What is an informative replication?

Duygu Uygun-Tunc

University of Chicago

1. Introduction

The social sciences have been engaged for about two decades in a debate about low replication rates (Camerer et al., 2018; OSC, 2015) and how to interpret replication failures (Bryan et al., 2019). Replication studies are widely regarded as one of the benchmarks of scientific rigor (Simons, 2014), but there is vast disagreement as to what makes a study a replication of another, what function replications serve, and what kinds of replication studies there can be. In this paper I propose a general account of replication from a novel theoretical perspective on what it means for a hypothesis test to be severe. Assessing the severity of a test involves determining whether and to what extent the test is informative. This assessment relies on a range of methodological strategies scientists use to address underdetermination. Hence, I argue that the epistemic function of replication is best understood as the reduction of underdetermination surrounding hypothesis tests, and that replication increases test severity by probing specific auxiliary assumptions.

I distinguish two aspects of underdetermination — holistic and contrastive — and reconstruct each as a methodological rather than a logical problem, elaborated in Section 3. Replication studies and their informativity have chiefly to do with the problem of holistic underdetermination, so I will dwell mostly on this aspect. I reframe holistic underdetermination as a problem of managing epistemic risks arising from the misspecification of auxiliary assumptions that mediate between a theoretical hypothesis and empirical data. In scientific practice, holistic underdetermination reflects the difficulty of allocating the impact of negative results between a hypothesis and various auxiliary assumptions that are made in devising its test. This problem applies directly to the current dissensus about the success conditions of replications, because the interpretation of replication success or failure is similarly underdetermined. Reducing holistic underdetermination reduces the risks of erroneously rejecting or maintaining a hypothesis due to false auxiliary assumptions or a false ceteris paribus clause. Replications reduce epistemic risks due to holistic underdetermination across multiple tests by revealing if the corroboration or refutation of a hypothesis is conditional on certain auxiliary assumptions. From this perspective, an informative replication is a test of certain alternative explanations of the findings that are associated with certain auxiliary assumptions of the original test.[1]

[1] See also Shadish et al. (2001) for a resonant perspective with a focus on validity. They speak of "the utility of ruling out the alternative explanations that practicing scientists in a given research area believe could compromise knowledge claims."

Underdetermination provides a unified framework for defining both test severity and the epistemic function of replications. In general terms, severity can be conceived as the capacity of a test to confer empirical support to a hypothesis relative to its alternatives. What undermines this capacity is the underdetermination of the hypothesis by test results. I propose that the severity of a hypothesis test increases to the extent that we reduce the epistemic risks posed by underdetermination.

This conception improves on approaches that rely on vague similarity criteria between tests. Beyond the still standing conceptual difficulties of defining and evaluating replication success or failure, the epistemic value of replication studies is not derivative of and goes much beyond the evaluation of the reliability of original results. But the ongoing methodological debate on which kind of replication should be preferred (e.g., close or conceptual) is not productive, because we still need a comprehensive and principled analysis of the kind of information provided in each kind of subsequent test. We can both assess and maximize the epistemic value of replications by identifying their specific contribution to reducing underdetermination based on which auxiliary assumptions they can test. This approach also gives us the necessary tools to create a taxonomy of replication studies in each scientific domain based on what kind of auxiliary assumptions they test.

The first two sections introduce severity and underdetermination. In the third section I introduce my account of severity formulated in terms of what I call the methodological problem of underdetermination. Next, I introduce my account of replication as an alternative method to increase severity, namely across multiple tests. In the last section, I go into how this function is realized in different kinds of replication studies categorized with regard to the kind of auxiliary assumptions they test.

2. The Two Tasks of Severity Assessment

According to Popper, the severity of a test is its ability to falsify a theory.[2] Popper then uses the concept of severity to define corroboration: the degree of severity of the tests which a theory has passed gives us its degree of corroboration. Popper's conception of severity is one-sided and, for this reason, rather inadequate. First and foremost, Popper does not acknowledge that the ability of a test to falsify a hypothesis can be understood in two different ways: It might mean that the test is a proper test of the hypothesis, or it might mean that the test is a difficult one for the hypothesis in question to pass so that it will most probably fail it. In other words, severity-assessment might address two different questions:

(i) *whether* a test is able to lend *any* empirical support to a hypothesis, and

(ii) *how much extra* empirical support it lends to a hypothesis vis-à-vis alternatives.

Popper's formulation of severity addresses only the second and ignores the first question, which is implicitly presupposed.

[2] More formally, severity of a test is the "ratio of the probability of observing evidence (e) given the hypothesis (H) and background knowledge (B), to observing evidence (e) given only the background knowledge (B)" (Claesen et al., 2022).

Mayo applied the concept of severity to statistical inference, where she acknowledged and unified its two functions explicated above. Mayo (2018) defines severity as a measure of how a test prevents erroneous interpretations of the results. According to her error statistical account, a test will not lend any empirical support to a hypothesis if it does not reduce certain frequentist error probabilities below a threshold of acceptability. Assessing this in turn can give a measure of the relative empirical support of alternative theories if we compare the error rates of the respective tests they passed. While Mayo's criterion helps determine whether a hypothesis has been severely tested, it does not offer much guidance regarding how to achieve severity. This is because severity is a holistic quality that emerges out of a combination of several different properties of a hypothesis-test that reduce underdetermination, which Mayo overlooks in focusing only on the indicators of severity. Most importantly, the error-statistical concept of test severity cannot realize what Popper originally intended; namely, evaluating the degree of corroboration of theoretical or substantive hypotheses.[3] This is because statistical test severity is only a component of the severity of a hypothesis test. We cannot evaluate the informativity of hypothesis tests beyond the statistical level without a more general conception of severity.[4]

This general perspective can be formulated in terms of the *epistemic risks due to underdetermination of hypotheses by data* — namely, the epistemic risks pertaining to erroneously abandoning or maintaining a hypothesis, where the source of error is conceptualized as the misspecification or falsity of one or more of the various kinds of auxiliary assumptions that mediate between a theoretical hypothesis and data.[5] Finally, reinterpreting severity in terms of epistemic risks due to underdetermination requires that we understand underdetermination as a gradational concept, which grounds the degree of (relevant kind of) risk in the degrees of (relevant kind of) underdetermination. Degrees of underdetermination will, in abstract terms, be a function of the number and importance of the auxiliary assumptions that present uneliminated alternative explanations of the data patterns.

Mayo's account insightfully emphasizes that severe testing proceeds through the piecemeal identification and probing of potential sources of error. Scientists draw on experimental knowledge to design tests capable of detecting ways in which the hypothesis might be false. However, the formal definition of severity in the error-statistical framework applies most directly to statistical hypotheses, where the relevant error probabilities can be explicitly modeled. Many hypothesis tests involve qualitatively heterogeneous sources of error that arise at multiple levels of modelling, including theoretical assumptions linking constructs to observable variables, experimental assumptions governing measurement and design, and statistical assumptions concerning sampling and inference. These sources of error often cannot be captured within a single probabilistic model. What unifies them is not their statistical form but their role as alternative explanations of the

---

[3] For an attempt to apply the error-statistical framework to large-scale theory testing, see Mayo (2009). While there are some conceptual commonalities between this attempt and the framework offered in the present paper, the aim here is to give the concept of severity a single, unified interpretation that can be applied to theoretical conjectures directly.
[4] There are also Bayesian approaches to statistical test severity. These emphasize increasing the precision of predictions and the amount of information conferred by the evidence in terms of the change in posterior probabilities and likelihood ratios of hypotheses. An adequate comparative discussion of these approaches is beyond the scope of this paper.
[5] This epistemic risk framework is clearly distinct from the inductive risk framework that deals with the non-epistemic consequences of scientific error. That being said, there is no in-principle barrier to adopting a similar approach in estimating inductive risks in policy-related contexts.

observed data patterns. The framework proposed here makes this structure explicit: the severity of a test increases to the extent that it reduces the degrees of underdetermination surrounding a hypothesis by probing such alternative explanations.

3. Underdetermination as a Logical vs. Methodological Problem

The problem of underdetermination refers to the impossibility of conclusively refuting or verifying hypotheses on the basis of empirical evidence. It has two distinct aspects: holistic and contrastive underdetermination (Stanford, 2017). I unpack each in turn.

Theories do not logically imply any testable predictions, as 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, scientific theories or hypotheses have empirical consequences only in conjunction with auxiliary assumptions that help bridging theoretical terms to their empirical instances. These auxiliary assumptions may concern 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 (or constructs) linked in the main hypothesis, to the implications of previous theories and the *ceteris paribus* clause (i.e., all other things being equal). In this regard, in every isolated empirical test, we actually pose several largely independent questions. Holistic underdetermination (Duhem, 1954; Quine, 1951; Ivanova, 2021) refers to the ambiguity of falsification: Given that we must make various auxiliary assumptions in formulating and testing hypotheses, no test can allocate blame uniquely to the hypothesis under test. By extension, the genuineness of confirming evidence can also be called into question, because the results can also be due to statistical, methodological, or instrumental artifacts. This finds an example in the dissensus about the success conditions and the value of replications.

Contrastive underdetermination (Laudan, 1990; Van Fraassen, 1980) refers to the inability of a test to uniquely support a hypothesis: Given that no hypothesis can be *entailed* by the evidence, the results can support multiple alternative hypotheses at once. Replication studies and their informativity have chiefly to do with the problem of holistic underdetermination. But decreasing holistic underdetermination is a first step in decreasing contrastive underdetermination, thus the two problems are closely related. In this regard, replications also contribute (although more indirectly) to managing epistemic risks due to contrastive underdetermination.

Strictly logical versions of the underdetermination thesis are intractable, because it is impossible to explicitly state and test all, potentially infinite, auxiliary assumptions made in hypothesis-formulation and test-design. But scientists have not only logical but also methodological tools for increasing the informational value of tests (see also Laudan, 1990). We can thus reconstruct the logical problem of underdetermination as a methodological problem. The methodological problem of holistic underdetermination is that of allocating the impact of negative results between a hypothesis and various auxiliary assumptions that are made in devising its test. The methodological problem of contrastive underdetermination is that of identifying rational criteria for preference between a given number of rival hypotheses that are compatible with the same evidence (see also Laudan & Leplin, 1991).

4. Severity as Management of Epistemic Risks Due to Underdetermination

As a methodological problem, underdetermination can be approached in terms of the management of various epistemic risks. Reducing holistic underdetermination reduces the risks of erroneously rejecting or maintaining a hypothesis due to false auxiliary assumptions or a false ceteris paribus clause. Managing both kinds of epistemic risks maximizes the informativity of hypothesis tests: When a hypothesis test is informative, both corroborative and non-corroborative evidence teaches us important things about the underlying assumptions and alternative hypotheses. I propose that severity increases to the extent that we reduce the epistemic risks posed by holistic and contrastive underdetermination. I call these holistic and contrastive severity.

The two considerations give us different qualifications. A hypothesis is well-tested when it is sufficiently articulated to be subjected to empirical scrutiny within a sufficiently valid testing framework, such that empirical results genuinely count for or against it, thereby mitigating holistic underdetermination. By contrast, a hypothesis is best-tested relative to a set of relevant alternatives, when the available data are capable, in principle, of discriminating among them. In such cases, rival conjectures are specified with respect to the same data model yet yield different predictions, allowing the evidence to favor one conjecture over its competitors and thereby mitigating contrastive underdetermination. Theories are complexes, so they may be well-tested regarding some of their implications while remaining untested regarding some others (Mayo, 2009).

Since replications have chiefly to do with holistic underdetermination, they can be seen at bottom as a methodological tool to increase holistic severity. In this form the methodological problem of holistic underdetermination offers a fruitful framework for reconsidering one of the two goals of severity-assessment: Assessing *to what extent* a test is able to confer *any* empirical support to a hypothesis is tantamount to the methodological problem of holistic underdetermination.

We should delineate how the methodological problem of underdetermination is distinct from the logical one. Firstly, among the plethora of different auxiliary assumptions existing in a hypothesis test only a certain subgroup of assumptions can be expected to meaningfully impact the results. There are infinitely many other auxiliary assumptions 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. Auxiliary assumptions that are thought to belong to 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 auxiliary assumptions — such as measurement reliability, construct validity, 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 the main hypothesis. The design elements (including the statistical analysis strategy) featured in a well-written methods section of a scientific paper can also be thought of as specifying the auxiliary assumptions associated with the test.

Holistic severity has to do with reducing epistemic risks due to holistic underdetermination (see also Oude Maatman, 2021). These risks increase to the extent that (i) the scientific model is misspecified (Uygun Tunç & Tunç, 2022) — i.e., one or more auxiliary assumptions are erroneously relegated to the ceteris paribus clause or the test features false auxiliary assumptions — and this misspecification influenced the test results. In this framework the problem of holistic underdetermination can be conceived as a (mis)specification problem regarding the scientific model that is being tested. In this respect, holistic severity is a function of how difficult it is to provide alternative explanations of the test results in reference to false auxiliary assumptions (Oude Maatman, 2021). The core function of replication studies is to investigate the possibility of misspecification in various ways.

We should note that the possibility of misspecification cannot be eliminated entirely. What methodological strategies can achieve instead is a gradual reduction of the range of alternative explanations compatible with the evidence. By designing tests that probe different components of the chain connecting theory to data — particularly those auxiliary assumptions that mediate between theoretical claims and observable variables — researchers can progressively narrow the scope of underdetermination. In this sense, improving the specification and testing of these mediating assumptions increases the severity of hypothesis tests by making it more difficult to explain the results in terms of alternative sources of error.

The distinction between holistic and contrastive severity also clarifies the epistemic role of replication studies. Replications primarily address holistic underdetermination by probing auxiliary assumptions that might generate the observed results independently of the hypothesis under test. In doing so they help determine whether empirical results genuinely bear on the hypothesis rather than on artifacts of measurement, design, or analysis. Only once this ambiguity has been sufficiently reduced can empirical tests meaningfully discriminate between rival hypotheses. In this sense, increasing holistic severity is a precondition for achieving contrastive severity. Replication therefore contributes indirectly to theory testing: by reducing the range of auxiliary explanations compatible with the evidence, it prepares the ground on which theoretical alternatives can be compared.

5. A New Taxonomic Principle for Replications

The epistemic risks associated with misspecification diminish as auxiliary assumptions are made explicit, independently examined, and effectively controlled within the design and execution of tests. Increasing test severity therefore requires not only more precise theoretical articulation but also improved understanding of the phenomenon under investigation and greater control over the testing environment. In the idealized case of an *experimentum crucis*, all relevant auxiliary assumptions would be explicitly specified and independently verified. In practice, however, such conditions are often unattainable. Scientific background knowledge is continually revised, and theories themselves evolve through the identification and correction of previously unnoticed assumptions. As a result, the possibility of misspecification can never be entirely eliminated.

These difficulties are particularly pronounced in the social and behavioral sciences. Theories in these disciplines typically underconstrain their own measurement: They do not specify how theoretical constructs should be operationalized, which leaves researchers with considerable latitude in choosing measures and little principled basis for preferring one over another (Folger, 1989; Meehl, 1978). This

problem is compounded by the fact that theoretical terms are often semantically open — vague, value-laden, or resistant to formalization (Green, 2019; Weber, 2017) — so that no particular operationalization clearly stands out as more faithful to the construct than its competitors. Each research tradition tends to develop its own measurement devices guided by its own theoretical commitments, which means that the auxiliary assumptions governing operationalization are rarely tested independently of the very theory they are meant to serve (MacCorquodale & Meehl, 1948; Eronen & Romeijn, 2020; Muthukrishna & Henrich, 2019). Researchers also exercise limited control over the environments in which social phenomena occur, and ambient factors generate background correlations that can masquerade as theoretically meaningful effects (Meehl, 1990; Orben & Lakens, 2020). The practical consequence of all this is not merely that replication is difficult, but that the choice of basic design elements substantially shapes the effects that are observed: When different research teams test the same theoretical hypothesis using their own preferred operationalizations, they regularly produce discrepant results that cannot be adjudicated by further replication alone (Landy et al., 2020). It is precisely under these conditions that probing the conditionality of findings with respect to specific auxiliary assumptions becomes indispensable rather than optional.

Given these conditions, single experiments should not be treated as the primary unit of hypothesis testing. Replication studies offer a more effective means of increasing severity across multiple tests. Historically, replications have often served precisely this function by addressing different aspects of the underdetermination problem. Close replications typically probe auxiliary assumptions related to measurement reliability or sampling error, while conceptual replications examine assumptions associated with operationalization or experimental context. However, replication results themselves remain ambiguous when they are not designed to test specific alternative explanations tied to particular auxiliary assumptions.

The key question, from the perspective developed here, is not whether a replication is close or conceptual but which auxiliary assumptions it is designed to test and what diagnostic information it can therefore yield. The so-called close replications are meant to test auxiliary assumptions that do not explicitly appear in the inference linking the hypothesis to data, i.e., those that have been implicitly relegated to the ceteris paribus clause on the assumption that they are inconsequential. These include assumptions about hidden moderators, cultural or contextual dependencies, researcher-specific factors such as procedural expertise, among others. Because such assumptions are not stated in the scientific model, their potential influence on the results is not acknowledged and therefore not controlled. A replication that reproduces the test under conditions that allow these background factors to vary can reveal whether the original result was conditional on them, that is, whether the scientific model was misspecified by excluding them from explicit consideration. A second, related function is to test the statistical auxiliary assumptions governing the original result: whether it reflects a genuine pattern in the target population or is an artifact of sampling variability. Replications that reproduce the test in independent samples with adequate power provide evidence bearing specifically on this class of background assumptions. Both functions share a common rationale, namely they probe assumptions that were treated as unproblematic in the original test, and conflicting results between the original study and such replications open a diagnostic space. The discrepancy may indicate Type II error in the replication, misspecification of the original model with respect to background assumptions, or Type I error in the original finding.

The replication failures surrounding the ego depletion effect illustrate what this diagnostic space looks like in practice (Hagger et al., 2016). When large-scale pre-registered replications failed to reproduce the effect originally reported by Baumeister and colleagues, the three-branch diagnostic remained genuinely open: the failures might reflect insufficient statistical power in some of the original studies, or they might indicate that the original scientific model was misspecified with respect to the auxiliary assumptions governing how depletion is operationalized and measured, or the original findings might have been false positives inflated by publication bias and flexible analytic choices.

Close replications give information on the stability of the scientific model as a whole. Whether they are successful or not, they cannot inform about the corroboration of the main hypothesis. Specifically, they cannot establish whether the outcome reflects the hypothesized relationship itself or some combination of that relationship with the explicit auxiliary assumptions built into the scientific model (the operationalizations, measurement instruments, and design choices that are stated in the methods and directly constitutive of how the hypothesis is tested). When a replication succeeds or fails, and the ceteris paribus assumption is sufficiently valid,[6] the remaining ambiguity concerns this second layer: corroboration or non-corroboration might be genuine, or it might be an artifact of the particular data-generating mechanism. Neither outcome resolves this. If the validity of particular auxiliary assumptions are themselves questionable, obtaining similar results in a subsequent test teaches us nothing, because both studies may be relying on the same problematic assumptions and converging for that reason rather than because the hypothesis is well-supported. A historical illustration of this risk is the case of N-rays. In 1903, the French physicist René Blondlot reported detecting a novel form of radiation, and the finding was subsequently replicated by numerous researchers within the French scientific community. The convergence was illusory: all replicators shared the same false auxiliary assumption about the reliability of the detection method, which relied on observers judging the brightness of a dimly lit thread by eye under conditions that made systematic self-suggestion nearly inevitable. The artifactual nature of the effect became apparent only when physicist Robert Wood covertly removed the prism that was supposed to be the radiation source during a public demonstration, while observers continued to report detecting N-rays uninterrupted. No accumulation of convergent results under the same conditions could have supplied this information. To make progress in this way, replications must be designed to determine whether the corroboration of a hypothesis is conditional on particular explicitly specified auxiliary assumptions.

This is not what so-called conceptual replications typically accomplish, despite their surface resemblance to what is being described here. Conceptual replications introduce methodological novelty, but rarely with the deliberate aim of testing whether a specific auxiliary assumption is necessary for the result to obtain. When they fail, discrepancy can be attributed to any number of differences between the original and the replication; when they succeed, nothing specific has been learned about the conditional structure of the hypothesis. The informational yield is low in both cases precisely because the auxiliary assumption being varied is not identified in advance as the target of the test.

---

[6] As Meehl (1990) put it, this roughly means that there are no *systematic* factors left unspecified.

Progress requires instead that each such replication be designed with a specific dependency relationship in mind: does the hypothesized effect hold when a particular operationalization is replaced, when a different population is sampled, when the experimental context is substantially altered? Discovering that corroboration is conditional on a particular auxiliary assumption is epistemically significant in two directions. It may delimit the practical and theoretical reach of the hypothesis — a finding that holds only under a narrow set of operationalizations or in a specific cultural context carries far less evidential weight than one that is robust across them. It may also, in stronger cases, constitute evidence against the generalizability claim implicit in the hypothesis, if the boundary conditions within which the effect holds turn out to be substantially more restricted than originally implied. Conversely, when a hypothesized relationship proves stable across systematic variation of its constitutive auxiliary assumptions, confidence in that relationship is correspondingly strengthened — not merely because the result has been reproduced, but because the conditionality structure of the hypothesis has been actively probed and the dependencies found to be limited.

Because numerous auxiliary assumptions are involved in any empirical test, the epistemic value of replication typically emerges cumulatively across multiple studies rather than from any single replication. Replication studies therefore function most effectively when organized in ways that systematically explore different auxiliary assumptions. Hence, replication studies become epistemically informative when they are designed to probe specific sources of underdetermination and thereby increase the severity of hypothesis tests.

6. Conclusion

This paper proposed a functional account of replication grounded in the concepts of underdetermination and test severity. The central claim is that a replication is informative to the extent that it reduces the degrees of underdetermination surrounding a hypothesis test by probing specific auxiliary assumptions and not to the extent that it reproduces or fails to reproduce a prior result.

This account has two implications worth underlining. First, it dissolves the debate over which kind of replication should be preferred. Close and conceptual replications are not competing strategies of unequal value; they are complementary probes of different classes of auxiliary assumptions, and their relative priority depends on which sources of underdetermination most threaten the evidential force of the hypothesis under investigation in a given research context. Second, the account shifts the locus of epistemic responsibility in replication practice. The question a replication researcher must answer before designing a study is not "how similar should this study be to the original?" but "which auxiliary assumption is the most pressing uneliminated alternative explanation of the original result, and how can this study be designed to test it?" Answering that question requires a level of theoretical and methodological analysis that is largely absent from current replication practice but that the framework developed here is intended to support.

More broadly, understanding replication through its epistemic function reframes what scientific progress through replication looks like. Progress is not measured by the rate at which results reproduce but by the degree to which successive studies reduce the range of alternative explanations compatible with the evidence. On this account, a well-designed replication that produces a null

result can be more informative than a successful one that leaves all auxiliary assumptions unexamined.

References

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

Bryan, C. J., et al. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *PNAS*, *116*, 25535–25545.

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*(9), 637–644.

Claesen, A., Lakens, D., Vanpaemel, W., & van Dongen, N. (2022). Severity and crises in science: Are we getting it right when we're right and wrong when we're wrong? https://doi.org/10.31234/osf.io/ekhc8

Duhem, P. (1954). *The aim and structure of physical theory* (P. W. Wiener, Trans., 2nd ed.). Princeton University Press.

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

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

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

Hagger, M. S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Ivanova, M. (2021). *Duhem and Holism*. Cambridge University Press.

Landy, J. F., et al. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*, 451–479.

Laudan, L. (1990). Demystifying underdetermination. In C. W. Savage (Ed.), *Scientific theories* (pp. 267–297). University of Minnesota Press.

Laudan, L., & Leplin, J. (1991). Empirical equivalence and underdetermination. *Journal of Philosophy*, *88*, 449–472.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*, 95–107.

Mayo, D. G. (2009). Learning from error, severe testing, and the growth of theoretical knowledge. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 28–57). Cambridge University Press.

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*, 195–244.

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

Open Science Collaboration. (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*, 238–247.

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

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

Quine, W. V. O. (1951). Two dogmas of empiricism. Reprinted in *From a logical point of view* (2nd ed., pp. 20–46). Harvard University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

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

Stanford, K. (2017). Underdetermination of scientific theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017 ed.). https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

Uygun Tunç, D. & Tunç, M. N. (2023). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. *Meta-Psychology, 7*. https://doi.org/10.15626/MP.2021.2756

Van Fraassen, B. (1980). *The scientific image*. Oxford University Press.

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

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393–472.