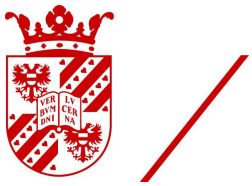

rijksuniversiteit groningen

Sophisticated Falsification in Psychology:

The Case of Automatic Social Behavior

Jacopo Zuppa

Master Thesis - Theory and History of Psychology

S3221148
July 2023
Department of Psychology
University of Groningen
Examiner/Daily supervisor:
dr. Maarten Derksen

Abstract

Psychological science is currently in a state of crisis due to the field's inability to replicate findings. Amidst this crisis, a fundamental issue is investigating what are the epistemological implications of failures to replicate previous findings. One suggestion is that replication failures could be interpreted as an opportunity to apply *sophisticated falsification*, a central notion in Lakatos' (1978) theory of scientific research programmes. In the present thesis, I discuss at length this idea as it applies to the research programme initiated by Bargh and colleagues (1996) on the phenomenon of automatic social behavior (ASB). Through the works of Lakatos (1978) and Meehl (1990), I demonstrate how sophisticated falsification cannot be applied to cases when experiments on ASB fail due to the deficient epistemological standards of the programme. Sophisticated falsification requires theoretical models to be well-corroborated by evidence. Presently, theories on ASB only make vague predictions in the form of directional estimates tested using null-hypothesis significance testing. Even if confirmed, these predictions are not corroborative and do not constitute good reasons to believe that a certain theory has high verisimilitude. As such, sophisticated falsification cannot be applied when experiments on ASB effects fail, and the corresponding research programme is most likely in a state of degeneration. In light of these considerations, I discuss how formal modeling could be a potential solution to improve the corroborative value of the predictions of psychological theories.

Introduction

A crisis of confidence in psychology

Over the last decade, the epistemic state of psychological science has been seriously called into question. What came under severe scrutiny was the ability of the field to generate reliable and meaningful findings due to the failure to replicate several well-known effects and phenomena.

In 2015, the research group Open Science Collaboration found evidence that a high proportion of studies published in major psychology journals could not be replicated according to a number of statistical criteria. Other large-scale replication projects since then found mixed results for the replicability rates of specific psychological effects, with some showing alarmingly low replication rates and effect sizes (Eerland et al., 2016; Hagger et al., 2016; Wagenmakers et al., 2016). These findings seem to have generated what has been dubbed a "crisis of confidence" in psychological science due to the central role that replication plays in scientific research (Zwaan, 2018).

The importance given to replication in science is rooted in the epistemological literature of the past and the present, and it is intimately connected to the idea of scientific objectivity. The importance of replication for science is generally explained in connection with the work of Karl Popper on scientific discovery. In discussing Kant's notion of objectivity, Popper (1959) noted how the systematic reoccurrence and resulting inter-subjective testability of events are two fundamental preconditions for producing truly scientific and objective theories. The idea that a certain experimental observation can be *repeated* by anyone following a specified procedure is what, according to Popper, increases our confidence that such an observation is not a "mere isolated coincidence" (p. 23). The idea of "mere coincidence" seems fitting in the current

research context, given that the role assigned to certain types of replication is that of revealing whether or not original findings were a random occurrence (Schmidt, 2009).

Modern scholars generally agree on the importance of experimental replication in the production of scientific knowledge. Schmidt (2009) describes it as a "methodological tool based on a repeated procedure that is involved *in the establishment of a fact, truth, or piece of knowledge*" [emphasis added] (p. 91). According to Schmidt, replication is crucial in establishing scientific knowledge in that it provides evidence for the stability of a relationship between objects. Put simply, replicating an experiment shows that a certain relationship exists in a stable and predictable manner, and this, in principle, allows scientists to formulate theories and laws describing such a relationship[1].

More specifically, different types of replications have different functions in scientific research. Replications that are sufficiently similar to the original experiment (generally referred to as "direct" or "close") are fundamental to ensure *a priori* that science can accumulate valid knowledge; the repeated observation of an experimental phenomenon is what ensures its basic existence, and as such, it is what allows to generate a consistent, cumulative body of knowledge about that phenomenon (Lebel and colleagues, 2017). On the other hand, experimental replications that deliberately introduce substantial variations to the research design[2] are important to the process of theory-building and to ensure the generalizability of research findings (Schmidt, 2009; Crandall & Sherman, 2016).

---

[1] In particular, Schmidt connects the epistemological value of replication with Hume's "Principle of uniformity of nature" as described by Dilworth (1996) in his review of the metaphysics of science. The principle essentially states that natural change occurs according to certain rules and it is therefore, to some extent, deterministically predictable. This very principle can be understood in terms of invariance, such that a relationship between two objects is assumed to remain relatively stable over time. As noted in the text, this allows us to generate reliable knowledge by formulating laws and theories describing such stable relationships, and by testing these theories through repeated experiments.

[2] Replications generally referred to as "conceptual" replications from Schmidt (2009).

In light of these considerations, a widespread inability to replicate research findings in direct replications is considered alarming in scientific fields. One of the primary reasons why the current low replication rates upset the psychology community is the idea that widespread replication failures might be evidence of a high prevalence of statistical type I errors. Given that the function of replication is that of providing evidence for the existence of a stable phenomenon, the inability to replicate an experiment could indicate that previous findings were, in fact, only a random occurrence, casting understandable doubt upon the credibility of knowledge in psychological science.

This interpretation of a widespread inability to replicate findings is motivated by what we could define as "methodology-based" explanations of the replication crisis. These explanations generally look at how research practices can negatively impact the reliability of findings, something then reflected in low replication rates. In a paper now considered a classic in the field of metascience, Simmons and colleagues (2011) conceptualized the notion of Researcher's Degrees of Freedom (RDoF) to show how a researcher's arbitrary decisions in data analysis and reporting greatly increase the likelihood of type I errors. Soon thereafter, John and colleagues (2012) found evidence that an alarming proportion of researchers in the field admitted to engaging in so-called Questionable Research Practices (QRPs), a series of practices closely related to the RDoF mentioned above. Just as with the employment of RDoF, the frequent engagement in QRPs by researchers can translate into high proportions of type I errors (John et al., 2012). The widespread adoption of QRPs is understood to be the result of a dysfunctional incentive system for doing scientific work. In particular, scientists work in a context with high pressure to publish, but where only novel and positive findings have a chance at ending up in scientific journals. This arguably creates a situation where researchers make decisions in the

research process according to what increases their chance of getting published, often at the expense of the accuracy and validity of findings (Pashler & Wagenmakers, 2012).

Methodology-based explanations can be contrasted with what we could generally refer to as "theory-based" explanations of the inability to replicate experimental findings. Some of these explanations include the historical contingency of certain psychological phenomena (Gergen, 1973), low a priori probability for a hypothesis due to deficient theorizing (Fiedler, 2017), or the fact that high theoretical flexibility results in psychological theories only weakly predicting specific empirical observations (Oberauer & Lewandosky, 2019; Szollosi & Donkin, 2021).

One research area particularly affected by this crisis is social priming research. Social priming can be loosely defined as the unconscious influence that the activation of social representations has on behavior[3] (Bargh, 2006). One prominent example of social priming is the series of studies by Bargh and colleagues (1996), where they investigated what they referred to as *automatic social behavior* (ASB). This type of behavior supposedly occurs whenever the mental representation of a social behavior is made more accessible by exposing a person to a stimulus related to a specific social category (Bargh et al., 1996). According to the original account, the activation of a specific behavioral representation has a direct influence on a person's actions by increasing the likelihood of observing behavior congruent with the activated representation.

This series of studies found evidence that participants primed with words related to the category "elderly" were more likely to walk slowly after being primed, supposedly due to an increased accessibility of the behavioral representation for "slowness". A second, well-known finding was that Caucasian students subliminally exposed to pictures of African American men

[3] Notably, the exact definition of this research area is still disputed, and the differences between "social" and "non-social" priming are not necessarily clear (Bargh, 2021).

were more likely to behave with hostility in a subsequent situation compared to students exposed to pictures of Caucasian men. The elderly stereotype activation study in particular could not be replicated in a subsequent study (Doyen et al., 2012). This failed replication study caused much turmoil and was arguably one notable event in generating skepticism around the validity of existing research (Yong, 2012).

In light of the common failures to replicate findings in social priming research and in psychology at large, it becomes particularly relevant to understand what exactly are the epistemological implications of such failures. In other words, what are the implications of widespread failures to replicate for the validity of existing knowledge? And how should we best approach replication failures?

In the current thesis, I will offer insights into these questions by investigating how philosopher of science Imre Lakatos' (1978) epistemological theory can be applied in the context of the replication crisis. In particular, Lakatosian theory could be useful in shedding light on the current epistemic state of certain research programmes, and in suggesting the conditions where replication failures could be interpreted as opportunities to apply *sophisticated falsification.*

Sophisticated falsification essentially reflects the idea that failing to confirm a theory through experimentation does not necessarily mean that the theory should be falsified. An admissible strategy in such a situation entails modifying the theory to incorporate the falsifying evidence and then re-testing it in a subsequent experiment. Hence the idea of a "sophisticated" version of falsification that does not require an immediate rejection of the theory upon finding disconfirming evidence.

In the analysis, I will primarily focus on experiments on automatic social behavior. This decision is motivated by two reasons. Firstly, I maintain that focusing on this specific effect is

particularly relevant in light of recent failures to replicate it, and generally on widespread failures to replicate effects in social priming research[4]. Secondly, the idea of applying sophisticated falsification to cases where replications of ASB effects fail is consistent with Cesario's (2014) proposal that interpreting such failures as evidence of type I errors might be premature in light of our limited theoretical understanding of these effects.

Although the idea of treating failures to replicate as an opportunity to apply sophisticated falsification seems appealing, I argue that in the case of research programmes like the one investigating ASB effects, this would be a misguided attempt. The fundamental reason is that reformulating a theoretical model by sophisticated falsification necessarily requires such a model to be adequately corroborated by empirical evidence. As it presently stands, the research programme studying ASB effects is not.

In what follows, I will explore this idea at length. I will first review the possible implications and meaning of replication failures in psychological research. Then, I will give an overview of Lakatosian theory and how sophisticated falsification in psychology might work in principle through the work of Paul Meehl (1990). By following Meehl's criticism of null-hypothesis significance testing (NHST), I will then show how it is not possible to apply sophisticated falsification in research on ASB effects due to the deficient epistemic state of the research programme.

What happens when replications fail?

As discussed in the previous section, successfully replicating research findings through experiments sufficiently similar to the original is what allows us to test the basic existence of the phenomena we study, and generally, to produce reliable knowledge. Conversely, whenever replications fail, some authors argue that we are instead justified in decreasing our confidence in

[4]See Lebel et al. (2017) p. 257 for a comprehensive summary of replication failures in social priming.

the veracity of the original effects and that such failures could be indicators of type I errors (Schmidt, 2009; Pashler & Harris, 2012; Nosek & Errington, 2020).

The idea of so-called "direct" replication failures decreasing our confidence in previous findings comes from the notion that by definition, direct replications are thought to replicate the *critical features* necessary for the production of an effect (Zwaan et al., 2018; Nosek & Errington, 2020). Just as a successful experimental procedure where all important features are recreated can increase our confidence in the existence of an effect, a failure in this experimental procedure can instead decrease it. Importantly, the description of these critical features is based on the current theoretical understanding of this effect, and as such, direct replications can be viewed as a "theoretical commitment" (Nosek & Errington, 2020) reflecting current beliefs about the necessary conditions for producing it.

On the other hand, in direct replications, several undeclared factors are allowed to vary. These are factors that are generally assumed to be unimportant for the production of the to-be-replicated effect and are relegated to what is known as the *ceteris paribus clause*[5]. In the case of automatic social behavior studies, examples of such factors could be the specific time of the day, the laboratory where the experiment takes place, or the gender of the experimenter.

In light of this idea of replications as recreating the critical features to produce an effect, it seems prima facie justified to decrease our confidence in its existence whenever a replication fails. Yet, some authors pointed out how single failed direct replications cannot be straightforwardly interpreted as indicating that previous findings were due to a random occurrence. One such criticism from Cesario (2014) rests on the idea that our current theoretical

[5] The ceteris paribus refers to a theoretical clause that is part of the deductive model of an empirical test (Meehl, 1990). It is a crucial auxiliary clause to the substantive theory under investigation that states that all perturbing factors are considered to be equal between the conditions of an experiment. As such, the empirical test is generally considered to be free of confounding elements. This technically allows us to isolate the factors under study and make inferences and generalizations about their causal relationship.

understanding of certain psychological phenomena might not be advanced enough to identify *all* the critical features necessary for producing such phenomena.

Cesario's argument specifically refers to effects studied in social priming research like that of Bargh and colleagues (1996), where it is likely that several, *unknown* environmental features play a role in the occurrence of priming effects. As such, failed direct replications of these effects that have not considered these unknown, "hidden moderators" cannot be interpreted as conclusive indications that previous findings were only a random occurrence, or as denying the existence or relevance of these effects (Cesario, 2014).

Let's take as an example Bargh and colleagues' (1996) third study on hostility priming. A direct replication of this study would be based on the theoretical understanding that motivated the formulation of the procedure in the original experiment. As such, certain critical features understood to be important are recreated in the replication, such as using the same pictures of African American men or staging the same hostility-provoking situation. On the other hand, several experimental features thought to be inconsequential are allowed to vary, such as the structural features of the environment where the experiment takes place. Upon failing to replicate the experiment, one then wonders whether any of these supposedly inconsequential features were instead important to the production of the effect, such that, for example, the structural features of the environment (e.g. the room's size) actually mattered in priming a hostile response. If that is the case, a failure to replicate could be due to the differences in the physical environment of the original compared to the replication. This latter hypothesis was successfully proposed and tested by Cesario and colleagues (2010), who found evidence of the moderating role of the environmental context in priming hostility.

Cesario's argument can be better understood when considering Uygun Tunç & Tunç's (2022) discussion of the scenarios we are confronted with when direct replications fail. The authors argue that non-corroborative findings in direct replications could be indicative of either 1) A type I error in the original experiment, 2) A type II error[6] in the replication, or 3) A misspecification of the auxiliary hypotheses of the theoretical model. In the third case, a *non-trivial auxiliary clause* was incorrectly relegated to the ceteris paribus clause upon conducting the replication. In other words, a relevant factor thought to be inconsequential was not included in the theoretical model. As such, the theory-derived experimental procedure to replicate the effect was missing a key component.

Cesario's (2014) idea that our theoretical understanding of a psychological phenomenon is not advanced enough to correctly reproduce it is intimately connected to the third possibility. In such cases, a certain moderating factor that was thought to be inconsequential might not have been considered among the critical features defining the replication procedure. Importantly, whenever we cannot replicate a certain effect, it is possible to "remove" this moderating factor from the ceteris paribus clause and include it in an auxiliary clause to the theory, a clause now understood to be "non-trivial" to explain how an effect occurs. In other words, what we would be doing in such a case is updating our theoretical understanding of the critical features necessary for producing that effect[7], by refining the theoretical model that describes it.

As proposed by Zwaan and colleagues (2018), the idea of reformulating theoretical models by including new moderators seems to fit Lakatos' (1978) notion of sophisticated falsification. In order to properly understand how such an idea could be applied to research

[6] That is, a scenario where one fails to correctly reject the null-hypothesis (i.e. a false negative), generally due to deficient statistical power.
[7] For example, including a clause on the structural features of the environment being important to explain social priming effects.

programmes in psychology, it is necessary to first give an overview of Lakatos' epistemological theory.

Sophisticated falsification in psychology

Lakatosian theory: the basics

Lakatos' idea of sophisticated falsification can generally be seen as evolving from the better-known version of falsificationism proposed by Popper (1959). According to Lakatos (1978), the central idea of Popperian falsificationism is that the scientific character of a theory comes from such a theory being *falsifiable*[8]: the scientific validity of a theory is independent of the evidence in support of it, and it is based instead on the theory specifying the logical case when it would be falsified. In particular, Popper emphasized the notion of a *crucial experiment* - a specific experimental procedure that, upon being performed repeatedly by anyone who has learned the relevant technique, allows us to tentatively accept or reject our theory.

In Popper's view, falsification is the mechanism by which scientific growth occurs, with the most accredited scientific theories being those that have resisted falsification attempts. Upon being falsified, previous theories are rejected and substituted with different theoretical models, effectively allowing scientific progress to take place.

According to Lakatos (1978), this kind of "strict" version of falsificationism suffers from a major shortcoming. That is, the requirement for any theory to forbid a specific observable state of affairs is untenable. This is due to the possibility of formulating post-hoc modifications to theoretical models in order to account for the evidence that initially falsified them. In the case of

[8] Popper's criterion of demarcation for which theories can be considered scientific was crucial in overcoming inductive epistemologies. The idea of falsificationism came from Popper's reflection that no finite number of observations could ever conclusively prove a theoretical proposition in an absolute sense (i.e. the famous "All swans are white" example, see Popper (1959), p.3 - "The problem of induction"), alongside the idea that all universal statements have zero probability of being true on inductive grounds, regardless of the amount of evidence available to us at any given moment. In particular, the "naive methodological" version of Popperian falsificationism identified by Lakatos holds that "Only those theories - that is, non-'observational' propositions - which *forbid* certain 'observable' states of affairs, and therefore may be 'falsified' and rejected, are 'scientific'" (p. 25) [emphasis added]

automatic social behavior, for example, Cesario's (2014) proposal that the physical structures of the environment could be considered as an explanation for a possible failure to replicate Bargh and colleagues' original experiment could be seen as one such post hoc modification.

The idea of finding post-hoc explanations for non-corroborative evidence is central to Lakatos' notion of sophisticated falsification. What Lakatos argues is that, when facing evidence that would normally falsify a theory, it is possible to come up with an auxiliary clause to the theory to account for such evidence. Once this auxiliary clause is implemented in the previous theoretical model, a new theory is created to account for the falsifying evidence. As such, the unit we use to evaluate scientific merit is not a single theory anymore, but rather a series of theories proposed to account for certain phenomena - what Lakatos defines as a *research programme*.

In the Lakatosian theoretical framework, a scientific research programme has several components. On the one hand, there is the *hard core* of the programme - a number of theoretical propositions that are the central component of the proposed theoretical model[9]. These propositions are then metaphorically encircled by a *protective belt* - a number of flexible, auxiliary hypotheses that are continuously updated in light of new evidence and upon which falsification can be deflected. Finally, the last crucial component of research programmes is what Lakatos defines as the *heuristic* - the "problem-solving machinery" (p. 4) that turns recalcitrant evidence into confirming evidence.

According to Lakatos, the main difference between *sophisticated* falsification and the Popperian "naive" version is that the scientific character of theories is not associated with their falsifiability, given the impossibility of falsifying a theory through single empirical observations.

[9] An example Lakatos offers of the *hard core* is the four central laws of Newtonian physics - the three laws of motion and the law of universal gravitation.

Instead, theories, or better, research programmes, acquire scientific legitimacy when the serial formulation of new theoretical models yields *excess empirical content*.

Lakatos identifies two fundamental aspects of the idea of excess empirical content in research programmes. First off, a research programme, through its successive modifications in light of recalcitrant evidence, must be *theoretically progressive*. The idea of theoretical progressiveness reflects the notion that whenever we modify our previous theoretical models by including an auxiliary clause, such modifications must *anticipate new facts*. They must predict something new, rather than simply offering a reinterpretation of previous evidence (i.e. what Popper (1959, p. 20) defined as mere "ad hoc" modifications to a theory). According to Lakatos, this type of progressiveness can be evaluated a priori by logical analysis, without requiring any empirical confirmation.

For instance, Cesario and colleagues' (2010) outline of a new, refined version of previous theoretical models of automatic social behavior would meet this condition. Their newer model implies the existence of a number of possible moderators to ASB (e.g. environmental context), effectively anticipating new empirical observations. An example of such observations could be a situation where the increase in hostile behavior observed by Bargh and colleagues (1996) only occurs when participants find themselves in a confined space.

Secondly, in order to be considered scientifically valid and fruitful, a research programme ought to be *empirically progressive*. Research programmes can be considered empirically progressive whenever the excess empirical content that made them theoretically progressive is partly corroborated. That is, whenever the newly proposed auxiliary clause that accounted for previous recalcitrant evidence is tested and adequately supported by empirical evidence (Lakatos, 1978).

Although we can establish a priori that modifications of the theoretical model explaining ASB effects meet Lakatos' requirement of theoretical progressiveness, I will show in the next sections that the requirement for such modifications to also be *empirically* progressive is hardly met. In doing this, I will mostly appeal to the work of clinical psychologist and philosopher of science Paul E. Meehl (1920-2003) who wrote extensively about the possible application of Lakatosian epistemology in the social sciences.

By referencing Meehl's work, I will discuss how empirical tests are generally characterized in psychology, how researchers incorrectly claim empirical support for their theories by using null-hypothesis significance testing, and what criteria might instead confer empirical progressiveness to a scientific research programme.

Sophisticated falsification in psychology

From his in-depth knowledge of the epistemological, methodological, and statistical aspects of psychological experiments, Meehl (1990) laid out a comprehensive account of a typical deductive model as a premise for his reflections on applying Lakatos' ideas to psychological science. According to Meehl, whenever a substantive theory is subjected to an empirical test, the deductive model used to test the theory can be represented as follows:

$$1. \quad (T \cdot At \cdot Cp \cdot Ai \cdot Cn) \rightarrow (O1 \supset O2)$$

In the model (1), the left-hand side (LHS) represents a conjunction of metatheoretical concepts from which we deduce an observation. T is the theory under test, and can be subdivided in a "Lakatosian manner" into core and peripheral components[10]. The remainder of the LHS

[10] By Meehl's own admission, formal criteria to precisely identify "core" and "peripheral" components of a theory are absent. As further explained in the discussion of the ASB programme in the last section, this distinction can be done "intuitively" by defining the "core" components as that set of statements that cannot be renounced if one is to

conjunction is a number of logical terms that reflect the specific components of an empirical test[11]. From this conjunction of logical terms, a *material conditional* (i.e. the horseshoe) of two observations is deduced. This conditional captures the idea that roughly speaking, if the theory is true, upon observing $O1$, we will observe $O2$.[12]

Taking as an example Bargh and colleagues' (1996) third study, their substantive theory T is that category-related stimuli (e.g. dark-skinned faces) activate behavioral representations, and activated representations directly influence behavior. Factoring in all other logical terms alongside T (e.g. auxiliaries, ceteris paribus), we can *deduce* that upon (subliminally) exposing a Caucasian person to the picture of an African American man (i.e. $O1$), we will observe an increase in hostility in that person (i.e. $O2$).

In this testing situation, falsification occurs by modus tollens, such that if we *fail to observe* the material conditional ($O1 \supset O2$) on the right-hand side (RHS) of (1), we falsify the LHS conjunction. In logical notation, this can be represented as either:

$$2.\ \neg(O1\ \supset\ O2) \rightarrow \neg(T\ \cdot\ At\ \cdot\ Cp\ \cdot\ Ai\ \cdot\ Cn)$$

or equivalently

identify as a "member" of the research programme. Meehl offers as examples those of a disciple of Skinner rejecting the idea of reinforcement or a disciple of Freud rejecting the idea of unconscious mental processes. Such theoretical propositions simply cannot be rejected without renouncing the research programme altogether.

[11] At is what Meehl defines as the *theoretical auxiliaries*, a set of statements needed "to make the derivation to observations go through" (p. 109) and that importantly, includes the operational definition of the constructs under study. Cp is the ceteris paribus clause discussed in the previous section, the assumption that no confounds are at play in our empirical test of the theory. Ai are the *instrumental auxiliaries* or "devices of control", a set of statements necessary to define the instrumentation and set-up of the experiment, and that by definition, do not include any psychological terms. And finally, Cn represents the "particulars", a statement regarding the specific, experimentally realized conditions in the test.

[12] Or conversely, that if the theory were false, the antecedent probability of O1 conditional upon O2 is small. As we will see later on, this is Meehl's (1990) proposed explanation for the notion of *corroboration.*

$$3.\ \neg(O1\ \supset\ O2) \rightarrow\ \neg T\ \vee\ \neg At\ \vee\ \neg Cp\ \ \vee\ \neg Ai\ \vee\ \neg Cn$$

Therefore, whenever we fail to observe an increase in hostility upon exposing Caucasian people to pictures of African American men, we falsify the whole conjunction of logical terms that constituted our empirical test. From model (1), we can see how falsification of T is complicated by its conjunction with all other terms in the context of an empirical test. Meehl (1990) explains this by noting how logically, the negation of the material conditional ($O1 \supset O2$) is necessarily equivalent to the disjunction of the negations of all terms in the metatheoretical bundle, as represented in (3). This essentially means that whenever the consequent ($O1 \supset O2$) is negated, any one of the logical terms could be "responsible" for this negation, but from a logical standpoint, it is not possible to establish whether the theory, auxiliaries, or the ceteris paribus is the "culprit" in such a scenario[13].

Meehl's logical representation of an empirical test comes in handy to understand the various components of a Lakatosian research programme and how sophisticated falsification works. On the one hand, we have the theory T whose central theoretical statements form the Lakatosian *hard core* of the programme. On the other, we have the rest of the conjunction, which together with the peripheral parts of T form the Lakatosian *protective belt*. As mentioned earlier, the protective belt serves the function of "deflecting" falsification away from the hard core and onto the belt[14]. Looking at Meehl's model, whenever the material conditional is negated, we can direct falsification onto any of the other terms of the conjunction. We could question the

[13] This situation well exemplifies the main idea of *holistic underdetermination*, alternatively known as the Duhem-Quine Thesis (DQT) (Stanford, 2023). The essential problem posed by this type of underdetermination in science comes from Duhem's and Quine's observation that in empirical tests (and according to Quine, for knowledge in general), it is impossible to ever test a hypothesis in isolation. Any hypothesis stemming from a substantive theory is always tested in conjunction with a number of other assumptions and clauses. As such, straightforward falsification of a theory is not possible in light of this inevitable conjunction.

[14] What Lakatos defines as the "negative heuristic" (p. 47).

auxiliaries, we could reject the ceteris paribus, or we could deny the particulars. Importantly, this process works outside the realm of formal logic and is instead, in Meehl's own words, a matter of "*scientific strategy*" [emphasis added] (p. 110).

One such strategic choice would be Cesario's (2014) proposal that our deficient theoretical understanding of effects like automatic social behavior is responsible for a failure to replicate them. In the context of a replication failure, instead of directing falsification onto the theoretical model proposed by Bargh and colleagues (1996), we could say that Cesario directed it toward the protective belt, in this case toward the ceteris paribus. In so doing, he effectively "spared" Bargh and colleagues' (1996) theoretical model by challenging the ceteris paribus, refining the model according to additional clauses positing the influence of new moderating variables. As noted earlier, this seems to show that the idea of "hidden moderators" might fit well in a context where sophisticated falsification is applied to improve theoretical understanding of psychological effects.

Now the question naturally arises, when exactly is the deflection of the modus tollens from the core to the protective belt epistemologically justified? We mentioned earlier that, in Lakatos' view, a process of sophisticated falsification is admissible whenever a modification of our theoretical model anticipates new facts that are later corroborated by new evidence. In such a case, the ad hoc modification of a theory to fit the falsifying evidence is permitted, and a research programme can be defined as progressive. Conversely, whenever successive post-hoc modifications do not anticipate new facts or are not corroborated by new evidence a research programme becomes *degenerating*. But this answer necessarily raises the question of what exactly constitutes *valid* corroborating evidence in psychological science and warrants a critical discussion of how evidentiary support is generally characterized in our discipline.

The epistemic state of psychology

What counts as evidence - the problems of null-hypothesis significance testing

In psychology, the manner in which theories are commonly tested is by employing null-hypothesis significance testing. In the case of Bargh and colleagues (1996), we have a certain theory that attempts to explain how Caucasian people exposed to pictures of African American men will show more hostility compared to the same people being shown pictures of Caucasian men. The general idea is that to test the substantive theory explaining the relationship between variables X (e.g. type of prime) and Y (e.g. level of hostility), we deduce a hypothesis (generally referred to as "alternative" or "counter-null") that two groups differing on X, will also systematically show a difference in the group averages of the second variable Y[15].

Now, upon observing a difference between groups in the variable Y, we ask ourselves: is this difference observed in our sample something we would find if we could, hypothetically, measure the whole statistical population? (i.e. all Caucasian people exposed to the different pictures). To provide an answer, we assume that the null hypothesis saying that no actual difference in variable Y exists at the level of the population is true. With that assumption, we then ask: with what relative frequency would we find a difference of the magnitude we observed if no such difference truly existed? Or alternatively, how unlikely is it to find a difference like the one we observed if the null-hypothesis were true? Finally, we arbitrarily choose a certain value of distance to the null point[16]. Upon exceeding this value, we reject the null hypothesis and we conclude that our relationship is statistically significant[17].

---

[15] Importantly, within this testing situation we admit the part of the difference between the two groups on Y could be either due to measurement errors or to sampling error.

[16]The null point refers to a specific numerical value (generally zero) assumed to be true under the conditions stated in the null-hypothesis that no significant difference between groups exists at the level of the population.

[17] For an in-depth discussion of the use of null-hypothesis testing in psychology and associated flaws, see Meehl (1967).

Problems arise, however, the moment we want to treat such differences as actual *evidentiary support* for our theories. Meehl (1990) criticized the common inference psychologists make when using NHST as based on clear logical flaws. He characterizes the situation of a statistical hypothesis test as a double deductive chain. From the theory T we deduce the statistical hypothesis H that two groups will differ. In turn, from the statistical hypothesis H, we then deduce that we will observe O, an actual difference in two observable variables. The chain can be graphically represented as:

Deduction

$$T \rightarrow H \rightarrow O$$

Whenever we fail to observe O we are justified in rejecting H by modus tollens. In equivalent manner, upon rejecting H we will also reject T. This reasoning is formally valid, and it leads to the conclusion outlined in (2), that upon not observing the material conditional ($O1 \supset O2$), we will reject the metatheoretical bundle ($T \cdot At \cdot Cp \cdot Ai \cdot Cn$).

Actual problems arise whenever we require our observation O to constitute *evidence* for the theory T. This, according to Meehl (1990) necessarily works by two inferential steps, represented as:

Inference

$$T \leftarrow H \leftarrow O$$

The first step requires us to infer that upon observing a significant difference O, our hypothesis H be true. The second inference requires us to infer that upon accepting our hypothesis, we believe our theory T to be true. Even assuming that our experiment reaches methodological and statistical perfection so that no error is present in our test[18] and so that this first inference is well-justified, the second inference is not logically tenable.

The fallacious nature of this type of inference can be explained by the simple consideration that upon observing O and confirming H, several *alternative* theories could provide a suitable explanation for our hypothesis H. According to Meehl (1990), this is a common mistake on the part of psychologists, such that by conflating the substantive theory T with our statistical hypothesis H, we affirm that O *provides evidence* for the veracity of T, something not logically (and arguably not even methodologically) justified.

Furthermore, what is particularly problematic is that in so-called "soft"[19] areas of psychology like social priming, substantive theories only make very vague predictions about observations to be made. As noted above, the typical prediction in psychology takes the form of a *directional estimate* (Meehl, 1990). Such estimates essentially only make predictions on how two groups or two experimental conditions (e.g. X1; X2) will differ on a variable Y in reference to whether this difference would be greater or lower than a null point[20]. Given that psychological theories in areas like the study of ASB only make loose predictions regarding the direction (lower or greater than zero) of a certain relationship, Meehl (1990) notes that the idea that observing a significant difference in that direction *strongly corroborates* these theories seems unjustified, given that many other possible theoretical alternatives could be equally

[18] Something arguably unachievable, and certainly not reflecting methodological and statistical standards in psychology.
[19] See Meehl (1978) for a definition.
[20] This state of affairs is radically different from other scientific disciplines (those commonly referred to as "hard" sciences), whereby the typical predictions take the form of *point estimates*.

well-supported by the same observation. In the case of ASB, an observed difference in levels of hostility after exposure to pictures of African American men could have alternative explanations other than an association between African Americans and hostile behavior. For instance, faces of African American men might simply be less familiar than faces of Caucasian men to Caucasian participants, increasing the likelihood of a hostile response.

The situation outlined above is one of considerable concern for psychology. The notion that multiple theoretical alternatives are equally well-supported by the same evidence is formally known as the problem of *contrastive underdetermination* (Stanford, 2023), an epistemological riddle that seems to forbid that any scientific theory can ever be conclusively proven true or trusted in an absolute sense. Given the considerations above, this type of underdetermination seems to be particularly relevant for psychological theories, at least in certain areas of research. Importantly, it makes Lakatos' requirement that research programmes be empirically progressive hard to meet in the present state of affairs. Put simply, sophisticated falsification requires a research programme to be well-corroborated. In turn, corroboration requires the inference from our hypothesis H to our theory T to be well-justified. Directional estimates like the ones predicted by theories such as the one explaining ASB can be easily accounted for by alternative explanations. As such, they do not seem to be corroborative enough and they suffer from a considerable degree of contrastive underdetermination.

In the following section, I will further elaborate on this idea and I will explore how some degree of empirical progressiveness could be ascribed to research programmes in psychology. I will do so by discussing Meehl's proposal of the *Lakatos-Salmon Principle* and how it could be applied to research programmes investigating automatic social behavior.

What counts as "good" evidence - the Lakatos-Salmon principle

In the previous section, I discussed why the way *evidentiary support* for theories is often characterized in psychological science is quite unsound. Psychologists make use of NHST to claim empirical support for the substantive theories under test. As pointed out by Meehl (1990) now more than 30 years ago, this method of empirical testing is highly problematic. Psychological theories only make vague predictions (e.g. directional estimates) about what exactly we should expect regarding the relationship between two variables of interest. As such, this leads to a situation whereby these theories are underdetermined by the findings of empirical studies. In other words, successfully finding significant results when testing a theory through NHST does not in any way warrant strong beliefs in a theory's truth-value, given that several alternative explanations could be equally well-supported by the same findings.

Given Lakatos' requirement that research programmes be empirically progressive, the idea of viewing replication failures as opportunities to apply Lakatosian sophisticated falsification is not reasonable in light of the current epistemological standards in psychology. Not just the new, modified theory would not be empirically progressive under these circumstances, but even the original theory could hardly be said to be well-supported by evidence in the first place. As Meehl puts it: "If the test consists of mere refutations of the null-hypothesis, [...] it is not rational to adopt the Lakatosian heuristic and engage in strategic defensive retreat, because we had feeble grounds to appraise the theory as it stood *before* it began to ran into apparent falsifiers" (p. 115).

This state of affairs is worrying, and it must be seen as a call to make some decisive improvements in the way theories are formulated and tested in psychological science. Sophisticated falsification can be a fruitful avenue in designing strong and epistemologically

sound research programmes in psychology. A priori, the idea of strategically interpreting replication failures as an opportunity to apply sophisticated falsification is still warranted.

At the same time, it is clear that if NHST is the primary way by which theories are tested in psychology, such sophisticated falsification would not lead to empirically progressive research programmes. In simple terms, empirical progressiveness requires the evidence to be "good", while the evidence we obtained through experiments relying on NHST can hardly be characterized as such.

Now, what kind of changes will be needed in order to make sophisticated falsification a useful strategy in case experimental replications fail? It is once again Meehl (1990) who offered great insights into what our epistemic goals should be when designing and testing theoretical models, and which meta-scientific principles we should follow in order to address underdetermination in psychological science.

In Meehl's view, in order to be trusted (and therefore less underdetermined) scientific theories should have high *verisimilitude*. By Meehl's own admission, a clear, formal definition of the concept of verisimilitude is missing in the metascientific literature. Nonetheless, high verisimilitude could be roughly defined as the notion that a theory's description of a phenomenon strongly resembles what that phenomenon actually looks like in objective reality (Meehl, 1990). It is a good description of what the phenomenon truly is *from an ontological perspective.*[21]

The epistemological counterpart of verisimilitude is the idea of *corroboration*. Going back to Meehl's proposed deductive model of an empirical test, we regard a theory T as well-corroborated by evidence whenever *absent* the theory T, the conditional probability of an observation $O2$ upon a second observation $O1$ (as described by the theory) is very low. In other

[21]If the reader is interested, Meehl (1990) nonetheless attempts to give a definition of the concept in the section "Excursus: the concept of verisimilitude".

words, a theory is well-corroborated by observational evidence when such evidence would be very unlikely if the theory were false.

In the case of Bargh and colleagues (1996), their theoretical explanation of the priming effect they observed would be well-corroborated if absent this theoretical explanation, observing an increase in hostility of Caucasian students after exposing them to pictures of African American men would be very unlikely. On the contrary, the theory would not be highly corroborated by the observed evidence whenever absent their theory, observing such an increase in hostility by Caucasian students would *nonetheless* be very likely. As such, this observation would not be a strong indicator of the verisimilitude of our theory. Importantly, this latter case would be an example of high contrastive underdetermination. In such a case, the fact that the observation ($O1 \supset O2$) could be easily explained by a bunch of alternative theories would result in such an observation being likely despite the substantive theory T being false.

This explanation shows the intimate relationship between corroboration and underdetermination. The more findings corroborate a substantive theory, the less the theory is underdetermined. Equivalently, the fewer alternative explanations there are for the findings, the more corroborating those findings are with respect to the theory. From this reasoning, we preliminarily conclude that what we might call "good" evidence are findings that strongly corroborate our theories. But once again, this consideration begs a further question, namely: when can we say that a prediction is sufficiently unlikely so as to be corroborative? And in the context of such a question, when is sophisticated falsification or *Lakatosian defense,* as Meehl referred to it, warranted in light of recalcitrant evidence?

Meehl (1990) attempts to give an answer to this question based on a precept he defines as the *Lakatos-Salmon principle* (LSP). This principle is in turn a conjunction of two others. The

first one is what Meehl defines as the "*Money in the bank*" principle and it reflects the idea that a theory can be said to be well-corroborated whenever it has passed a number of very *stiff* tests[22]. Having passed such tests means the theory has a good "track record" (p. 115). According to Meehl, in light of an apparent falsification, if a theory has accumulated money in the bank (i.e. it has successfully survived a number of very stiff tests), the "Lakatosian defense" of such theory is preliminarily justified.

The stiffness of a test is in turn defined through the second component of the LSP, the epistemological criterion defined ironically by Wesley Salmon (1984) of certain experimental results being a "*damn strange coincidence*". The criterion of damn strange coincidence essentially reflects the meaning of corroboration we outlined above. An experimental observation can be said to be highly corroborative whenever its antecedent probability absent the substantive theory is low. Put simply, if our theory successfully predicts something that a priori is very unlikely (and I would add, if this prediction is replicated a number of times), this evidently constitutes an *empirical strange coincidence*. In turn, the stiffness (or riskiness) of our theory-testing procedure is defined by how unlikely the predictions derived from my theory are.

Very unlikely predictions would mean that such a procedure is, in principle, very intolerant: our theory can easily be falsified. Passing a number of tests that made very unlikely predictions then warrants our belief that there's gotta be something to the theory, those unlikely successful predictions cannot just be "damn strange coincidences". Therefore, whenever this happens we say that our theory is well-corroborated, and we bestow upon it high verisimilitude[23]. In turn, when we encounter recalcitrant evidence, it seems reasonable that - in light of the

[22] This is the "Lakatos" part of the Lakatos-Salmon principle, as Meehl refers to it as the "Big Lesson" (p. 115) Lakatos had to teach us.

[23] In Meehl's view, the credit accumulated by a theory through past successes in passing stiff tests is in a *direct stochastic relationship* with verisimilitude. As such, we are warranted to believe that the more such tests were passed by a theory, the more likely the theory is to have high verisimilitude.

theory's past success in passing very intolerant tests - we "cut the theory some slack" so to say, temporarily deflecting falsification upon the Lakatosian protective belt.

Now once again, it is necessary to go on a little tangent so as to define what exactly an improbable, corroborative prediction would be in the context of psychological science. Meehl makes a number of suggestions that have inspired arguments from recent scholars on how to make better predictions, improve our theories, and claim stronger corroboration.

Notably, there is the idea that experimental results would constitute damn strange coincidences whenever very similar numerical results are found from *qualitatively different observational avenues*. In such a case, the fact that experiments in qualitatively different domains yielded the same (or very similar) numerical results would be quite surprising and would indeed constitute a case of strong corroboration.

Imagine a situation whereby Bargh and colleagues' theory (1996) would make predictions on the *actual extent* to which hostility in Caucasian students increases as a function of exposure to pictures of African American male faces, such that these predictions turn out to be exact (or sufficiently close) to the observed value in studies where hostility is measured in qualitatively different manners. If, for instance, we could correctly predict a precise increase in hostility as measured by different operationalizations[24] (e.g. observation, self-report, physiological measures, etc.), that would indeed be quite a "coincidence". In the case of a replication failure, having observed such empirical agreement in previous studies would lead us to conclude that temporarily amending our theory on automatic social behavior (by, for instance, conceiving the existence of a hidden moderator) would be an epistemologically justified strategy.

[24] What would be defined as "conceptual replications" in Schmidt (2009) or as "Far replications" in Lebel et al. (2017).

It is important to note how this metascientific criterion for strong corroboration requires theories to make precise numerical predictions. Meehl notes that the situation where a precise point estimate can be derived from a theory is rare in psychological science, and particularly so in areas of "soft psychology" (e.g. social psychology). Concurrently, other formal criteria can be set up in order to make predictions unlikely enough to make our theory-testing procedures sufficiently intolerant.

In Meehl's view, predicting that the value of a certain variable of interest would fall within a specific numerical interval could make for an "unlikely enough" prediction whenever such interval is judged against the *a priori range of logical possibilities*[25] for that value. This allows us to judge just how intolerant our prediction is in light of all the possible values that the predicted variable can assume. In a very simplified manner, if our predicted interval is very small compared to the range of all possible values, our theory-testing procedure can be said to be very intolerant. In other words, in such a case it would be very unlikely to get a "hit", and therefore our testing-procedure would be risky enough to claim strong corroboration if successful.

The idea of strengthening the evidentiary support of findings through the design of severe tests is appealing and suggests an important way to build solid epistemic grounds for research programmes in psychology. Yet, the problem still remains how exactly severe tests could be designed given the apparent inability of theories in soft areas of psychology to strongly imply corroborative predictions.

Theories in these areas are generally specified in purely verbal terms. As such, they do not allow us to strongly deduce specific observations of how the phenomenon they are supposed to represent will behave. According to Robinaugh and colleagues (2021), this is because expressing a theory's structure in words is necessarily "limited by the imprecision of natural

[25] What Meehl refers to as the *Spielraum*.

language" (p.727). Following Meehl's initial criticism of the detrimental value of NHST for epistemic progress, the authors instead propose *formal models* as a promising way forward to design better theories in soft areas of psychological science.

Their focus rests mostly on the importance of the explanatory accuracy of such models, but they do emphasize that a formalized version of a theory generates specific and strongly implied predictions, something well-aligned with Meehl's criteria for designing severe tests and obtaining corroborative findings.

They show this by attempting a formalization of the well-known vicious-cycle theory of panic attacks proposed by Clark (1986). In particular, they propose to formalize the relationship between the components of the vicious cycle (i.e. level of arousal and perceived threat) by means of a difference equation. This allows in principle to define how these two elements influence each other over discrete units of time. They propose four distinct formalized models, defined by different configurations of the relationships between the two components mentioned above as either linear or sigmoidal functions.

Once the initial verbal theory is formalized, the system can be then simulated as a computational model, yielding precise function forms of how the vicious cycle between arousal and perceived threat plays out according to the different configurations. The simulated function forms are what the authors referred to as the *theory-implied behavior* of the target system[26] being represented by the formal theory. In essence, this implied behavior is the actual prediction of how the phenomenon of panic attack would occur under the conditions specified by one or another formalized version of Clark's theory (1986).

[26] The target system refers to the real-world components and their relationships giving rise to a specific observable phenomenon. In the case of Clark's theory (1986), the target system would refer to arousal, perceived threat, and how they interact with each other in order to give rise to a panic attack.

In order to test the precision with which a formal model can represent real-life phenomena, theory-implied behavior must then be compared with empirical data of how the phenomenon plays out in a real-world context. Discussing the full epistemic merits and shortcomings of this proposal falls outside the scope of this thesis. Yet, what the idea of formalization clearly shows is the capability of formal models to improve the inherent vagueness of verbal theories describing effects like automatic social behavior.

Formal models can generate highly precise predictions of how these phenomena would occur were the theory true. As such, they seem to meet the Lakatos-Salmon principle outlined by Meehl. Actually observing the relationship between arousal and perceived threat playing out in a way that resembles the function form implied by one or another formalized version of the model would be quite a "damn strange coincidence" in Salmonian terms. Consequently, an experiment that tests whether the specific mathematical behavior of the predicted target system looks anything like the observable phenomenon would then constitute a fairly stringent test of the theory. Following Meehl's reasoning, a theory that *correctly* predicts such specific observation (or that comes sufficiently close) could then be said to be well-corroborated, and that very specific observation would constitute appropriately good evidence for the theory in question.

A closer look at the research programme on automatic social behavior

In the following section, I wish to take a closer look at studies that have further investigated the phenomenon of automatic social behavior after the original paper published by Bargh and colleagues (1996). This will serve the purpose of offering a preliminary analysis of the epistemic state of the research programme, understanding which theoretical modifications have been proposed to explain these effects, and further answering the question of whether sophisticated falsification can be applied to cases where replications fail in ASB studies.

In the years following the original work investigating automatic social behavior, different studies have replicated and expanded findings in this research area. Notably, an overwhelming majority of these findings have been "conceptual" (Schmidt, 2009) or "far" replications (Lebel et al., 2017). As such, they cannot speak for the actual reliability of the original effect.

Of the studies originally conducted by Bargh and colleagues (1996), study 2 on the influence on behavior of priming the category "elderly" seems to have been the one attracting the largest number of replication studies. Among these replications, we include findings on performance differences in the expected direction between an elderly prime and priming the category "basketball players" in a ball-throwing task (Follenfant et al., 2005), effects of an elderly prime on performance speed but not on manual dexterity in fine-motor tasks (Ginsberg et al., 2012, study 2 and 3), and effects of the elderly prime on unique elements of an action sequence (Banfield et al., 2003).

In terms of moderating influences, Dijksterhuis and colleagues (2000) seemingly found evidence for the potential role of self-reported contact with elderly people on the effect of an elderly prime on behavior in a recall task (study 2a), while Cesario and colleagues (2006) provided evidence for the role of attitude toward elderly in producing priming effects on behavior.

Finally, one series of studies by Doyen and colleagues (2012) attempted a direct replication of the original work. In study 1, a methodologically similar replication of the original experiment yielded no significant results despite a larger sample size and improved measurements. In study 2, they found evidence for the possible moderating role of the experimenter's expectations on the effect of elderly prime on walking speed. That is, the

decrease in walking speed observed in the original study was only observed when the experimenter was led to believe that participants would walk slower following a prime.

Study 3 of the original paper (on the influence of priming the category “Black people” on non-African American participants) was also investigated in a number of replication studies. As mentioned earlier, Cesario and colleagues (2010) found evidence for a possible moderating effect of the physical structures of the environment and of resource-holding potential on the influence of an African American prime on the accessibility of fight-related words, aggressiveness, and preparation to fight. Hagiwara and colleagues (2012) found evidence for the independent effects of skin tone and facial features on automatic affectivity (i.e. cognitive accessibility of positive and negative words) following an African American prime. Finally, DeMaree and Loersch (2009) reported evidence for the moderating role of self-focus as opposed to other-focus in producing priming effects on aggressive behavior after subliminal exposure to words related to the social category of African American people.

By close examination of the studies above, we can attempt to give a preliminary definition of what the ASB research programme looks like. We can do so by trying to identify its central theoretical statements - the *hard core*. Lakatos (1978) never gave explicit instructions on how to identify the hard core of a research programme. He generally referred to this set of statements as the ones that are “protected” in the process of sophisticated falsification.

We might think of the hard core as the part of the theory that simply cannot be rejected without renouncing the research programme altogether, which is Meehl’s (1990) preferred definition of the hard core. In this respect, I maintain that the research programme investigating ASB can be associated with the following two propositions:

P1: A stimulus related to a certain social category will increase the accessibility of behavioral and trait representations associated with that category.

AND

P2: The increased accessibility of a certain behavioral and trait representation automatically influences behavior.

These two statements can generally be thought of as the hard core of the programme. No researcher investigating ASB effects would reject them, because together, they reflect the very essence of what automatic social behavior represents. That is, the idea that subtle environmental cues can influence behavior without the person knowing it.

I argue that the theoretical formulations of different accounts of ASB effects generally differ in how P2 is interpreted through further auxiliary clauses. The original account by Bargh and colleagues (1996) holds that the activation of a behavioral representation has a *direct* influence on behavior. The authors explain this mechanism in terms of William James' (1890) principle of ideomotor action, which they define as the idea that "the mere act of thinking about a behavior [increases] the tendency to engage in that behavior" (p. 231). This particular interpretation is endorsed by a majority of the research teams who studied ASB effects[27].

On the other hand, there are alternative accounts that reformulated the theoretical model by including new auxiliary clauses to explain the second proposition, rejecting the idea that activated representations directly influence behavior with no other mechanism involved.

[27] Of the ones mentioned above Dijksterhuis et al. (2000), Banfield et al. (2003), Follenfant et al. (2005), and Ginsberg et al. (2012).

Cesario and colleagues (2010) proposed what they refer to as a "motivated preparation account" of automatic social behavior. Differently from the original account by Bargh and colleagues (1996), they maintain that automatic effects on behavior elicited by a category prime are better explained by seeing these behaviors as a consequence of participants' preparation to interact with a member of the primed category. As such, their theoretical accounts include two additional auxiliary clauses:

A1: Accessible trait and behavioral representations related to a social category activate specific goals of interaction.

AND

A2: The selection of a specific behavioral response to pursue a goal depends on a number of other relevant factors[28].

The inclusion of these two clauses then provides an alternative explanation of how P2 plays out. Accessible trait representations activate a social goal which in turn leads to the selection of a specific behavioral response when considered along other relevant factors (e.g. physical surroundings). As noted above, this reformulation can be understood as a challenge to the ceteris paribus clause. By defining further auxiliary clauses, the authors effectively deny that

[28]For instance, positive as opposed to negative attitudes toward elderly people would produce opposite effects on behavior in participants primed with a category-related stimulus. Participants with positive attitudes toward the elderly will supposedly prepare themselves for a friendly interaction with a member of the category and would therefore slow down following a prime. On the contrary, participants with negative attitudes toward the elderly will not be motivated to interact with the category member, and would therefore not slow down following a prime.

no other factors other than category-related stimuli played a role in the production of ASB effects.

Loersch and Payne (2011) also offer an alternative and more nuanced explanation of automatic social behavior through what they refer to as the "situated inference model". They also included further clauses in the model in order to give a specific explanation for how behavior is influenced by environmental cues. In their account, the following two auxiliary clauses are included:

A3: Accessible trait and behavioral information is misattributed to one's own natural response to a specific environmental stimulus.

AND

A4: The misattributed information is used to answer a specific question afforded by the environment.

As stated in these two clauses, the automatic production of specific behavior is mediated by a process of misattribution and depends on what behavioral responses are afforded by the environment. This last feature is shared with Cesario and colleagues' (2010) model, and also implies the influence of specific moderating factors on ASB effects[29].

Now, both these novel accounts undoubtedly offer a more nuanced view of automatic social behavior. They do posit the influences of additional factors such as environmental

[29] Indeed, both groups of researchers reference Gibson's (1997) theory of affordances in visual perception in outlining their respective models.

characteristics, attitudes, and focus of attention on the way ASB plays out. Arguably, this can lead to observations that falsify Bargh and colleagues' original account[30], and to predictions that are somewhat more specific in light of newly proposed moderators.

In a simplistic manner, these alternative accounts could explain why direct replications based on the original account fail, as proposed by Cesario (2014). Yet, I argue that in light of the requirements of sophisticated falsification, this interpretation is ultimately ill-informed. The evidentiary support for these new models does not meaningfully differ from the one used to support old models. Notably, virtually all studies, both maintaining the original model and proposing a new version, make use of null-hypothesis significance testing. They characterize the relationship between variables in vague terms that translate to a statistical hypothesis about the difference between group means being greater or lower than a null point. These types of predictions do not lead to highly corroborative evidence in a theory-testing situation (Meehl, 1990). They fail to appropriately meet the criteria of the Lakatos-Salmon Principle laid out by Meehl: the observations they lead to are not "unlikely enough" to constitute a Salmonian damn strange coincidence; as such, substantive theories like the ones outlined above are not tested by experiments *risky* enough to testify to their value, and can therefore hardly be taken as evidence of the verisimilitude of a certain theoretical explanation.

Relatedly, the predictions of these theoretical models lead to a situation where upon successfully confirming them, even in the unlikely situation where no methodological or statistical errors are present, the evidence produced grossly underdetermines the models proposed to account for it. In other words, there are no strong epistemological reasons to prefer these models over alternative accounts of automatic social behavior.

---

[30] For an example, see Cesario and colleagues' (2010) study one on the behavioral effects of priming the category "gay people", where the prime produced a response not in line with Bargh and colleagues' (1996) original account, but that could be explained through the motivated preparation model the authors propose.

Therefore, the idea of operating a form of Lakatosian defense upon a failure to replicate would be misguided in this epistemological context. Reconceptualizations of theoretical models in this research programme are theoretically progressive - they do lead to novel, potentially testable predictions. But following Meehl's reasoning, they fail a priori to be empirically progressive, given their incapability of yielding predictions that would produce truly corroborative evidence.

In light of its epistemological standards, I argue that the research programme studying ASB effects can most likely be defined as *degenerating* in a Lakatosian sense. Its occasional falsifiers in the form of replication failures can be dealt with by ad-hoc modifications, but the predictions that such modifications yield are not specific enough to constitute stringent and corroborative tests.

Meehl (1990) pointed this fact out in his discussion of research programmes in psychology. He maintained that the true problem of such programmes is the "weak epistemic linkage" (p. 127) between theories and hypotheses. The deficient corroborative value of a hypothesis in the context of NHST does not give us good reasons to believe in the theory that produced it. Importantly, this is a problem completely unrelated to the possible statistical shortcomings of such experiments. It is a matter of what is logically admissible and epistemologically justified. As Meehl put it: "No statistical ingenuity can cure a logician's complaint about the third figure of the implicative syllogism, that the theory is a sufficient but not necessary condition for the fact, by *casting doubt on the fact*" (p. 127)

Conclusion

In the previous sections, I discussed at length the possible application of Lakatos' notion of sophisticated falsification to current research programmes in psychology. This idea was

briefly proposed by Zwaan and colleagues (2018) and reflected Cesario's (2014) argument that the current theoretical understanding of ASB effects does not allow us to interpret replication failures as evidence of type I errors. By focusing on the research programme on automatic social behavior started by Bargh and colleagues (1996) I showed how the idea of interpreting failures to replicate as an opportunity to apply sophisticated falsification is not justified in light of the current epistemic state of the programme.

In the analysis, I primarily focused on the theoretical aspects of the programme. The experiments investigating ASB and social priming effects have been primarily criticized for their statistical and methodological shortcomings (e.g. low power). These aspects of an experiment generally complicate the inference from observations to hypotheses. I argue that a further problem lies in the unjustified inference from statistical hypotheses to theoretical models.

As noted by Meehl (1990) many decades ago, the common practice of conflating hypotheses with theories leads psychologists to claim that experimental observations provide evidence for theories. Such a position is ill-informed. On the one hand, it is generally not justified in light of the problem of induction: theories imply observations, but observations do not imply theories. On the other hand, inductive inferences are not *epistemologically* justified in light of how theoretical models are constructed and tested in certain areas of psychology.

Theories on ASB have always been tested through null-hypothesis significance testing. This testing procedure characterizes observational predictions in terms of directional estimates, magnifying the inherent vagueness of purely verbal theories. As such, even if predictions are confirmed by observational evidence (i.e. the null-hypothesis is rejected), that evidence is not *corroborative* enough; even if observed, predictions are not "unlikely enough" to constitute good reasons to believe that a certain theoretical explanation has verisimilitude.

This absence of corroboration goes hand in hand with the notion that theories in psychology suffer from high degrees of contrastive underdetermination. That is, the same observational evidence can be easily explained by alternative theoretical explanations in light of its high antecedent probability.

As noted above, the research programme on ASB effects can hardly be characterized as progressive from a Lakatosian perspective. Lakatos required successive modifications of a research programme to yield corroborated empirical content. As we saw in previous sections, theoretical models like Bargh and colleagues' (1996) or Cesario and colleagues' (2010) fail *a priori* to be well-corroborated in light of their vague predictions.

Given the current epistemological standards in the programme, I believe it is possible to preliminarily conclude that as it currently stands, the research programme on ASB effects is most likely in a state of degeneration. Further, I argue that so long as its epistemological standards are the norm in many areas of psychological science, this idea might generalize to several other research programmes.

A promising way forward might be relying on *formalized* versions of theoretical models. As we noted above, formal models are at least capable of yielding specific, strongly implied predictions. As such, they better align with the criteria laid out by Meehl on how to construct severe, corroborative testing procedures. Notably, the use of formal models might entail other problems, especially in connection with the fact that predictions from formalized models necessarily need to be compared with empirical data obtained through measurements with questionable validity (Oude Maatman, 2021).

In addition, I believe that future research efforts should be invested in exploring the idea proposed by Uygun Tunç & Tunç (2022) of developing *systematic replication frameworks* (SRF)

as forms of Lakatosian research programmes. From an epistemological perspective, the notion of controlled variation of experimental procedures proposed by the authors constitutes a promising methodological tool. Such variations allow us to disentangle the differential effects that specific features of successive replication studies have on a phenomenon.

In light of the considerations put forward in this thesis, a sounder implementation of the SRF they propose might consider the use of formalized versions of theoretical models in order to yield more corroborative and stringent tests. This would allow us to better perform the function proposed for SRFs of revealing if "the corroboration of the [theoretical hypothesis] is restricted to particular versions of the model with particular [auxiliary hypotheses]" (p. 12).

The present thesis might suffer from a number of limitations that prompt further reflection on the arguments proposed. Notably, one possible limitation concerns the limited explanation given for how corroboration and underdetermination relate to each other. In the thesis, I argued that the two notions are connected in that findings that are more corroborative reduce the extent to which a theory is underdetermined by those same findings. I believe that although prima facie justified, this explanation would benefit from further theoretical elaboration and possibly from a formalized description.

One further limitation is that the thesis only focuses on the epistemological aspects of scientific practice from a traditional, individualistic perspective. The conclusions of the analysis might not be relevant in a context where research practices are analyzed from a social epistemological perspective. In light of the problems posed by underdetermination, social epistemology generally rejects the idea of identifying a priori criteria like falsification or corroboration to define why scientific research is an epistemically valid practice, focusing

instead on how the social nature of science explains the success of scientific practice as an epistemic enterprise (Oreskes, 2019, Longino 2022).

In conclusion, Lakatosian sophisticated falsification in scientific research programmes constitutes a valid scientific strategy to improve theoretical understanding of psychological phenomena. This must necessarily be preceded by the construction of theories that allow to obtain corroborative results. Psychological science is undergoing major changes in light of the problems evidenced by the replication crisis in the last decade. Amidst this period of change, I believe that a valid implementation of Lakatosian research programmes in the architecture of scientific practice would be a decisive step forward in improving the ability of psychology to produce solid and reliable knowledge.

Reference list

Banfield, J., Pendry, L., Mewse, A., & Edwards, M. (2003). The Effects of an Elderly Stereotype Prime on Reaching and Grasping Actions. *Social Cognition*, 21, 299-319. https://doi.org/10.1521/soco.21.4.299.27002

Bargh, J. A. (2006). Agenda 2006: What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, 36(2), 147-168. https://doi.org/10.1002/ejsp.336

Bargh, J. A. (2021). All aboard! 'social' and nonsocial priming are the same thing. *Psychological Inquiry*, 32(1), 29-34. https://doi.org/10.1080/1047840X.2021.1889326

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71(2), 230-244. https://doi.org/10.1037/0022-3514.71.2.230

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, 9(1), 40-48. https://doi.org/10.1177/1745691613513470

Cesario, J., Plaks, J. E., & Higgins, E. T. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology*, 90(6), 893-910. https://doi.org/10.1037/0022-3514.90.6.893

Cesario, J., Plaks, J. E., Hagiwara, N., Navarrete, C. D., & Higgins, E. T. (2010). The ecology of ole of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 12(1), 46-61. https://doi.org/10.1177/1745691616654458

Clark, D. M. (1986). A cognitive approach to panic. *Behaviour Research and Therapy*, 24(4), 461-470. https://doi.org/10.1016/0005-7967(86)90011-2

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66, 93-99. https://doi.org/10.1016/j.jesp.2015.10.002

DeMarree, K. G., & Loersch, C. (2009). Who am I and who are you? Priming and the influence of self versus other focused attention. *Journal of Experimental Social Psychology*, 45(2), 440-443. https://doi.org/https://doi.org/10.1016/j.jesp.2008.10.009

Dijksterhuis, A., Aarts, H., Bargh, J. A., & van Knippenberg, A. (2000). On the Relation between Associative Strength and Automatic Behavior. *Journal of Experimental Social Psychology*, 36(5), 531-544. https://doi.org/10.1006/jesp.2000.1427

Dilworth, C. (1996). *The metaphysics of science. An account of modern science in terms of principles, laws and theories.* Dordrecht: Kluwer.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, 7(1). https://doi.org/10.1371/journal.pone.0029081

Eerland, A., Sherrill, A. M., Magliano, J. P., & Zwaan, R. A. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, 11(1), 158-171. https://doi.org/10.1177/1745691615605826

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 12(1), 46-61. https://doi.org/10.1177/1745691616654458

Follenfant, A., Légal, J.-B., Dinard, F., & Meyer, T. (2005). Effet de l'activation de stéréotypes sur le comportement: Une application en contexte sportif. *European Review of Applied Psychology*, 55, 121-129. https://doi.org/10.1016/j.erap.2005.02.002

Gibson, J.J. (1977). The theory of affordances. In R. Shaw & J.Bransford (Eds.), *Perceiving, acting, and knowing: toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Lawrence Erlbaum.

Ginsberg, F., Rohmer, O., & Louvet, E. (2012). Priming of disability and elderly stereotype in motor performance: similar or specific effects? *Perceptual and Motor Skills*, 114(2), 397-406. https://doi.org/10.2466/07.17.Pms.114.2.397-406

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, 26(2), 309-320. https://doi.org/10.1037/h0034436

Hagiwara, N., Kashy, D., & Cesario, J. (2012). The Independent Effects of Skin Tone and Facial Features on Whites' Affective Reactions to Blacks. *Journal of Experimental Social Psychology*, 48, 892. https://doi.org/10.1016/j.jesp.2012.02.001

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546-573. https://doi.org/10.1177/1745691616652873

James, W. (1890). *Principles of psychology*. New York: Holt.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524-532. https://doi.org/10.1177/0956797611430953

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes: Philosophical Papers* (J. Worrall & G. Currie, Eds. Vol. 1). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511621123

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, 113(2), 254-261. https://doi.org/10.1037/pspi0000106

Loersch, C., & Payne, B. K. (2011). The Situated Inference Model: An Integrative Account of the Effects of Primes on Perception, Behavior, and Motivation. *Perspectives on Psychological Science*, 6(3), 234-252. https://doi.org/10.1177/1745691611406921

Longino, H. E. (2022). What's Social about Social Epistemology? *Journal of Philosophy*, 119(4), 169-195.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, 34(2), 103-115. http://www.jstor.org/stable/186099

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(4), 806-834. https://doi.org/10.1037/0022-006X.46.4.806 (Methodology in Clinical Research)

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108-141. https://doi.org/10.1207/s15327965pli0102_1

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biol*, 18(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596-1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 1-8. http://search.ebscohost.com.proxy-ub.rug.nl/login.aspx?direct=true&db=psyh&AN=2015-40514-005&site=ehost-live&scope=site

Oreskes, N. (2019). Perspectives from the History and Philosophy of Science. In *Why Trust Science?* (pp. 15-68). Princeton University Press. http://www.jstor.org.proxy-ub.rug.nl/stable/j.ctvfjczxx.5

Oude Maatman, F. (2021, July 12). Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It. *PsyArXiv.* https://doi.org/10.31234/osf.io/puqvs

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531-536. https://doi.org/10.1177/1745691612463401

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6), 528-530. https://doi.org/10.1177/1745691612465253

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Basic Books. http://search.ebscohost.com.proxy-ub.rug.nl/login.aspx?direct=true&db=psyh&AN=1961-02882-000&site=ehost-live&scope=site

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science*, 16(4), 725-743. https://doi.org/10.1177/1745691620974697

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

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(2), 90-100. https://doi.org/10.1037/a0015108

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Stanford, K., (2023). Underdetermination of Scientific Theory. In E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4), 717-724. https://doi.org/10.1177/1745691620966796

Uygun-Tunç, D., & Necip Tunç, M. (Accepted/In press). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. *Meta-Psychology*, *XX*(X). https://doi.org/10.31234/osf.io/pdm7y

Yong, E. (2012). A failed replication attempt draws a scathing personal attack from a psychology professor. *National Geographic*. Retrieved from https://www.nationalgeographic.com/science/article/failed-replication-bargh-psychology-study-doyen

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E. M., Bulnes, L. C., Caldwell, T. L.,

Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., . . . Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917-928. https://doi.org/10.1177/1745691616674458

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41. https://doi.org/10.1017/S0140525X17001972