, MP.2021.2756
https://doi.org/10.15626/MP.2021.2756
Article type: Original Article
Published under the CC-BY4.0 license

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Nick Brown
Reviewed by: Maarten Derksen, Sophia Crüwell
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/MZ87J

A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework.

Duygu Uygun Tunç
Eindhoven University of Technology

Mehmet Necip Tunç
Tilburg University

Abstract

Auxiliary hypotheses (*AH*s) are indispensable in hypothesis-testing, because without them specification of testable predictions and consequently falsification is impossible. However, as *AH*s enter the test along with the main hypothesis, non-corroborative findings are ambiguous. Due to this ambiguity, *AH*s may also be employed to deflect falsification by providing "alternative explanations" of findings. This is not fatal to the extent that *AH*s are independently validated and safely relegated to background knowledge. But this is not always possible, especially in the so-called "softer" sciences where often theories are loosely organized, measurements are noisy, and constructs are vague. The Systematic Replications Framework (SRF) provides a methodological solution by disentangling the implications of the findings for the main hypothesis and the *AH*s through pre-planned series of systematically interlinked close and conceptual replications. SRF facilitates testing alternative explanations associated with different *AH*s and thereby increases test severity across a battery of tests. In this way, SRF assesses whether the corroboration of a hypothesis is conditional on particular *AH*s, and thus allows for a more objective evaluation of its empirical support and whether post hoc modifications to the theory are progressive or degenerative in the Lakatosian sense. Finally, SRF has several advantages over randomization-based systematic replication proposals, which generally assume a problematic neo-operationalist approach that prescribes exploration-oriented strategies in confirmatory contexts.

Keywords: Auxiliary Hypotheses, Duhem-Quine Thesis, Empirical Underdetermination, Falsificationism, Adversarial Collaboration

Introduction

Some of the problems that social and behavioral sciences tackle have far-reaching and serious implications in the real world. Among them one could list very diverse questions, such as "Is exposure to media violence related to aggressive behavior and how?", "Do the differences in intelligence test scores represent a true difference in cognitive abilities between various ethnic groups?", "Does willpower draw on a finite supply of resources that can dry up?", "What are the main dimensions through which we form our impressions about other human beings?", "Are emotions distinct entities demonstrating natural-kind-like properties (e.g.

having clear neurological and physiological markers)?" Apart from all being socially very pertinent, substantial numbers of studies investigated each of these questions. However, the similarities do not end here. Curiously enough, even after so much resource has been invested in the empirical investigation of these almost-too-relevant problems, nothing much is accomplished in terms of arriving at clear, definitive answers (see Barrett et al., 2019; Ellemers et al., 2020; Hilgard et al., 2017; Lin et al., 2020; Wicherts et al., 2010). If we take the first in the list as an example, we began the inquiry with three logical possibilities regarding how media violence can influence aggression, namely: 1) it increases aggression, 2) it decreases aggression, 3) it does not affect aggression. After decades of investigation, endless discussions, and what seems to be a yearly updated series of conflicting meta-analyses, one can argue that we are not far from where we started (Hilgard et al., 2017).

This is a depressing state for any scientific discipline to be in, as the aim of science is not to accumulate (contradicting) observations for its own sake, but to explain how the universe works or to make reliable predictions about its future states (Lakatos, 1978). Besides, the scientific enterprise differs from other types of nomothetic inquiry (e.g., mythological, philosophical) in that it puts its postulations to empirical tests in the hope of eventually selecting theories with higher verisimilitude (Popper, 2002a). Research programs or disciplines which fail in these tasks of providing valid explanations and accurate predictions or weeding out the bad seeds would have a hard time maintaining their scientific credibility in the long run. This problematic situation has been going on for a considerably long time in the social and behavioral sciences, which renders an old observation of Meehl still relevant; namely, that theoretical claims often do not die normal deaths at the hands of empirical evidence but are discontinued due to a sheer loss of interest (Meehl, 1978). Observing this state, Lakatos maintained decades ago that most theorizing in the social sciences risks making merely pseudo-scientific progress (Lakatos, 1978, p. 88-9, n. 3-4).

Any entity that experiences such a crisis of (self-)confidence has every right to question its core assumptions. Given the seriousness of the issue, there might indeed be great value in reflecting on the age-old problems of the established norms of scientific inquiry. Here, we investigate how the current undesirable state is related to the problem of empirical underdetermination and its disproportionately detrimental effects in the social and behavioral sciences. We then discuss how close and conceptual replications can be employed to mitigate different aspects of underdetermination, and why they might even aggravate the problem when conducted in isolation. The Systematic Replications Framework we propose consists in logically connected series of close and conceptual replications and will provide a way to increase the informativity of non-corroborative results and thereby effectively reduce the ambiguity of falsification.[1]

Duhem-Quine Thesis and the ambiguity of falsification

Falsificationist methods are widely regarded by the scientific community as the most useful tools in testing the comparative merits of theoretical claims (Dienes, 2008; Hull, 1999; LeBel et al., 2017; Tarantola, 2006). In essence, the falsificationist strategy consists in deriving empirical predictions (P) from a theory (T) and to search for instances that contradict these predictions and thereby refute the theory from which they are derived via *modus tollens*: $(T \rightarrow P \wedge \neg P) \rightarrow \neg T$. It is built on the fundamental asymmetry between confirmation and falsification: While acquiring supportive evidence is trivial and even a huge number of observations do not give us sufficient reason to accept a theory, a single counterevidence is (at least potentially) enough to reject it (Popper, 2002b).

However, this straightforward falsificationist strategy is complicated by the fact that theories do not logically imply any testable predictions. This is because theoretical terms themselves are not observable (only their empirical instances are), and theoretical terms and their empirical instances are not directly linked (Woodward, 1989). So, as the Duhem-Quine Thesis (DQT) famously propounds, scientific theories or hypotheses have empirical consequences only in conjunction with other hypotheses or background assumptions (Laudan, 1990) that help bridging theoretical terms to their empirical instances. For example, for testing a theory concerning intelligence and social class link, you first need to define how intelligence and social class look like in the real world and doing that requires you to make certain assumptions (i.e., auxiliary hypotheses) regarding the nature of these abstract theoretical constructs. These "auxiliary" hypotheses (AH) range from various assumptions regarding the qualities and the execution of the research design and the reliability of the instruments being used, the assessment and/or creation of the experimental conditions, the accuracy of the measurements, the validity of the operationalizations of the theoretical terms linked in the main hypothesis, to the implications of previous

[1] The present paper offers a methodological solution to the problem of underdetermination from a sophisticated methodological falsificationist perspective. For other, such as Bayesian, proposals for dealing with underdetermination, see e.g., Holcombe and Gershman, 2018; Strevens, 2001.

theories and the ceteris paribus clause (i.e., all other things being equal). These assumptions are not necessarily directly derivable from the main theory to be tested. Consequently, it is impossible to test a theoretical hypothesis in isolation. For this reason, falsification is necessarily ambiguous. That is, it cannot be ascertained from a single test if the hypothesis under test or one or more of the auxiliary hypotheses should bear the burden of falsification (see Duhem, 1954, p. 187; Quine, 1969, p. 79; also Strevens, 2001, p. 516).[2] Likewise, Lakatos maintained that absolute falsification is impossible, because in the face of a failed prediction, the target of the modus tollens can always be shifted towards the auxiliary hypotheses and away from the theory (Lakatos, 1978, p. 18-19; see also Popper, 2002b, p. 20).

In the context of single hypothesis testing, we have at the minimum two such auxiliary hypotheses, because the simplest falsifiable scientific proposition hypothesizes a certain relation (e.g. causal or correlational) between two terms (say, $X \rightarrow Y$). More precisely, we need a hypothesis (say, AH_{pre}) that links the theoretical predictor X_t (e.g., 'intelligence') to the observable predictor X_o (e.g., 'academic aptitude, measured through SAT scores') and another hypothesis (say, AH_{out}) that links the theoretical outcome Y_t (e.g., 'social class') to the observational outcome Y_o (e.g., 'control over means of production, measured through occupation').

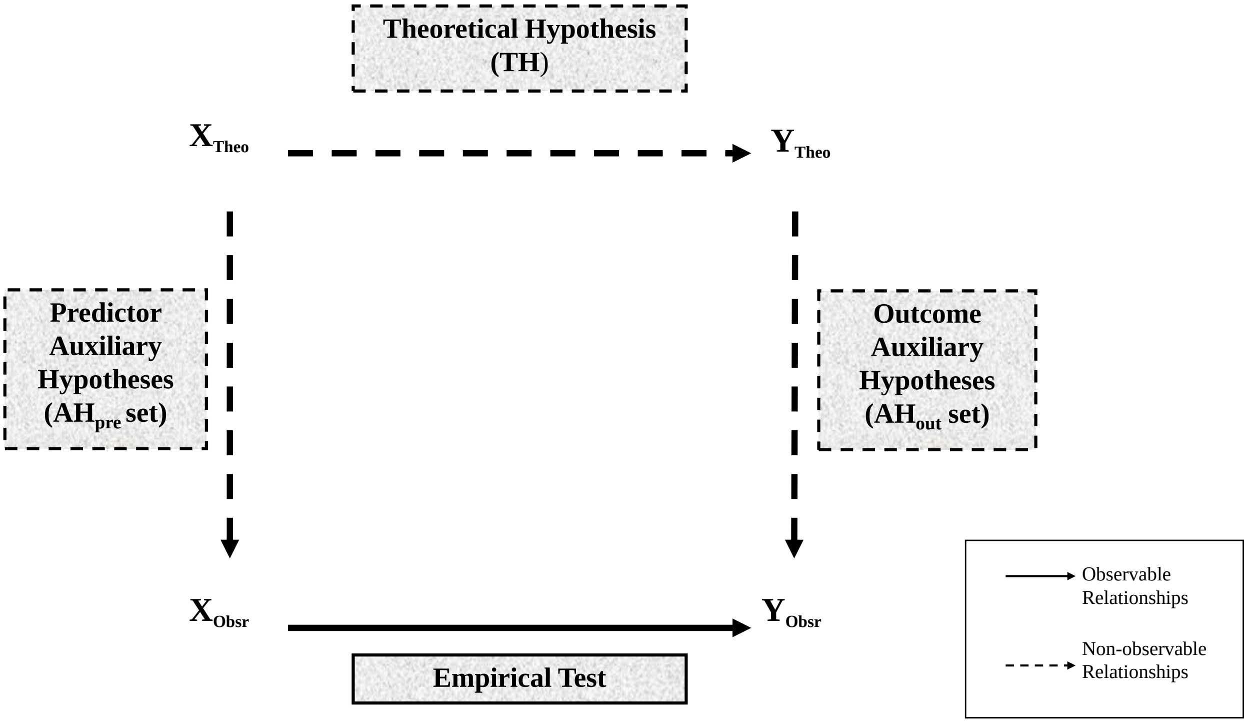
When we reformulate the *modus tollens* of falsification accordingly, our antecedent clause in the first premise becomes a bundle containing at least three elements (TH, AH_{pre}, AH_{out}). If the test results are in disagreement with our prediction, then the conclusion of the *modus tollens* inference would be a negation of the whole bundle. Thus, the ambiguity of falsification as implied by the DQT can be expressed minimally as such: $\neg TH$ or $\neg AH_{pre}$ or $\neg AH_{out}$ (see Figure 1). In this regard, to every isolated empirical test we pose at least three largely independent questions such as, (i) "Does intelligence predict social class?", (ii) "Do SAT scores measure intelligence?", and (iii) "Does occupation capture social class?". And to all we receive just a single answer. Moreover, while the AH_{pre} and AH_{out} can be treated as unitary hypotheses for simplicity, they actually consist in two sets of various assumptions (for instance, the AH_{pre} set comprises “Academic aptitude reflects intelligence,” “SAT scores have adequate reliability,” “Test familiarity is not an issue” etc.). Different assumptions that constitute an *AH* set may become individually highly relevant in designing and interpreting empirical tests and replication studies. Thus, when speaking of the falsity or invalidity of an *AH* set, we also have to take into account that some of its constituent assumptions may still be true or valid.

Popper was aware of the necessity of auxiliary hypotheses and the difficulties they present whenever we try to falsify a theoretical claim. However, Popper relegates *AH*s to unproblematic background knowledge, which the scientist needs to demarcate from the theory under test by independently testing and corroborating them and taking certain methodological decisions (see e.g., Churchland, 1975; Lakatos, 1978; Popper, 2002a, p. 238-239; Popper, 2002b, sections 19-20 and p. 23-28; Uygun Tunç et al., 2023). While Popperian methodological falsificationism does not deny the role of *AH*s in deriving empirical predictions from theories, it suggests that we set up our investigation so that there would be little reason to regard them as part of the empirical test situation (for instance, the measures might be well validated in other independent studies, so even when they are in the test bundle, they can be considered as not contributing to underdetermination). Then we can be in a position to regard the empirical test as a fight exclusively between a theoretical claim and evidence. Accordingly, methodological falsificationism condemns the allocation of blame to *AH*s after a failed test as an inadmissible ad hoc maneuver.

In the social and behavioral sciences, depending on the state of a particular literature or the nature of the construct, it may sometimes be the case that some *AH*s or their constituent assumptions are preferable to their alternatives on independently established theoretical grounds, in reference to widely endorsed disciplinary norms or for directly observational reasons. For instance, in a subsequent replication of the "elderly-slow" priming effect (Doyen et al., 2012), the outcome variable (walking speed as leaving the lab) was measured via sensors instead of handheld stopwatches that were used in the original study (Bargh et al., 1996). Clearly, it is possible to infer on theoretical and empirical grounds that sensors offer higher precision as a measurement instrument than handheld stopwatches. Therefore, the particular component of the AH_{out} concerning the novel method of measurement (i.e., the accuracy of laser sensors) can be more easily regarded as an “unproblematic background assumption.”

[2] This problem is also called “holist” underdetermination (Stanford, 2017). Empirical underdetermination of theories also has serious implications for the issue of theory choice, since the same body of evidence can support alternative theories equally, which is customarily called “contrastive” underdetermination. This paper addresses empirical underdetermination only as it bears on falsification.

Figure 1. Testing a bundle consisting of *TH* and *AH*s


Theoretical Hypothesis (TH)
X_{Theo}
Y_{Theo}
Predictor Auxiliary Hypotheses (AH_{pre} set)
Outcome Auxiliary Hypotheses (AH_{out} set)
X_{Obsr}
Y_{Obsr}
Empirical Test
Observable Relationships
Non-observable Relationships

Not-so-unproblematic background assumptions

However, promoting the components of AH_{pre} and AH_{out} sets to the rank of unproblematic background assumptions is often a quite formidable task. For example, it might be the case that the suspect *AH*s are not of the sort that can be independently corroborated (cf. Rowbottom, 2010) or not embedded in some well-established theory or widely accepted theory of measurement (See e.g. Muthukrishna and Henrich, 2019. The problem is further complicated by the possibility that an *AH* can receive blame not merely to save a theory from refutation by an ad hoc maneuver, but rightly so (Lakatos, 1978); for instance, when a malfunctioning instrument or coincidental choice of an extreme sample prevent the predicted effect from being realized.

In the social and behavioral sciences, treating *AH*s as unproblematic background assumptions is particularly difficult, and consequently the implications of the DQT are particularly relevant and crucial (Meehl, 1978; 1990). For several reasons we need to presume that problematic *AH*s nearly always enter the test along with the main theoretical hypothesis (Meehl, 1990). Firstly, in the social and behavioral sciences the theories are so loosely organized that they do not say much about how the measurements should be (Folger, 1989; Meehl, 1978). Secondly, *AH*s are seldom independently testable due to being heavily theory-laden or difficult to experimentally isolate (Meehl, 1978; 1990). Consequently, often no particular operationalization qualitatively stands out. Moreover, in these disciplines, theoretical terms are often vague (Green, 2019; Meehl, 1978; 1990), value-laden (Weber, 2017), and hard to formalize (Eronen and Romeijn, 2020) or to semantically close (MacCorquodale and Meehl, 1948). Together with the fact that researchers have less control on the environment of inquiry, this implies that hypothesized relationships can be expected to be spatiotemporally less reliable (Leonelli, 2018), and covered in ambient noise (i.e., crud factor; see Orben and Lakens, 2020). Moreover, in the absence of a strong theory of measurement that is informed by the dominant paradigm of the given scientific discipline (Muthukrishna and Henrich, 2019), the selection of *AH*s is usually guided by the assumptions of the very theory that is put into test. Consequently, each contending approach develops its own measurement devices regarding the same phenomenon, heeding to their own theoretical postulations. Attesting to the threat this situation poses for the validity of scientific inferences, it has recently been shown that the differences in research teams' preferences of basic design elements drastically influence the effects observed for the same theoretical hypotheses (Landy et al., 2020). For all these reasons, the problem of underdetermination is usually more acute in social and behavioral sciences.

The problem of underdetermination as regards replication studies

It can be argued that one of the main functions of replication studies has always been tackling various aspects of the problem of underdetermination. While close replications test auxiliary hypotheses such as the reliability of the instruments or that the original finding is not a statistical fluke, conceptual replications test other auxiliaries such as the ones that pertain to the operationalization of variables of interest. This is arguably one of the main reasons why the scientific community came to regard replications as a “cornerstone of science” (Moonesinghe et al., 2007; Simons, 2014) or even as the "gold standard" (Bonett, 2012).

However, the results of single replication studies are similarly ambiguous because they too rely on isolated tests to rule out at least three independent hypotheses at once (i.e., those associated with the AH_{pre}, the AH_{out}, and the *TH*), and there is no way to reach a definitive answer as to which of the three was corroborated or disconfirmed by the observation. It is argued that falsifiability goes hand in hand with replicability (e.g., Earp and Trafimow, 2015). But if replications also at best only diagnose the truth value of the *TH* & *AH*s bundle without indicating whether the *TH* itself or any number of *AH*s are chiefly responsible for the observed results, they might even aggravate the ambiguity of falsification.

Although not necessarily addressing the implications of the DQT, similar arguments have already been voiced with respect to close and conceptual replications. For instance, conceptual replications, and particularly the ones that yield non-corroborative results, are purported to be relatively uninformative and susceptible to be easily brushed aside by the original author (Nosek et al., 2012; Pashler and Harris, 2012), since it is not clear if the differences between the original study and replications indicate a problem with the *TH* or the *AH*s in the replication study. Due to the problem of underdetermination, unsuccessful close replications also cannot provide the scientific community with definitive answers regarding which element in the test bundle is responsible for the results, as the discussions about hidden moderators, sampling characteristics and sundry other differences between the original and replication studies following failed close replications illustrate (see Stroebe, 2019 for a summary). The problem of underdetermination is not dissolved when close replication attempts are successful either, since the observed effect might be an artefact of particular operationalizations of the predictors and outcomes, and hence close replications cannot be regarded as the ultimate test of a hypothesis (Shadish et al., 2002; Stroebe and Strack, 2014). Still others have argued against the very association between replicability and the truth (or verisimilitude) of theoretical claims, maintaining that studies with false results might be highly replicable (e.g., Devezer et al., 2021; Hacking, 1992; Mayo, 2018; Shadish et al., 2002).

We think that the main methodological function of both close and conceptual replications is to reduce the epistemic risks associated with the problem of underdetermination. However, each type of replication is effective in investigating only particular types of alternative explanations associated with problematic *AH*s. In order to establish a stronger connection between the theory and its test, we need to garner the advantages of both close and conceptual replications while controlling for their respective weaknesses. In this way, it can be possible to dissociate the *TH* and the *AH*s to a certain extent by organizing replications into a pre-planned series whose parts are designed so as to systematically vary the *AH*s associated with predictor and outcome variables.

Systematic Replications Framework

The *TH*s under examination in the social and behavioral sciences often are submerged in a complex bundle of potentially confounding *AH*s which cannot be relegated to unproblematic background knowledge due to the reasons mentioned in the previous sections. In this kind of situations, it is still possible to investigate conditional relationships between particular theoretical models (i.e., consisting of *TH*, AH_{pre}, AH_{out}) and test results. Learning more about these relationships would allow us to incrementally identify the most problematic parts of the complex bundle of hypotheses under examination. This, in turn, would significantly reduce the ambiguity of falsification due to underdetermination. Consequently, hypothesis tests in social and behavioral science would become much more informative.

The hypothesis testing and replication framework we propose (Systematic Replications Framework or SRF) is a methodological procedure for investigating such conditional relationships. SRF consists in a systematically organized series of replications that function collectively as a single research line. The main idea behind SRF is coordinating replication studies in such a way that they reveal if the corroboration of the *TH* is restricted to particular theoretical models with particular *AH*s in the AH_{pre} and AH_{out} sets. By revealing such conditional relationships, it is also possible to track how a theory responds to recalcitrant instances, so SRF is also a tool for assessing if a theory acquires a progressive or degenerative character over time (see Lakatos, 1978).

This is an achievable aim, we believe, because in each (close and conceptual) replication study only a limited number of *AH*s in the AH_{pre} and AH_{out} sets are varied. By tracking how the results change in tandem with the exact set of *AH*s utilized in each individual replication, we can significantly reduce underdetermination.

The AH_{pre} and AH_{out} sets

Since selecting the *AH*s to be varied in the AH_{pre} and AH_{out} sets is is at the heart of this strategy, we need to explicate in more detail what we mean by these sets before we lay out SRF. In any hypothesis testing situation, the number of *AH*s is potentially infinite (Quine, 1951; Lakatos, 1978). So, accounting for all of them is practically unattainable. For example, in the social and behavioral sciences (depending on the particular questions investigated) it is usually assumed that the exact color of the lab walls, the elevation of the lab above the sea level, the exact design of the chairs used by the subjects, the humidity of the room that the study takes place or many other minute details do not significantly influence the study outcomes. It would be a Herculean task even to identify each and every one of such hypotheses, and it is impossible to make even the simplest observation in the absence of them. Therefore, since in both original and replication studies we are potentially dealing with an infinite number of *AH*s, we cannot practically keep track of all *AH*s that are altered. Therefore, one might think that isolating the effect of individual *AH*s by monitoring the changes in *AH* sets in the original and replication studies is a task doomed from the start.

However, among the plethora of different *AH*s existing in a hypothesis test only a certain subgroup of *AH*s can be expected to meaningfully impact the results. There are infinitely many other *AH*s that presumably do not exert a meaningful enough influence on the results due to being completely inconsequential, or only barely consequential so that their influence can be safely ignored to a certain extent, or coinciding with opposing factors that always nullify the potential effect, and so forth. *AH*s that are thought to belong this category are relegated to the ceteris paribus clause (Meehl, 1978). As long as they are deemed to belong to the ceteris paribus clause, they are not explicitly stated, and thus are not tested and (tentatively) accepted as they are.

The remaining *AH*s such as the reliability of the measures used in the study, whether the variables are operationalized in a way that is true to the theoretical construct (i.e., accurate and exhaustive) and various other factors that are associated with sample and treatments/measures interactions (e.g., if measures are appropriate to be used in a particular cultural context) are all crucially influential in testing a *TH*. It is such hypotheses that comprise the AH_{pre} and AH_{out} sets. The design elements (including the statistical analysis strategy) featured in a well-written methods section of a scientific paper can also be thought as specifying the AH_{pre} and AH_{out} sets.

A practical example of how the AH_{pre} and AH_{out} sets diverge from other *AH*s that fall under the ceteris paribus clause can be found in the very idea of close replications. It has been rightly pointed out that no close replication is an exact copy of an original study (i.e., it is not possible that all *AH*s are the same in two different studies), yet close replications still serve as an important part of cumulative science because of the role they play in establishing intersubjective agreements on facts (Brandt et al., 2014). Close replications can play this role despite being "different" from the original study in many different ways, because for the theoretical model under investigation only the *TH* and the AH_{pre} and AH_{out} sets are expected to matter, while the other *AH*s under the ceteris paribus clause are not. So, even if exact resemblance between the original and the replication study is beyond the bounds of possibility, close replications can establish intersubjective testability by reiterating the elements that were expected to matter in the original study (i.e., the AH_{pre} and AH_{out} sets).

From a methodological perspective, the problem of underdetermination can be conceived as a (mis)specification problem regarding the AH_{pre} and AH_{out} sets. That is, in a perfectly specified theoretical model (in terms of the AH_{pre} and AH_{out} sets and the ceteris paribus clause) only the *TH* can be held to account for the test results. However, this is almost never the case. Firstly, one or more elements in the AH_{pre} and AH_{out} sets can be false, invalid or in conflict with other elements. Secondly, nontrivial *AH*s might be erroneously relegated to the ceteris paribus clause by the theoretical model. This latter kind of misspecification is always a logical possibility no matter how severely we test our theoretical model. What this means is that the possibility of misspecification can never be conclusively excluded. That being said, by probing these two sources of misspecification we can significantly diminish the degrees of underdetermination.

We can utilize mainly three types of probes in investigating the possibility of misspecification, whereby we would decrease the epistemic risks due to underdetermination and increase test severity: 1) The theoretical model (i.e., the *TH* and the AH_{pre} and AH_{out} sets) should be associated with relatively stable observations, 2) the boundary conditions of a theory defined by the particular AH_{pre} and AH_{out} sets should not be too limiting, and 3) if certain *AH*s are featured in the AH_{pre} and AH_{out} sets of a contending theory while they were relegated to the ceteris paribus clause in another theory, the difference should be accounted for. We can call these the stability probe, boundary conditions probe, and the contending theories probe. Next, we describe how SRF proceeds and examine how the three misspecification probes are used in SRF via systematically organized close and conceptual replications.

SRF: The procedure

SRF is a testing framework with 6 distinct steps. The first and the second steps consist in an original study and its close replications. Next, series of conceptual replications are conducted for testing the AH_{pre} and AH_{out} sets which are also followed by associated close replications (steps 3 to 6). A visual summary of SRF can be seen in Figure 2.[3] The figure has two parts. The upper part illustrates how observable variables are alternated in different steps for comparing the effects of individual AH elements in the AH_{pre} and AH_{out} sets specified by the theoretical model. In the first 2 steps, the same sets of AH_{pre} and AH_{out} are being tested with different samples. From step 3 to step 6, the AH_{pre} and AH_{out} sets are varied. The lower part of the figure describes how this variation of the AH_{pre} and AH_{out} sets is done in detail. The lowercase letters in the lower part of the figure stand for individual auxiliary hypotheses that constitute the AH_{pre} and AH_{out} sets. The sets are represented as finite Venn diagrams and a representation of the ceteris paribus clause is omitted for the sake of simplicity. The intersection represents the *AH*s that are not directly related to the predictor or outcome variables but pertain to the test as a whole, such as the assumptions of the statistical model (e.g., sampling related assumptions). Highlighted lower letters stand for the elements that are varied in that step.

SRF starts with close replications of an original finding and close replications are conducted in 3 out of 6 steps in the procedure. Close replications are indispensable for scrutinizing if the theoretical model (i.e., the *TH* and the AH_{pre} and AH_{out} sets) is associated with relatively stable observations (Schmidt, 2016; Simons, 2014). So, close replications correspond to the stability probe that we introduced earlier. Stability probes investigate if the theoretical model in the original study erroneously relegated some potentially problematic *AH*s (e.g., hidden moderators, cultural context dependency, “flair” or expertise of the researcher who conducts the study[4]) into the ceteris paribus clause, so the replication study unintendedly varied it and ended up making different observations than the original study. Investigating if a theoretical model can generate stable observations is also a test for statistical (i.e., sampling related) *AH*s. Therefore, in the context of SRF, if we obtain conflicting results between the original study and its close replications, the possible implications are: 1) the non-corroborative findings in the replication are due to type 2 error (the higher powered the close replications are the less chance that this is true), 2) theoretical model in the original study is misspecified in terms of *AH*s that are relegated to the ceteris paribus clause or 3) the original corroborative finding was due to type 1 error and the theoretical model is wrong.

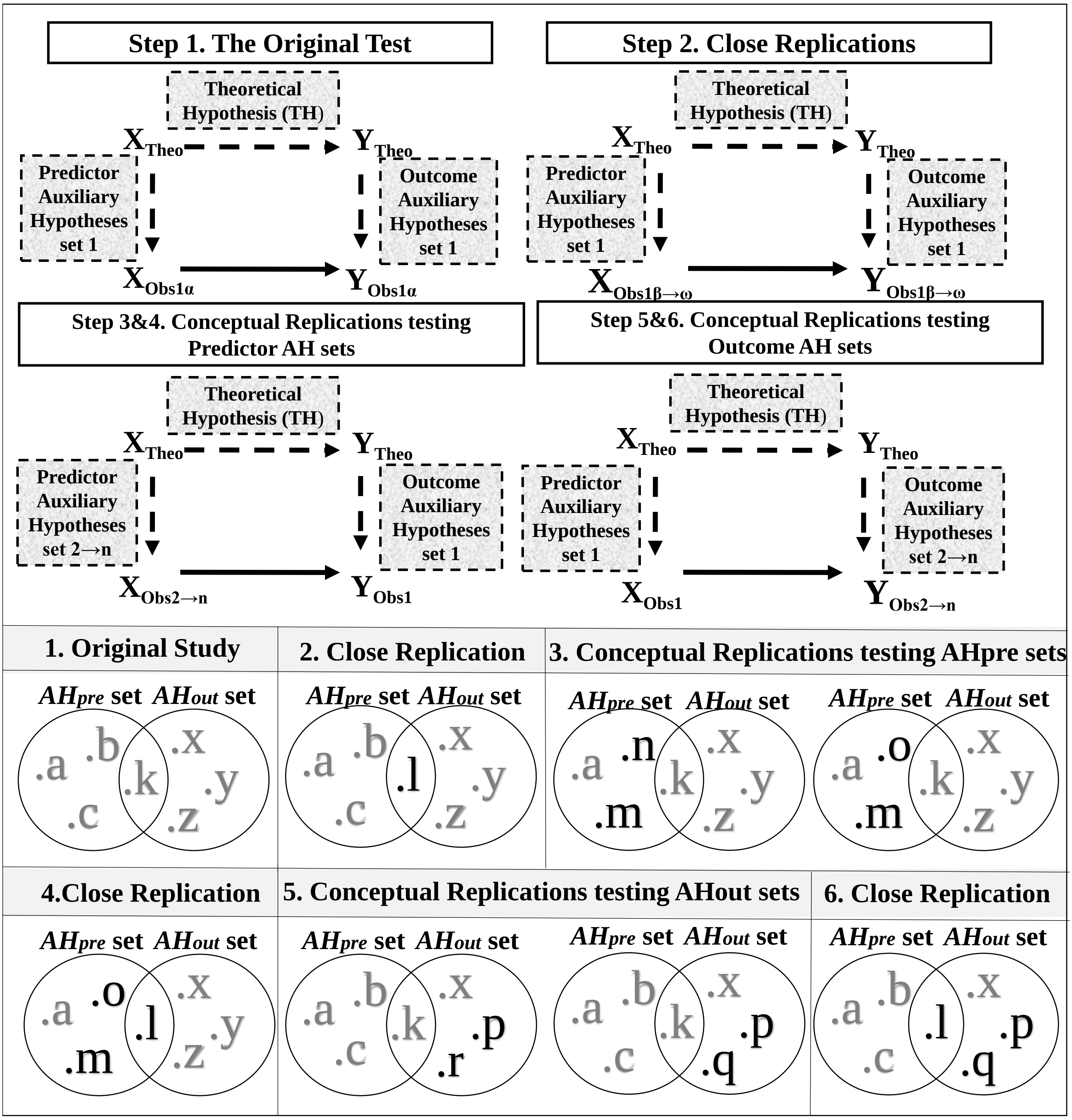
In the face of non-corroborative replications, if the implication 1 is deemed tenable (i.e., type 2 error in replication), then further replications should be conducted. If implication 2 is thought to have more credibility (i.e., the AH_{pre} and AH_{out} sets are misspecified), the route one should take is reformulating the theoretical model with additional elements in the AH_{pre} and AH_{out} sets (e.g., by either describing the limiting instances or defining the hidden moderators). If implication 3 is thought to be more plausible (i.e., the theoretical model is wrong), then seriously modifying or abandoning the theoretical model should be considered. In cases where modifications take place in the *TH* or the AH_{pre} and AH_{out} sets the process should start afresh from the very beginning.

However, for the reasons we discussed so far, neither success nor failure in close replications provides sufficient evidence for reaching a verdict on the corroboration of a *TH*. Due to misspecifications in the model with respect to the *AH*s, artefactual findings might be perfectly replicable or a true effect might persistently elude us. This is also true where a *TH* seems to be repeatedly confirmed by a finding which it completely misrepresents, such as the theory of phlogiston which

[3] A decision guide explicating in more detail how to proceed in different research scenarios can be found in the supplementary materials.

[4] It can be suggested that the “skill or expertise of the replicators” might not be as easily dismissible as an unreliable auxiliary hypothesis as something like “flair.” Admittedly, most scientific studies require expertise in a particular method or domain of research to be properly conducted. That being said, if one assumes that intersubjective testability or replicability is a desirable property for confirmatory research, the term “skill” can be understood only as objectively defined adequate experience and training, which might amount to having an academic qualification (e.g., a degree in a relevant field) or demonstrable expertise in certain techniques (e.g., demonstrable previous experience with the online experiment software). Understood thus, however, “lack of skill” explanation cannot be justifiably applied to other researchers who have been conducting studies in the same field, if one does not forego the claim for intersubjective testability and thus objectivity altogether.

Figure 2. The Systematic Replications Framework


Step 1. The Original Test
Theoretical Hypothesis (TH)
X Theo
Y Theo
Predictor Auxiliary Hypotheses set 1
Outcome Auxiliary Hypotheses set 1
X Obs1α
Y Obs1α
Step 2. Close Replications
Theoretical Hypothesis (TH)
X Theo
Y Theo
Predictor Auxiliary Hypotheses set 1
Outcome Auxiliary Hypotheses set 1
X Obs1β→ω
Y Obs1β→ω
Step 3&4. Conceptual Replications testing Predictor AH sets
Theoretical Hypothesis (TH)
X Theo
Y Theo
Predictor Auxiliary Hypotheses set 2→n
Outcome Auxiliary Hypotheses set 1
X Obs2→n
Y Obs1
Step 5&6. Conceptual Replications testing Outcome AH sets
Theoretical Hypothesis (TH)
X Theo
Y Theo
Predictor Auxiliary Hypotheses set 1
Outcome Auxiliary Hypotheses set 2→n
X Obs1
Y Obs2→n
1. Original Study
AHpre set AHout set
.a .b .c .k .x .y .z
2. Close Replication
AHpre set AHout set
.a .b .c .l .x .y .z
3. Conceptual Replications testing AHpre sets
AHpre set AHout set
.a .n .m .k .x .y .z
AHpre set AHout set
.a .o .m .k .x .y .z
4.Close Replication
AHpre set AHout set
.a .o .m .l .x .y .z
5. Conceptual Replications testing AHout sets
AHpre set AHout set
.a .b .c .k .x .p .r
AHpre set AHout set
.a .b .c .k .x .p .q
6. Close Replication
AHpre set AHout set
.a .b .c .l .x .p .q

mistook oxygen for dephlogisticated air (Trafimow and Earp, 2016). If this is the case, even an infinite number of close replications would fail at identifying which *AH*s are erroneously included in or excluded from the AH_{pre}

and AH_{out} sets. To do that, it is necessary to conduct further tests that connect close and conceptual replications in a logically systematic way, which would allow the researchers to identify if or to what extent the corroboration of the *TH* is conditional on particular AH_{pre} and AH_{out} sets.

This is practicable if conceptual replications are conducted with the aim that they will have consequences for distinct *AH*s in the AH_{pre} and AH_{out} sets. If either an element in the the AH_{pre} or the AH_{out} set is changed while the elements in the other set are kept constant, we can track changes in the results to discern which set or element may be chiefly responsible for the difference. For example, a researcher can first keep the operationalization of the predictor variable the same (i.e., keeping AH_{pre1} constant) while using various different outcome variables (i.e., varying AH_{out} sets: $AH_{out2} \rightarrow n$). In the next step, a similar diversification procedure is applied to the variable that was kept constant in the previous step (i.e., varying AH_{pre} sets: $AH_{pre2} \rightarrow n$), and this time the variable that was being varied in the previous step is kept constant (i.e., keeping AH_{out1} constant). If a theoretical claim survives all these steps without being falsified, it can be said that it is severely tested and largely corroborated. When it faces mixed results, SRF allows researchers to relatively isolate the effects of different *AH*s (i.e., different elements in *AH* sets), and to see if their *TH* is conditional on particular operationalizations (i.e., particular AH_{pre} or AH_{out} sets).

Step 3 and 5 in SRF feature such an investigation of conditional relationships. This is the boundary conditions probe we introduced earlier. It serves to identify conditional relationships between the *TH* and the elements of the AH_{pre} or AH_{out} sets, which might significantly delimit the application of the *TH* if not directly undermine it. That being said, a *TH* whose corroboration is strictly dependent on particular AH_{pre} or AH_{out} sets would have only a very limited theoretical and practical use.

We would like to emphasize here that the systematic variation in the *AH* sets is not envisioned to be a random process in SRF. The *AH* sets to be tested should be decided with a view to severely test the main hypothesis (see Mayo, 2018). Test severity, in turn, is inversely related to underdetermination (Oude Maatman, 2021): The more there are probable alternative explanations of the results in relation to individual *AH* elements, the less severe our test will be. Since underdetermination is a relative matter and admits of degrees (Oude Maatman, 2021), severity is also a relative property of tests and can increase or decrease. Aiming for higher test severity means in this context, then, to critically examine the most probable alternative explanations that arise in relation to individual *AH* elements. So, for example, if an *AH* associated with a particular manipulation is suspected to be chiefly responsible for the previous findings (e.g., using handheld watches instead of laser sensors in a priming experiment), then the variation should be targeted at that hypothesis. Examining probable alternative explanations associated with different *AH* sets would be a useful method for selecting the most severe test available at the time and thus it would potentially provide the strongest corroboration for the *TH*.

Another way to increase the test severity by determining the most problematic *AH*s in the testing situation and isolating their effects in subsequent tests is to compare the theoretical model put forward by one theory to the theoretical models of the contending theories. That is, when particular *AH*s that are unspecified (i.e., relegated to the ceteris paribus clause) in a theoretical model are included in the explicit AH_{pre} and AH_{out} sets of a contending theory as crucial factors for the investigated effect, these particular *AH*s should be considered problematic and further investigated. This is the contending theories probe we mentioned earlier. The basic idea behind this probe is that theories might display self-serving bias in selecting the particular *AH*s to be included in AH_{pre} and AH_{out} sets. Utilizing the distinct preferences of contesting theories for nontrivial, explicitly specified *AH*s can offer a cost-effective way to tackle with the self-serving selection of *AH*s to be included in the theoretical model.

In this regard, SRF will find a particularly significant and effective application in the case of contested theoretical claims and questions, especially if it is employed as a framework for hypothesis testing through adversarial collaboration. Contested questions such as the ones we mentioned in the beginning are extremely difficult to definitively answer in the present context, because the scientific community lacks clear criteria for falsifying points of view and disagrees on key methodological issues—a situation which comes close to what Tetlock described as an "epistemic hell" (2006). The idea of adversarial collaboration has been articulated a few times in the recent past (Tetlock, 2006; Mellers et al., 2001) to organize empirical testing of such contested questions. However, it did not find realization except for a couple of cases (e.g., Bateman et al., 2005; Doherty et al., 2019; Matzke et al., 2015). And even when it did, the studies conducted as adversarial collaborations have been isolated tests, so they were plagued with the same problem of underdetermination we discussed throughout. Adversarial collaborations are for resolving disputes, but this very problem renders it hard to reach a rational consensus on what the results mean when they are undertaken for conducting isolated tests and particularly if they produce mixed results (e.g., Doherty et al., 2019). However, in direct contrast to this, in the context of SRF adversarial collaborations would facilitate an active confrontation of conflicting theories in regard to which *AH*s can be safely relegated to the ceteris paribus clause and which others should be seen as a crucial part of the testing situation. In an SRF-based adversarial collaboration, even if the researchers cannot reach consensus regarding the AH_{pre} and AH_{out} sets, they can at least agree on conditionals and thus reach consensus in the appraisal of the outcomes of the whole scheme.

Conventionalist and methodological falsificationist strategies

What happens when the *TH* faces a non-corroborative observation at any point during this procedure? A theoretical model is preferable to its alternatives to the extent that it is proficient in solving the puzzles posed by non-corroborative instances (i.e., "If the *TH* is true how can this anomalous case be explained?"). There are broadly two ways to deal with such puzzles; namely conventionalist and methodological falsificationist strategies (see Popper, 2002b, sections 19-20). The conventionalist strategy, on the one hand, is degenerative/deflationary in the sense that it involves taking the *TH*'s explanatory claim back from some classes of phenomena to save the theory from refutation (e.g., from the *TH* "All swans are white" one retreats to the claim that the whiteness-swanness link envisioned by the *TH* only applies to swans living in Istanbul, or redefines the category 'swan' so that any other color would indicate a different species). The methodological falsificationist strategy, on the other hand, involves commitment to the level of generality to which the *TH* initially put its claim on. That means, if an unexpected observation is reported and the theory is still believed to have some merit, the observation should be assimilated by either appealing to the greater theory (or theories) which the *TH* is dependent upon and proposing a testable hypothesis that explains the discrepancy (e.g., irregularities in Uranus' orbit with regard to Newton's gravitational theory led to the discovery of Neptune), or making some peripheral adjustments to the theory (e.g., when the temporal distribution of fossil records of different species do not lend support to the traditional gradualist approach in Darwinian evolution, punctuated equilibrium model was proposed), or demonstrating the existence of hidden moderators. That requires additional studies (from step 3 to 6) to be conducted to test the new theoretical model that introduces novel *AH*s into the AH_{pre} or AH_{out} sets. If these responses fail to be corroborated by the evidence in subsequent tests, the *TH* should be discarded as a refuted theoretical claim. While both conventionalist and methodological falsificationist strategies for tackling non-corroborative findings in conceptual replications are acceptable on purely logical grounds, the latter is empirically more justifiable.

SRF, theory-ladenness, and the experimenter's regress

Ultimately all decisions regarding which *AH*s are to be relegated to ceteris paribus clause are theory-laden (See Kuhn, 1996), which might lead one to think that all hypothesis tests are in a way circular. That is, theories choose their own benchmarks, and those benchmarks are used in testing theories, leading to what Harry Collins calls "experimenter's regress" (1992). What would make this problem intractable is the absence of theory-independent (i.e., external) epistemic success criteria. Building upon earlier suggestions concerning piecemeal-testing and calibration (Franklin, 1999, p.470-491), we believe regress problem is not intractable and SRF provides a way out of this conundrum. The three probes that we introduced above are precisely conceived as theory-independent methods for investigating theory misspecification. Firstly, when theory misspecification is due to an erroneous relegation of some crucial *AH*s to the ceteris paribus clause, divergent results in close replications can be an indication of such unspecified *AH*s (e.g., hidden moderators). In this regard, the stability probe is not embedded in the theory under test and constitutes an external success criterion. Secondly, the boundary condition probe is by definition theory-independent, as it involves testing the theory with *AH*s that were not included in the initial theoretical model specified by the theory. Consequently, the boundary condition probe can even be conceived as a tool for generating incompatible observations for the theory. Thirdly, the contending theories probe is already the most stringent test of theory-ladenness, since it involves employing AHs that are specified by the rivals of the theory under test. Thus, SRF is impervious to the criticism of theory-ladenness and circularity of testing to the utmost extent this is possible.

To summarize, SRF reduces the degrees of underdetermination and thereby the ambiguity of test results in original studies as well as in close and conceptual replications. Primarily, it allows for non-corroborative evidence to have differential implications for the components of the *TH* & *AH*s bundle. Although empirical underdetermination may never be eliminated (i.e., we may never conduct "crucial tests"), it can thereby be reduced to a sufficient degree that the scientific community can rationally converge on a verdict of falsification or high corroboration. This can be achieved because by reducing the degree of underdetermination we can devise more informative tests and demarcate justified post hoc revisions of theories from merely ad hoc maneuvers to save them from falsification. If research is planned and executed in compliance with SRF, the community can tell with substantially increased safety when rejecting a falsification leads to the discovery of new factors, clarification of conceptual relationships or the improvement of measurement techniques, and when it comes at the price of losing explanatory and predictive power. Thus, SRF facilitates factual and theoretical consensus formation by increasing the informativity of hypothesis tests.

In cases where it is not possible to achieve corroboration, SRF allows demarcating on which pairings from possible AH_{pre} and AH_{out} sets the truth-value of the *TH* is conditional. In all cases, the confounding effects deriving from the *AH*s can be relatively isolated. Lastly, SRF can enable to approximate to an ideal test of a theoretical hypothesis within the methodological falsificationist paradigm by embedding alternative operationalizations and associated measurement approaches into a severe testing framework (see Mayo, 1997; 2018).

Alternative proposals and their underlying philosophies of science

The suggestion that tests should be logically interconnected might not appear entirely new to the reader. Sidman (1960), for example, uses the notion of systematic replication. The idea behind Sidman's systematic replication is that changing one particular research design element at a time (such as the sampling strategy) in successive studies can allow researchers to test the internal consistency and generalizability of their original findings. Lykken's (1968) constructive replication is another example, where researchers replicate the original study with different operationalizations of the same constructs. By getting beyond the limitations of particular operationalizations, it is suggested that researchers will be able to test the hypothesis of "real interest;" that is, the hypothesis that links the theoretical constructs (hence the name "constructive" replication). There are other similar, more recent suggestions for designing meta-studies, where independent experimental variables are indiscriminately randomized (Baribault et al., 2018), or different operationalizations are introduced as random factors into studies (Yarkoni, 2022; see also Barr et al., 2013 on random effects). Triangulation, another concept, also indicates the need for diversifying and connecting replications (Munafò and Davey Smith, 2018).

However, despite the superficial similarity, the underlying philosophy of science and relatedly the concrete objectives of these methods are very different from those of SRF. First, SRF differs from the methods that rely on randomization in regard to the role they assign to *AH*s in science. Operationalism, which largely constitutes the philosophical framework that randomization-based approaches operate in, purports that the meaning of a concept is exhausted by the empirical justification provided for the existence of its referent (Bridgman et al., 1927, p. 5). In other words, a concept consists in nothing but the set of operations used to empirically measure or manipulate its referent. Thus, the set of operations is not a sign, more particularly an index, of a theoretical entity or property that is conceptually represented in a construct—operations do not measure or manipulate anything beyond themselves. Randomization-based approaches remain faithful to the basic tenets of operationalism, but extend the definitions of concepts (i.e., operational definitions) to all possible operationalizations, arguably in order to address the surplus meaning problem.[5] It is quite obvious that no particular operationalization can perfectly capture the underlying concept as each individual operationalization introduces some random and systematic error to the measurement, but operationalists assume that collectively they can do the job. But it is obviously a practical impossibility to identify, let alone test every possible operationalization of a concept. How can one, then, empirically capture a scientific concept definitively? The solution offered to this problem by randomization-based approaches is to "randomly" select a sample of operationalizations from an imagined universe, in the hope that the errors associated with each operationalization would cancel each other out. This, in turn, would reveal the true nature of the links between concepts, freed from the confounding effects of different sets of operations. We can thus call the philosophical framework offered (though rather implicitly) by randomization-based approaches neo-operationalist.

This neo-operationalism, however, does not really address the problems of classical operationalism previously raised by numerous critics. Among these, a quite serious one is the inherent circularity of how concepts and their measurements are conceived in the operationalist framework –a true chicken and an egg situation (Bickhard, 2001). So, without first arriving at a definition of a concept that incorporates test-independent

[5] The surplus meaning problem is that no matter how meticulously you devise the list of all the possible relevant operationalizations, it is logically and empirically impossible to rule out that the meaning of the construct is not exhausted by them (see Leahey, 1980).

(i.e., non-operational) qualities, it is impossible to decide when and how different measurements can be meaningfully grouped into a concept (Vessonen, 2021). For example, the argument that intelligence is what intelligence tests test is circular, so without a theory of intelligence it is impossible to define the construct intelligence only using the existing "measures of intelligence".

The neo-operationalist thinking behind the randomization-based approaches has its unique problems as well. One of them is how to define the universe of all possible operationalizations of a concept (classical operationalism limits the meaning of a concept to established operations), which is actually a problem more intractable than it first appears to be. For example, it might not be ideal to include a measure that is known for its poor psychometric qualities in that universe just because of its connection to the concept (Köhler and Cortina, 2021). Or we can always (and often do) imagine that future researchers will come up with a much better, previously unthought of measure of a concept that would clearly win out over its existing alternatives (you may think of Popper's black swan in terms of measurement). Therefore, the sampling at any given time might not be sufficiently random (it might be biased towards white swans/hypothesis-confirming measures) as the error associated with these operationalizations are often systematic rather than purely random. Thus, we can never be sure whether the results obtained via existing operations reflect the true underlying relationship between the concepts. It is particularly problematic to cluster good and bad operationalizations together, thinking that the associated errors are always normally distributed and will cancel each other out if random selection is applied.

Furthermore, randomization-based approaches can be said to adhere to a kind of thinking that share peculiarly many features with enumerative induction. As in enumerative induction, the number of confirming instances will be interpreted as the magnitude of supporting evidence for the conclusions reached. Still more problematically, mistakenly believing in the possibility of defining a universe of operationalizations and in the effectiveness of randomly selecting a set of operationalizations in eliminating the error associated with them, these approaches might lead researchers to a false sense of certainty regarding the "true nature" of the relationships between concepts. In this sense, these approaches seem to prescribe a practice of enumerative induction on steroids, so Popper's logical criticisms of verificationism (2002b, p. 1-7; 133-208) apply even more strongly here.

SRF as a severe testing procedure and a tool for assessing Lakatosian progressiveness

Following largely the sophisticated methodological falsificationism of Lakatos (1978), SRF has a very different idea about the role we should assign to *AH*s in science. According to this view, theoretical statements lend themselves to empirical tests only with the help of *AH*s, because they connect core theoretical concepts and relationships to observations. As auxiliary hypotheses, operationalizations do not substitute or collectively exhaust theoretical concepts and relationships. *AH*s can also function as a protective belt that saves the core theory by taking the burden of falsification on themselves. The prevalence of one of these two different roles which *AH*s can play (i.e., increasing testability vs. deflecting falsification) can help us identify respectively whether modifications to theories vis-à-vis accumulating evidence are of a progressive or degenerative character (see also Popper, 2002a, p. 240f.). In progressive research programmes (consisting of successive versions of a theory), *AH*s predominantly increase empirical content by increasing the explanatory and predictive power, and hence generating more potential falsifiers for the core theory, while in degenerative research programmes they often serve a content-decreasing function by putting forward ad hoc alternative explanations that do not suggest any empirical discoveries or novel research questions. Researchers may avoid falsification of the *TH*, on pain of losing explanatory or predictive power and giving their research programme a degenerative character, by continuously refining its terms according to whether particular *AH*s yield corroborative or non-corroborative results, for instance by delimiting the boundary conditions of the *TH* to a pair of operationalizations that work (see also Laudan, 1990, p. 276). Such refinements often result in decreasing the theory's scope, precision or narrowing its semantic reference, and consequently diminishes its empirical support and makes it increasingly less rational to stick to it.

Using an earlier distinction that we made regarding the possible "solutions" in the face of non-corroborating evidence, conventionalist strategies are associated with degenerative research programmes while methodological falsifiactionist strategies are associated with progressive research programmes. In this regard, SRF is also a method for identifying if and to what extent a research programme can be deemed progressive, by tracking how the researchers respond to non-corroborative results (see the supplementary materials for a more detailed exposition). If (or to the extent that) the corroboration of *TH* is made increasingly dependent on certain operationalizations, then the set of *AH*s that comprises these operationalizations can be said to play a falsification-deflecting role. In this respect, SRF facilitates an objective assessment of Lakatosian progressiveness of a research programme.

In SRF the systematic variation of design elements is not a bottom-up and random procedure, but rather is organized with a view to examine the most probable alternative explanations associated with different *AH*s and thereby to increase test severity. In this sense, what we understand from replication is quite akin to "constructive replication" of Köhler and Cortina (2021), where the succeeding replications are conducted with the objective of improving the measures/operationalizations. However, because of the reasons we explained before, it is usually not possible to justify the superiority of one measure over other in social sciences. Under these conditions, the best we can do is to map out on which particular *AH*s the main hypothesis is conditional. By providing a way to accomplish this, SRF increases the transparency of how *AH*s influence "(non-)corroborative evidence," and allows us to evaluate post hoc modifications to theoretical claims vis-à-vis evidence. This in turn can potentially foster progressive theory development and the discovery of novel effects by revealing the weak spots of theories.

Consequently, SRF can be said to have certain theoretical and practical advantages over other systematic replication approaches. The main difference lies in the philosophical commitments. Randomization-based approaches seem to follow a neo-operationalist and inductivist philosophy of science, while SRF rests on sophisticated methodological falsificationism. The objective of hypothesis testing in randomization-based approaches is to collect confirming evidence ("hyper-powered" through randomization), and to inductively verify generalizability of findings as such, while in SRF the aim is to severely test hypotheses by examining the most plausible alternative explanations associated with *AH* sets (for the distinction, see Mook, 1983). In terms of interpretation, confirming results in randomization-based approaches might lead researchers to mistakenly believe that their *TH* reflects the true nature of the relationship between the concepts, despite it is logically invalid to draw such an inductive conclusion no matter how big your sample of operationalizations is (see Popper, 2002b). However, in SRF confirmatory results are interpreted only as further corroboration, and the door is never closed for possible alternative explanations and discovery of systematic errors due to particular *AH*s. Non-confirmatory results are also very hard to interpret in randomization-based approaches, as it is impossible to know the sample characteristics of a given set of randomly chosen operationalizations without having a justifiable opinion about the universe from which they are selected. Whereas in SRF, being a falsificationist method that aims to disentangle *AH* dependencies, non-confirmatory results are much more informative. SRF shares the core advantage of falsificationist frameworks in confirmatory hypothesis-testing settings: Trying to devise a testing situation that maximizes the chances of finding a falsifying instance if there is any (i.e., conducting severe tests) is a more attainable goal than collecting verifying examples and/or weighing the rational belief in a hypothesis by enumerating such instances (Mayo, 2018), even when collecting all the verifying examples is not deemed necessary thanks to randomization. Lastly, unlike some randomization-based approaches, SRF also does not require conducting mega studies and allows hypothesis testing to be realized in a step-by-step fashion, which also provides flexibility. That being said, we do not completely reject that random sampling of operationalizations might have a use. The famous distinction of Reichenbach (1938, p. 7) between the context of justification and the context of discovery is to the purpose here. The present falsificationist criticism of the inductivist tendencies in neo-operationalist, randomization-based approaches only applies if these methods are implemented in the justification context, thus in confirmatory studies. Hypothesis generation is not bound by the strict logical validity criteria of hypothesis testing. In the context of discovery, hyper-powered exploration via random selection of operationalizations can be considered perfectly kosher. However, the context of justification necessitates logically valid inferences, which is exactly what SRF aims to facilitate.

Practical Implications

As it stands, SRF can be said to have practical implications for three broad domains of scientific inquiry, namely 1) Hypothesis testing via providing a severe testing framework for self-replications or collaborative projects, 2) Replication studies via coordinating close and conceptual replications into a more coherent, informative and critical body of investigations and 3) Literature reviews via offering an alternative structure of clustering the existing findings in terms of the *AH* sets that generate them.

We already examined how SRF can help us in disentangling *AH* and *TH* driven effects in a systematic series of close and conceptual replications under different research scenarios (see also the supplementary material), and how this replication effort might be most fruitfully realized through adversarial collaboration. Now we discuss how a similar systematic approach can be implemented in organizing self-replication attempts or collaborative research projects, and also in straightening up an existing body of findings into a meaningful network of relationships in a systematic literature review.

SRF can be implemented as a hypothesis-testing procedure in place of a single study in order to substantially increase test severity across a pre-planned battery of systematically organized tests. One obvious way to do this is by conducting self-replications as an integral part of the hypothesis-testing procedure. Replicating an initial finding before publication (i.e., self-replication) has long been considered among the best practice (Cesario, 2014; Roediger III, 2012). Nevertheless, the DQT-related problems (which render the results of isolated close/conceptual replications nothing but tentative) are also relevant for self-replication efforts. Since the problem of underdetermination equally applies to self-replication studies, organizing them into a logically connected set of replications that systematically vary sets of AH_{pre} and AH_{out} will substantially increase their informativity.

A self-replication attempt planned in compliance with the requirements of SRF follows a similar procedure as we described for other replication studies. So, also herein an initial hypothesis test should be re-examined with a close replication. Then, the hypothesis should be further investigated by conceptual replications that systematically vary the *AH*s. The main idea again is to link close replications to conceptual replications with a view to increase tests severity by decreasing the degree of underdetermination; that is, to become able to determine if the inconsistent results are driven by one or more of the *AH*s or by the *TH* (thus suggest that we modify or abandon the *TH*).

We also recommend pre-registering the whole SRF plan before the data collection. At present, the common practice is to pre-register only a single study (or a single set of studies) where the operationalization of variables (and hence the *AH*s) are kept constant. This conventional practice of pre-registering only a single set of operationalizations might pave the way for a setting that condones conducting multiple studies and selectively reporting the studies that corroborated the *TH*. Pre-registering SRF in the context of self-replication can decrease the researcher degrees of freedom (see Simmons et al., 2011). Realistically it would be a tentative plan, but it would still inform both the researchers and their audience about the initial expectations. And since SRF-compatible pre-registrations can offer broader protection against researcher degrees of freedom, a separate badge (that is similar to the ones awarded for pre-registration or open data/code) can be bestowed on studies that satisfy the criteria.[6]

That being said, it is important to note here that self-replications never quell the need for independent replications, as whether the experimenter's bias (Rosenthal and Fode, 1963) influences the results is an *AH* that needs almost always to be taken seriously. A still more rigorous strategy to hypothesis-testing would be to follow SRF as a collaborative research project, where various steps are distributed among collaborating researchers or teams. This research strategy will be similar to applying SRF as a replication procedure, but instead of evaluating earlier findings and resolving disputes it will serve to severely test novel hypotheses. Since conducting highly controlled studies is more difficult in social and behavioral sciences, increasing test severity over a systematically organized battery of tests is a viable alternative that would serve the same purpose.

Another potential practical implication of SRF lies in using the same strategy of logically connecting different *AH* bundles in conducting and interpreting systematic literature reviews (particularly when the previous findings are mixed). Such a strategy can help researchers distinguish the effects that seem to be driven by certain *AH*s from the ones in which the *TH* is more robust to such influences. To put it differently, in a contested literature there are already numerous conceptual replications that have been conducted, and at least some of these replications rely on the same *AH*s in their operationalizations. Therefore, to the extent that they have overlaps in their *AH*s, their results can be organized in such a way that resembles a pattern of results that can be obtained with a novel research project planned according to SRF. The term "systematic" in systematic literature review already indicates that the scientific question to be investigated (i.e., the subject-matter, the problem or hypothesis), the data collection strategy (e.g., databases to be searched, inclusion criteria) as well as the method that will be used in analyzing the data (e.g., statistical tests or qualitative analyses) are standardized. However, for various reasons (e.g., to limit the inquiry to those studies that use a particular method), not every systematic literature review is conducive to figuring out whether the *TH* is conditional on particular *AH* sets. An SRF-inspired strategy of tabulating the results in a systematic literature review will also help researchers in appraising the conceptual networks of theoretical claims, theoretically relevant auxiliary hypotheses and measurements. Thus, it can eventually help in appraising the empirical support of the *TH* by revealing how it is conditional on certain *AH*s, and can lead to the

[6] Preregistering only individual steps of SRF (as independent studies) will not provide the level of test severity that will be achieved with the whole procedure. For the relation between preregistration and test severity, see Lakens (2019).

reformulation or refinement of the *TH* as well as guide and constrain subsequent modifications to it.

Coda

In this paper, we have suggested, firstly, a methodological procedure that will considerably bolster the social and behavioral sciences' ability to address the problem of empirical underdetermination. While theories are always underdetermined by empirical evidence, we argued that in the context of hypothesis testing it can be possible to reduce certain researcher degrees of freedom with respect to auxiliary hypotheses and thus to facilitate decision making. Achieving this requires, first and foremost, that researchers pay substantially more attention to the auxiliary hypotheses they assume to be true in designing empirical tests. Moreover, it requires that they acknowledge that individual tests cannot investigate the epistemic worth of single scientific hypotheses, let alone of theories.

On a more general note, opting for a series of systematically interconnected tests instead of single studies in deciding the fate of scientific theories implies a more critical process of scientific inquiry, which would also require a relatively higher investment of time and resources than a typical empirical study in the social sciences. We should, however, weigh this extra investment not against a single study, but against the current situation where the same amount of time and resources yield numerous independent studies which are much less informative individually as well as collectively. Moreover, SRF-style research strategies can be implemented much faster and with much less burden falling on individual researchers by way of scientific collaboration; that is, through collective testing and appraisal of scientific theories. Today researchers have at their disposal more opportunities for large scale collaborations, such as the Psychological Science Accelerator (Moshontz et al., 2018).

Clearly, the methodological decision between more rigorous tests and quicker decisions on the empirical worth of theories is bound to be a collective one, which reflects our collective take on scientific priorities. We can generally speak of two central missions of scientific inquiry; namely, extending the established body of knowledge to include novel phenomena (i.e., science's exploratory mission) and to weed out false theories via testing and replication (i.e., science's critical mission).[7] Depending on the state of a particular discipline or research programme, one or the other of these two missions might be more accentuated. While in expansionist periods accumulation of novel hypotheses is prioritized over severe tests, replications of earlier studies or critical assessment of literature, during moments of crisis the need for disciplinary self-reflection might overcome that for novelty and growth. The decade-long discussion on a replicability and confidence crisis in several disciplines of social, behavioral and life sciences (e.g., Camerer et al., 2018; Open Science Framework, 2015; Ioannidis, 2005) has identified the utilization of methods that would be most fitting for accomplishing the exploratory mission in the context where the critical mission should be prioritized as being one of the main causes of the problem. This diagnosis led to proposals for slowing science down (Stengers, 2018), applying more caution in giving policy advice (IJzerman et al., 2020), and inaugurating a credibility revolution (Vazire, 2019). All potential contributions of SRF will be part of a strategy to prioritize science's critical mission on the way towards more credible research in social, behavioral, and life sciences. This would imply that the scientific community focuses less on producing huge numbers of novel hypotheses with little corroboration and more on having a lesser number of severely tested theoretical claims. Successful implementation of SRF also requires openness and transparency regarding both positive and negative results of original and replication studies (Nosek et al., 2015) and demands increased research collaboration (Landy et al., 2020). Ideally, this would also take the form of adversarial collaboration.

Author Contact

Duygu Uygun Tunç Philosophy & Ethics Group / Eindhoven University of Technology, the Netherlands Philosophy Department/ Middle East Technical University, Ankara, Turkey ORCID: https://orcid.org/0000-0003-0148-0416

Mehmet Necip Tunç Social Psychology Department/ Tilburg University, the Netherlands ORCID: https://orcid.org/0000-0002-1350-174X

Conflict of Interest and Funding

There are no conflicts of interest associated with this research. This research is partly funded by the European Union and the Turkish Scientific and Technological Research Council under the Horizon 2020 Marie Skłodowska-Curie Actions Cofund program Co-Circulation2.

Author Contributions

Both authors have contributed equally to conceptualization, project administration, visualization, original draft preparation, review, and editing.

[7] A similar distinction is made in Longino (1990).

Acknowledgements

The authors would like to thank Mark Brandt, Willem Sleegers, Daniël Lakens, Choi Hong Hui, Agnes Bolinska, Nico Formanek, the two reviewers at Meta-Psychology Sophia Crüwell and Maarten Derksen and the editor Nick Brown, the participants of two invited talks at Eindhoven University of Technology and Boğaziçi University for critical comments and suggestions on earlier versions of this paper, which helped to substantially improve it.

Open Science Practices

This article is theoretical and as such received no Open Science badges. The entire editorial process, including the open reviews, is published in the online supplement.

References

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, *71*(2), 230.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, *20*(1), 1–68.

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*(8), 1561–1580.

Bickhard, M. H. (2001). The tragedy of operationalism. *Theory & Psychology*, *11*(1), 35–44.

Bonett, D. G. (2012). Replication-extension studies. *Current Directions in Psychological Science*, *21*(6), 409–412.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Bridgman, P. W., Bridgman, P. W., Bridgman, P. W., & Bridgman, P. W. (1927). *The logic of modern physics* (Vol. 3). Macmillan New York.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., & Pfeiffer, T. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on psychological science*, *9*(1), 40–48.

Churchland, P. M. (1975). Karl popper's philosophy of science. *Canadian Journal of Philosophy*, *5*(1), 145–156.

Collins, H. (1992). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society open science*, *8*(3), 200805.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.

Doherty, J. M., Belletier, C., Rhodes, S., Jaroslawska, A., Barrouillet, P., Camos, V., Cowan, N., Naveh-Benjamin, M., & Logie, R. H. (2019). Dual-task costs in working memory: An adversarial collaboration. *Journal of experimental psychology: learning, memory, and cognition*, *45*(9), 1529.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS one*, *7*(1), e29081.

Duhem, P. (1954). *The aim and structure of physical theory* (Vol. 13). Princeton University Press.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621.

Ellemers, N., Fiske, S. T., Abele, A. E., Koch, A., & Yzerbyt, V. (2020). Adversarial alignment enables competing models to engage in cooperative theory building toward cumulative science. *Proceedings of the National Academy of Sciences*, *117*(14), 7561–7567.

Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, *30*(6), 786–799.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *American Psychological Association*.

Franklin, A. (1999). *Can that be right?* Springer.

Green, B. (2019). The essential ambiguity of the social. *Philosophy of the Social Sciences*, *49*(2), 108–136.

Hacking, I. (1992). The self-vindication of the laboratory sciences. *Science as practice and culture*, *30*.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of anderson et al.(2010). *Psychological Bulletin*, *143*, 757–774.

Holcombe, A. O., & Gershman, S. J. (2018). Bayesian belief updating after a replication experiment. *Behavioral and Brain Sciences*, *41*.

Hull, D. L. (1999). The use and abuse of sir karl popper. *Biology and Philosophy*, *14*(4), 481–504.

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., Vazire, S., Forscher, P. S., Morey, R. D., & Ivory, J. D. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*(11), 1092–1094.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Köhler, T., & Cortina, J. M. (2021). Play it again, sam! an analysis of constructive replication in the organizational sciences. *Journal of Management*, *47*(2), 488–518.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge University Press.

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis., *62*(3), 221–230.

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., & Gronau, Q. F. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451.

Laudan, L. (1990). Demystifying underdetermination in savage CW (ed.), scientific theories (pp. 267–297).

Leahey, T. H. (1980). The myth of operationism. *The Journal of Mind and Behavior*, 127–143.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of personality and social psychology*, *113*, 254–261.

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. *Including a symposium on Mary Morgan: curiosity, imagination, and surprise*.

Lin, H., Saunders, B., Friese, M., Evans, N. J., & Inzlicht, M. (2020). Strong effort manipulations reduce response caution: A preregistered reinvention of the ego-depletion paradigm. *Psychological science*, *31*(5), 531–547.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological bulletin*, *70*(3), 151.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological review*, *55*(2), 95.

Matzke, D., Nieuwenhuis, S., Van Rijn, H., Slagter, H. A., Van Der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*(1), e1.

Mayo, D. G. (1997). Duhem's problem, the bayesian way, and error statistics, or "what's belief got to do with it?" *Philosophy of Science*, *64*(2), 222–244.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological reports*, *66*(1), 195–244.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? an exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275.

Mook, D. G. (1983). In defense of external invalidity. *American psychologist*, *38*(4), 379.

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false—but a little replication goes a long way. *PLoS medicine*, *4*(2), e28.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., & Antfolk, J. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Munafò, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., & Christensen, G. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Open Science Framework. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Orben, A., & Lakens, D. (2020). Crud (re) defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247.

Oude Maatman, F. (2021). Psychology's theory crisis, and why formal modelling cannot solve it. https://doi.org/https://doi.org/10.31234/osf.io/puqvs

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Popper, K. (2002a). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Popper, K. (2002b). *The logic of scientific discovery* (2nd ed.). Routledge.

Quine, W. V. O. (1969). Epistemology naturalized. (pp. 38-114). In *Ontological relativity and other essays* (pp. 38–114). New York: Columbia University Press.

Quine, W. (1951). Two dogmas of empiricism. , 60, 20-43. *The Philosophical Review*, *60*, 20–43.

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge*. The University of Chicago Press.

Roediger III, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*. https://www.psychologicalscience.org/observer/psychologys-woes-and-a-partial-cure-the-value-of-replication

Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*(3), 183–189.

Rowbottom, D. P. (2010). Corroboration and auxiliary hypotheses: Duhem's thesis revisited. *Synthese*, *177*(1), 139–149.

Schmidt, S. (2016). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. In *Methodological issues and strategies in clinical research* (pp. 581–596). American Psychological Association. https://doi.org/10.1037/14805-036

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin; Company.

Sidman, M. (1960). *Tactics of scientific research*. Basic Books.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on psychological science*, *9*(1), 76–80.

Stanford, K. (2017). Underdetermination of scientific theory. In *The stanford encyclopedia of philosophy* (Winter 2017). The Metaphysics Research Lab. https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

Stengers, I. (2018). *Another science is possible: A manifesto for slow science*. John Wiley & Sons.

Strevens, M. (2001). The bayesian treatment of auxiliary hypotheses. *British Journal for the Philosophy of Science*, *52*(3).

Stroebe, W. (2019). What can we learn from many labs replications? *Basic and Applied Social Psychology*, *41*(2), 91–103.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Tarantola, A. (2006). Popper, bayes and the inverse problem. *Nature physics*, *2*(8), 492–494.

Tetlock, P. E. (2006). Adversarial collaboration: Least feasible when most needed? least needed when most feasible.

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on klein. *Theory & Psychology*, *26*(4), 540–548.

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory Psychology*, *33*(3), 403–423. https://doi-org.proxy.lnu.se/10.1177/09593543231160112

Vazire, S. (2019). Do we want to be credible or incredible? *APS Observer*, *33*.

Vessonen, E. (2021). Respectful operationalism. *Theory & Psychology*, *31*(1), 84–105.

Weber, M. (2017). *Methodology of social sciences*. Routledge.

Wicherts, J. M., Borsboom, D., & Dolan, C. V. (2010). Why national IQs do not support evolutionary theories of intelligence. *Personality and individual differences*, *48*(2), 91–96.

Woodward, J. (1989). Data and phenomena. *Synthese*, 393–472.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*.