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

Duygu Uygun Tunç[1]

Mehmet Necip Tunç[2]

[1]Middle East Technical University, [2]Tilburg University

Correspondence can be addressed to m.neciptunc@hotmail.com

Submitted to *Meta-Psychology*. Participate in open peer review by commenting through hypothes.is directly on this preprint. The full editorial process of all articles under review at *Meta-Psychology* can be found following this link: https://tinyurl.com/mp-submissions You will find this preprint by searching for the first author's name.

Draft Date: 06-01-2020

Word count: 9,130

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

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 problem is not fatal to the extent that *AH*s are independently validated and thereby safely relegated to unproblematic 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) we propose provides a methodological solution by disentangling the implications of the findings for the main hypothesis and the *AH*s through pre-planned series of logically interlinked close and conceptual replications. In this way, SRF provides an objective assessment of whether the corroboration of a hypothesis is conditional on particular *AH*s. SRF endorses a falsificationist, severe testing approach that facilitates testing alternative explanations associated with different *AH*s. It has several theoretical and practical advantages over previous randomization-based systematic replication proposals, which generally assume a philosophically problematic neo-operationalist approach and misleadingly prescribe exploration-oriented strategies in confirmatory contexts.

Introduction

At least 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, Engelhardt, & Rouder, 2017; Lin et al., 2020; Wicherts, Borsboom, & Dolan, 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, Engelhardt, & Rouder, 2017).

Resolving theoretical disputes is an important means to scientific progress because when a given scientific field lacks consensus regarding established evidence and how exactly it supports or contradicts competing theoretical claims, the scientific community cannot appraise whether there is scientific progress or merely a misleading semblance of it. That is to say, it cannot be in a position to judge whether a theory constitutes scientific progress in the sense that it accounts for phenomena better than alternative or previous theories and can lead

to the discovery of new facts, or is degenerating in the sense that it focuses on explaining away counterevidence by finding faults in replications (Lakatos, 1978). Observing this state, Lakatos maintained decades ago that most theorizing in social sciences risks making merely pseudo-scientific progress (1978, p. 88-9, n. 3-4). What further solidifies this problem is that most "hypothesis-tests" do not test any theory and those that do so subject the theory to radically few number of tests (see e.g., McPhetres et. al., 2020). This situation has actually been going on for a considerably long time, 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 (1978).

This is a depressing state for any scientific discipline to be in, as the aim of science is not to accumulate 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. Tellingly, it has even been argued in a widely discussed recent paper that much of the psychology should forgo its claims on being a quantitative enterprise and most of the academic psychologists would do better if they pursue alternative careers anyway (Yarkoni, 2020).

Any entity that experiences such a crisis of (self-)confidence has every right to question its core assumptions. Given the seriousness of the problems, there might indeed be great value in reflecting on the age-old problems of established norms of scientific inquiry. It is always possible that traditional approaches become obsolete in the face of some novel or

not-previously-thought-of problems. However, the outcome of such an undertaking would not necessarily dictate abandoning the prevailing norm. Upon closer inspection, one may discover that the “novel” problem is not that novel or sometimes relatively small amendments to the established norm are just what it takes to address the “novel” problems.

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 involves conducting 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.

The prescriptive norm: Falsificationism

Falsificationism is widely regarded by the scientific community as the methodological norm in testing the comparative merits of theoretical claims (Dienes, 2008; Hull, 1999; LeBel et al., 2017; Tarantola, 2006). Its most paradigmatic form, Popperian methodological falsificationism builds on a critique of induction. Firstly, there is no strictly valid logical procedure for inferring universal statements (such as theories and theoretical hypotheses) from singular statements describing observations. To use a textbook example, no finite number of observations of white swans logically warrant us to conclude the truth of the statement "all swans are white," since there can always be a hitherto unobserved non-white swan. Since confirmations are ubiquitous and trivial to obtain, they are "valueless and uninteresting" (Popper, 1974, p. 991). Popper similarly argued against probabilistic induction; namely, we cannot validly infer the probability of hypotheses from the probability

of events (see especially Popper 2002b, p. 252-267). Thus, observation statements can neither prove, nor inductively confirm theories (see also Lakatos, 1978, p. 11).

The falsificationist thesis thus regards the strategy of proving or justifying scientific theories by facts as a dead-end and directs our attention instead to the possibility of refuting or disconfirming them. In essence, the falsificationist strategy consists in deriving empirical predictions (*P*) from a theory (*T*) and to search for those instances that would contradict these predictions and thereby refute the theory from which they are derived via the valid *modus tollens* inference: $(T \rightarrow P, \sim P) \rightarrow \sim T$. 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 counter-evidence (e.g., observing a black swan) is potentially enough to reject it.

Duhem-Quine Thesis and the ambiguity of falsification

However, this straightforward falsificationist strategy is complicated by the fact that theories by themselves do not logically imply any testable predictions. As the Duhem-Quine Thesis (DQT from now on) famously propounds, scientific theories or hypotheses have empirical consequences only in conjunction with other hypotheses or background assumptions. These auxiliary hypotheses range from *ceteris paribus* clauses (i.e., all other things being equal) to various assumptions regarding the research design and the instruments being used, the accuracy of the measurements, the validity of the operationalizations of the theoretical terms linked in the main hypothesis, the implications of previous theories and so on. Consequently, it is impossible to test a theoretical hypothesis in isolation. In other words, the antecedent clause in the first premise of the *modus tollens* is not a theory (*T*) but actually a bundle consisting of the theory and various auxiliary hypotheses ($T, AH_1, \ldots, AH_n$). 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; also Strevens, 2001, p. 516).[1] 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 (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 (T, 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: $\sim TH$ or $\sim AH_{pre}$ or $\sim AH_{out}$ (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?", to all of which we receive 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

[1] The implied ambiguity of falsification is often referred to as Duhem's "problem." Empirical underdetermination of theories also has serious implications for the issue of theory choice, since the same body of evidence can support alternative, possibly inconsistent, theories equally (the problem is further aggravated when we also introduce unconceived alternatives). This paper addresses empirical underdetermination only as it bears on falsification.

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.

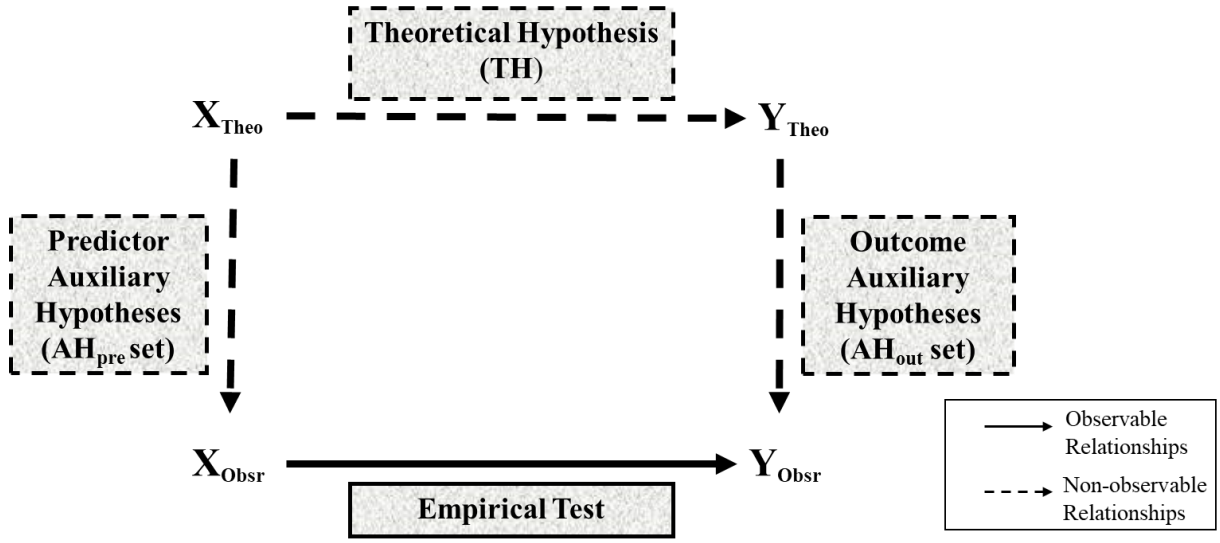

Fig. 1 The ambiguity associated with testing a bundle consisting of *TH* & *AH*s

Popper was aware of the necessity of auxiliary assumptions or hypotheses and the difficulties they present whenever we try to falsify a theoretical claim. However, Popper relegates *AH*s to unproblematic background assumptions, which the scientist needs to *demarcate* from the theory under test by taking certain *methodological decisions* (see e.g., Popper, 2002b, sections 19-20; Lakatos, 1978, p. 23-28; Churchland, 1975). 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.

Also in the social and behavioral sciences, depending on the state of particular literature or the nature of the construct, it may be the case that some *AH*s or their particular 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, Chen, & Burrows, 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."

Not-so-unproblematic background assumptions

However, the demarcation task Popper assigns to the researcher is often not as easy. 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 embedded in some well-established theory or widely accepted theory of measurement (See e.g. Muthukrishna & Henrich, 2019). The problem is further complicated when an *AH* receives blame not merely to save a theory from refutation by an *ad hoc* maneuver, but rightly so. For instance, when a contaminated tube, a malfunctioning instrument, or a misrepresentation of the initial conditions prevents the predicted effect from being realized.

In the social and behavioral sciences, relegating *AH*s to 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 *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 (Meehl, 1978) and, consequently, usually no particular operationalization qualitatively stands out. Besides, in these disciplines, theoretical terms are often necessarily vague (Qizilbash, 2003), and researchers have a lesser degree of control on the environment of inquiry, so hypothesized relationships can be expected to be spatiotemporally less reliable (Leonelli, 2018). Moreover, in the absence of a strong theory of measurement that is informed by the dominant paradigm of the given scientific discipline (Muthukrishna & 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).

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 for the validity of auxiliary assumptions such as the reliability of the instruments or that the original finding is not a statistical fluke, conceptual replications test for the validity of other auxiliaries such as the ones that pertain to the particular operationalizations of variables of

interest. This is arguably one of the main reasons why the scientific community came to regard replications as the "cornerstone of science" (Moonesinghe et al., 2007; Simons, 2014) or 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 (i.e., those associated with the AH_{pre}, the $AH_{out,}$ and the *TH*) at once, and there is no way to reach a definitive answer as to which of the three was corroborated or disconfirmed by the observation. As it is widely supposed, falsifiability goes hand in hand with replicability (Earp and Trafimow 2015; also Popper, 2002b, p. 22). 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, in the long run they might just aggravate the ambiguity.

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, Spies, and Motyl 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, 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, Cook, & Campbell, 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, Navarro, Vandekerckhove, & Buzbaş, 2020; Hacking, 1992; Shadish et al., 2002). Thus, to ground the link between replicability and falsification or corroboration of scientific hypotheses in a more satisfying manner, we need to garner the advantages of both close and conceptual replications while controlling for their respective weaknesses.

To this aim, it can indeed be possible to dissociate the main *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. Popper in fact hints at a possible solution, in a rather different context and in passing, but does not develop his germinal idea further into an effectively realizable methodological procedure (1960, p. 132, fn. 2):

> Duhem is right when he says that we can test only huge and complex theoretical systems rather than isolated hypotheses; but if we test two such systems which differ in one hypothesis only, and if we can design experiments which refute the first system while leaving the second very well corroborated, then we may be on reasonably safe ground if we attribute the failure of the first system to that hypothesis in which it differs from the other.

The hypothesis testing and replication framework we propose (Systematic Replication Framework or SRF) offers such a methodological procedure that significantly reduces the ambiguity of falsification stemming from the problem of underdetermination. Although empirical underdetermination may never be eliminated, it can thereby be reduced to a sufficient degree that the scientific community can rationally converge on a verdict of falsification or high corroboration, and can demarcate theoretically justified post hoc revisions from ad hoc maneuvers with substantially increased safety.

Systematic Replications Framework

SRF consists of a systematically organized series of replications that function collectively as a single research line. The basic idea is to bring close and conceptual replications together in order to weight the effects of the AH_{pre} and AH_{out} sets on the findings. SRF starts with a close replication, which is followed by a series of conceptual replications in which the operationalization of one theoretical variable at a time is varied while keeping that of the other constant and then repeats the procedure for the other leg.

SRF starts with a close replication of an original finding. The previous arguments supporting the necessity of performing close replications (see Schmidt, 2009; Simons, 2014) are to the purpose here, as they are also relevant for underdetermination-related problems. To explicate, it is impossible to assess the corroboration of a *TH* without first a) attaining consistent results that support the reliability of particular operationalizations (and thus the reliability of *AH*s) and b) ruling out sampling or context-related *AH*s—elements which are situated at the intersection of AH_{pre} and AH_{out} sets (see Figure 2). For example, if the results of a close replication diverge from the original study, the implication might be that the corroboration of the *TH* is contingent upon a context-related *AH* (e.g., a peculiarity in the lab where the original study is conducted, or the "flair" of the researcher who conducts the study). That is, in case we obtain conflicting results between the original study and its close replications, the corroboration of the *TH* would at best be conditional on unreliable AHs—which of course is not a very good state to be in. Therefore, reformulating or abandoning the *TH* altogether on the face of such results can be considered.

However, neither success nor failure in close replications provides sufficient evidence for reaching a verdict on the corroboration of a *TH*. Because, false findings might be perfectly replicable or a true effect might elude us due to unstable findings if our tests rest on unreliable *AHs*. Therefore, 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.

As we said earlier, AH_{pre} and AH_{out} are actually sets comprising individual auxiliary assumptions. So, if either an element in 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 $AH_{pre}1$ constant) while using various different outcome variables (i.e., varying AH_{out} sets to $AH_{out}2\rightarrow n$). 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 to $AH_{pre}2\rightarrow n$), and this time the variable that was being varied in the previous step is kept constant (i.e., keeping $AH_{out}1$ constant). This procedure allows the researchers to 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).

It is important to note here that the systematic variation in *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; that is, to examine the most plausible alternative explanations that arise in relation to individual *AH* elements. So, for example, if an auxiliary assumption associated with a particular manipulation is suspected to be chiefly responsible for the previous findings (e.g., using hand held watches instead of laser sensors in a priming experiment), then the variation should be targeted at that assumption. Examining alternative explanations associated with different *AH* sets would be a useful method for selecting the riskiest falsification test and thus it would potentially provide the strongest corroboration for

TH. A visual summary of SRF can be seen in Figure 2 and a decision guide explicating how to proceed in different research scenarios can be found in the supplementary materials.

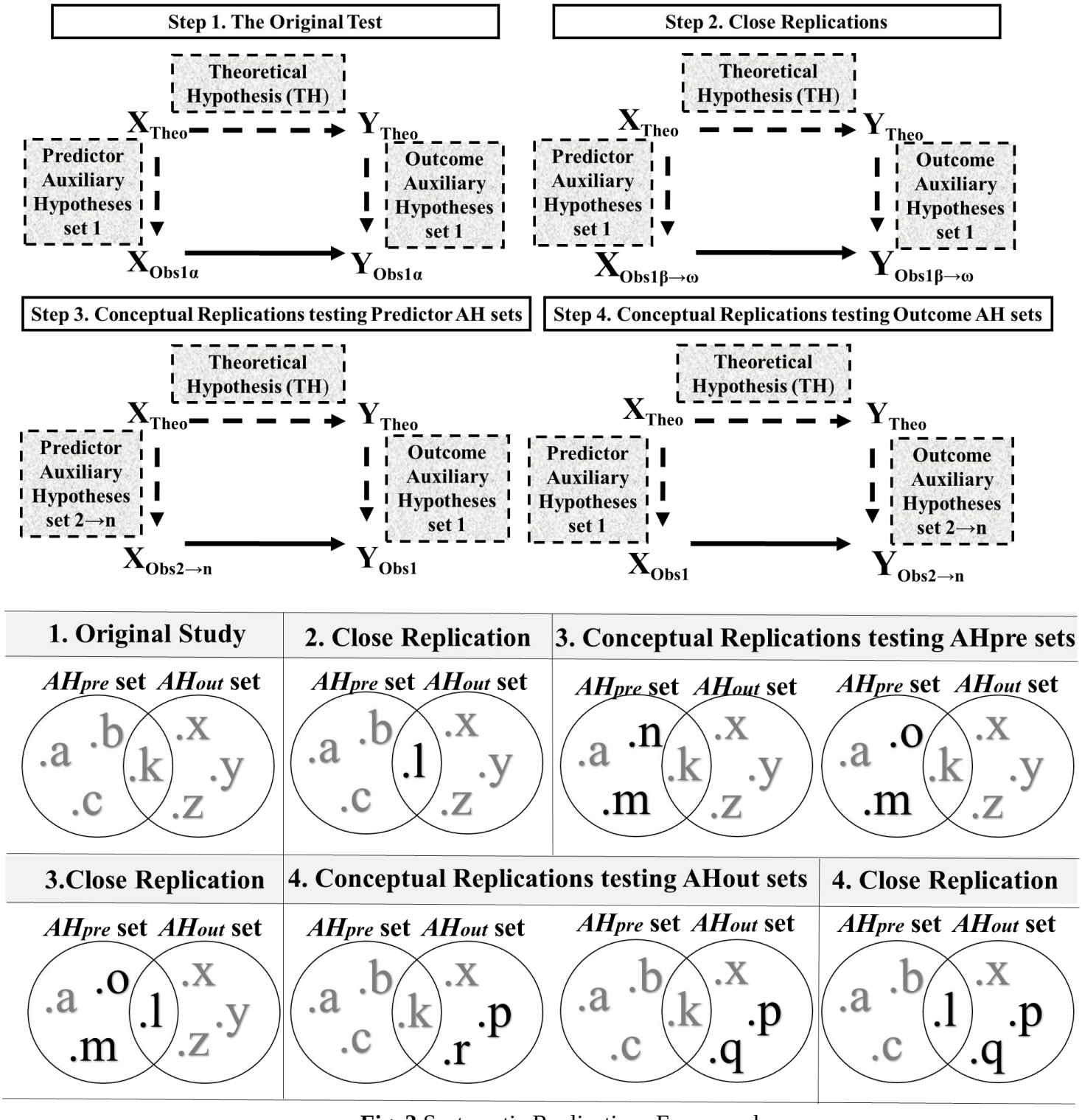

Fig. 2 Systematic Replications Framework

SRF reduces ambiguities implied by the DQT 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. Thereby these components can receive blame not collectively but in terms of a weighted distribution. In cases where it is not possible to achieve this, it 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 that we 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).

What is different in SRF?

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, 2019; see also Barr et al., 2013 on random effects).

Triangulation, another concept, also indicates the need for diversifying and connecting replications (Munafò & 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 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 (see Leahey, 1980). They seem to assume that no particular operationalization can perfectly capture the underlying concept but they can do so collectively. This is because each individual operationalization introduces some random error. 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 (i.e., non-operational) qualities, it is impossible to decide when and how different measurements can be meaningfully grouped into a concept (Vessonen, 2020).

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 & 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) and 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, following largely the sophisticated methodological falsificationism of Lakatos (1978), has a very different idea about the role we should assign to *AHs* in science. According to this view, theoretical statements lend themselves to empirical tests only with the help of *AHs,* because they connect core theoretical concepts and relationships to observations. As auxiliary assumptions, operationalizations do not substitute or collectively exhaust theoretical concepts and relationships. *AHs* 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 *AHs* 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. In progressive research programmes (consisting of successive versions of a theory), *AHs* predominantly increase empirical content by enabling novel observations 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 novel empirical discoveries or questions. Researchers may avoid falsification of the *TH*, on pain of 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). 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 *AHs* 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* plausible alternative explanations associated with different *AHs*. In this sense, what we understand from replication is quite akin to "constructive replication" of Köhler and Cortina (2019), 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 *AHs* the main hypothesis is conditional. By providing a way to accomplish this, SRF increases the transparency of how *AHs* influence "(non-)corroborating 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. Lastly, SRF shares the advantages of falsificationist frameworks: It is always more practical to try to find a falsifying instance then collecting verifying examples, even if collecting all the verifying examples is not deemed necessary because of randomization. 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.

Adversarial collaboration

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, Hertwig, & Kahneman, 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 underdetermination problem 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). Since SRF is an effective tool in addressing the problem of underdetermination, it can bring adversarial collaboration closer to

what it should be, namely a method to find solutions to contentious theoretical issues. SRF can better facilitate adversarial collaboration in hypothesis-testing also because in such a framework the parties do not need to agree on particular measures in order to collaborate: They can at least agree on conditionals and thus reach consensus in the appraisal of the outcomes of the whole scheme.

Practical Implications

As it stands, SRF can be said to have practical implications for three broad domains of scientific inquiry, namely 1) *Replication studies* via coordinating close and conceptual replications into a more coherent, informative and critical body of investigations, 2) *Hypothesis testing* via providing a severe testing framework for self-replication attempts, 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 and also in straightening up an existing body of findings into a meaningful network of relationships in a systematic literature review.

Replicating an initial finding before publication (i.e., self-replication) has long been considered among the best practice (Cesario, 2014; Roedinger 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} can mitigate the resulting ambiguity here as well.

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 testing 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 in a well-ordered way to partly circumvent the underdetermination problem; 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, Nelson, & Simonsohn, 2011). Realistically it would be a tentative plan, but it would still inform both the research team 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. 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 & Fode, 1963) influences the results is an *AH* that needs almost always to be taken seriously.

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 assumptions and measurements. Thus, it can eventually help in appraising the verisimilitude of the *TH* by revealing how it is conditional on certain *AH*s, and can lead to the 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 increased scientific collaboration and collective testing and appraisal of scientific theories. 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, replication, crucial experiments and the like (i.e. science's critical mission).[2] 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; OSC, 2015; Ioannidis, 2005) has identified the prioritization of the exploratory over the critical mission as one of the key causes, and led to proposals for slowing science down (Stengers, 2018), applying more

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

caution in giving policy advice (Ijzerman et al., 2020), and inaugurating a credibility revolution (Vazire, 2020). 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 Contributions

Conceptualization: Duygu Uygun Tunç and Mehmet Necip Tunç.

Funding Acquisition: Duygu Uygun Tunç and Mehmet Necip Tunç.

Project Administration: Duygu Uygun Tunç and Mehmet Necip Tunç.

Visualization: Duygu Uygun Tunç and Mehmet Necip Tunç.

Writing - Original Draft Preparation: Duygu Uygun Tunç and Mehmet Necip Tunç.

Writing - Review & Editing: Duygu Uygun Tunç and Mehmet Necip Tunç.

Conflict of Interest

This research is 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″, under project number 120C064.

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*, 230-244.

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*, 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*, 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*, 1-68.

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*, 1561-1580.

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

Bonett, D. (2012). Replication-Extension Studies. *Current Directions in Psychological Science, 21*, 409-412.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J., & 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. (1927). *The Logic of Modern Physics*. New York: Macmillan.

Camerer, C. F. et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2,* 637–644.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9,* 40-48.

Churchland, P. M. (1975). Karl Popper's Philosophy of Science. *Canadian Journal of Philosophy, 5,* 145–156.

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. Psychological Bulletin, 145, 610–651.

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Banaruee, H., Butcher, N., Cavallet, M., ... & Gygax, P. (2020). The Many Smiles Collaboration: A Multi-lab Test of the Facial Feedback Hypothesis. https://doi.org/10.31234/osf.io/cvpuw

Crusius, J., Gonzalez, M. F., Lange, J., & Cohen-Charash, Y. (2020). Envy: An adversarial review and comparison of two competing views. *Emotion Review, 12,* 3-21.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020, April 26). The case for formal methodology in scientific reform. https://doi.org/10.1101/2020.04.26.048306

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

Doherty, J. M., Belletier, C., Rhodes, S., Jaroslawska, A., Barrouillet, P., Camos, V., ... & Logie, R. H. (2019). Dual-task costs in working memory: An adversarial collaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45,* 1529-1568.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*, (P. W. Wiener, Trans.), Princeton, NJ: Princeton University Press. (Original work published 1908)

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

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

Ekkekakis, P. (2013). *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge: Cambridge University Press.

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*, 7561-7567.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin*, *106*, 155–160.

Hacking I. (1992). The Self-Vindication of the Laboratory Sciences. In A. Pickering (Ed.), *Science as Practice and Culture.* University of Chicago Press.

Hempel, C. G. (1945). Studies in the Logic of Confirmation I, *Mind*, *54*, 1–26.

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.

Hull, D. L. (1999). The use and abuse of Sir Karl Popper. *Biology and Philosophy*, *14*, 481-504.

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., ... & Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*, 1092-1094.

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

Qizilbash, M. (2003). Vague language and precise measurement: The case of poverty. *Journal of Economic Methodology, 10*, 41-58.

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

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes* (J. Worrall & G. Currie, Eds.). Cambridge University Press.

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... & Ly, A. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin, 146,* 451–479.

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). Re-thinking reproducibility as a criterion for research quality. http://philsci-archive.pitt.edu/id/eprint/14352

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*, 0956797620904990.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. New Jersey: Princeton University Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-9. https://doi.org/10.1037/h0026141

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,* e1.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

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,* 222-244.

McPhetres, J., Albayrak-Aydemir, N., Barbosa Mendes, A., Chow, E. C., Gonzalez-Marquez, P., Loukras, E., … Volodko, K. (2020, June 11). A decade of theory as reflected in Psychological Science (2009-2019). https://doi.org/10.31234/osf.io/hs5nx

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,* 195-244.

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

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-387.

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

Munafò, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature, 553,* 399–401.

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

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology, 114*, 657-664.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, …T. Yarkoni. (2015). Promoting an open research culture. *Science, 348*, 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*, 615-631.

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

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

Popper, K. (2002a). *Conjectures and refutations: The growth of scientific knowledge*. New York: Routledge. (Original work published in 1963)

Popper, K. (2002b). *The Logic of Scientific Discovery* (2nd edition). London, New York: Routledge. (Original work published in 1934)

Popper, K. (1974). Replies to my critics. In P. A. Schlipp (Ed.), *The philosophy of Karl Popper* (Vol. 2). La Salle, IL: Open Court.

Popper, K. (1961). *The Poverty of Historicism* (2nd edition). London: Routledge and Kegan Paul. (First publication 1957)

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

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

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90-100.

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

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

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

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76-80.

Stengers, I. (2018). *Another science is possible: A manifesto for slow science*. Cambridge: Polity Press.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768-777.

Strevens, M. 2001. The Bayesian Treatment of Auxiliary Hypotheses. *British Journal for the Philosophy of Science, 52*, 515‑537.

Stroebe, W. (2019). What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology, 41*, 91-103.

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

Tarantola, A. (2006). Popper, Bayes, and the inverse problem. *Nature Physics, 2*, 492-494.

Tetlock, P.E. (2006). *Adversarial collaboration: Least feasible when most needed? Least needed when most feasible?* Presentation to Board of Directors of Russell Sage Foundation, New York City.

Tetlock, P. E., & Mitchell, G. (2009). Adversarial collaboration aborted but our offer still stands. *Research in Organizational Behavior*, *29*, 77-79.

Vazire, S. (2020, January). Do We Want to Be Credible or Incredible?, *APS Observer*, https://www.psychologicalscience.org/observer/do-we-want-to-be-credible-or-incredible

Vessonen E. (2020, August). Respectful operationalism. *Theory & Psychology*, *31*, 84-105. doi:10.1177/0959354320945036

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917-928.

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

Yarkoni, T. (2019, November 22). The Generalizability Crisis. https://doi.org/10.31234/osf.io/jqw35