Meta-Psychology, 2025, vol 9, MP.2024.4224
https://doi.org/10.15626/MP.2024.4224
Article type: Commentary
Published under the CC-BY4.0 license

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

Edited by: Rickard Carlsson
Reviewed by: Eric Hochstein, Brian Haig
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/2YNGV

Psychology's Theory Crisis, and why formal modelling cannot solve it

Freek J.W. Oude Maatman
Department of Philosophy of Behavioural Science, Faculty of Social Science, Radboud University Nijmegen, the Netherlands
Department of Theoretical Philosophy, Faculty of Philosophy, University of Groningen, the Netherlands

In light of psychology's 'theory crisis', multiple authors have argued that adopting formalization and/or formal modelling would constitute a useful or even necessary step towards stronger psychological theory. In this article, I instead argue that formal modelling cannot solve the core problem the psychological 'theory crisis' refers to, which are the currently high degrees of contrastive and holistic underdetermination of our theories by our data. I do so by first introducing underdetermination as an explanatory framework for determining the evidential import of research findings for theories, and showing how both broader theoretical considerations and informal assumptions are key to this process. Then, I derive the aforementioned core problem from the current 'theory crisis' literature and tentatively explore its possible solutions. Lastly, I show that formal modelling is neither a necessary nor sufficient solution for either contrastive or holistic underdetermination, and that its uncritical adoption might instead worsen the crisis.

Keywords: theory, theory crisis, formal modelling, underdetermination, evidence

Introduction

Until 2018, discussion of psychology's replication crisis appeared to mostly concern issues of incentives, statistics and methodology, such as publication bias, pressure to publish, QRP's, p-hacking and HARKing (e.g., Nelson et al., 2018; Shrout and Rodgers, 2018). Since then the spotlight appears to have turned: the locus of attention now is on 'new' and unresolved issues with our measurement practices (e.g., Chester and Lasko, 2021; Flake and Fried, 2020, but also see Bringmann and Eronen, 2016) and theories (e.g., Proulx and Morey, 2021), which are argued to also contribute to our field's replicability problems.

Especially this former discussion of psychology's theories has attracted a lot of attention after the coinage of psychology's theory crisis (see Oberauer and Lewandowsky, 2019, but also e.g. Borsboom et al., 2021; Robinaugh et al., 2021). Oberauer and Lewandowsky (2019) argued that our theories are too 'weak': they do not strongly imply hypotheses, by which is meant that a hypothesis (e.g., a statistical hypothesis, an experimental hypothesis) can only be deductively derived from our theories when adding a large amount of (often unevaluated) auxiliary assumptions about research design, statistics or measurement techniques (see e.g., Earp and Trafimow, 2015; Meehl, 1990b). Such theories hereby inhibit their own testability: this lack of connection between theory and hypothesis entails that a failure to find predicted effects does not necessarily provide evidence against the theory, but is instead often taken to be a violation of auxiliary assumptions (Oberauer & Lewandowsky, 2019, p.1598). Psychological theories and hypotheses currently thus suffer from poor testability, and in turn, this allows almost all psychological theories to persist whilst also providing no incentive to improve upon them. Yet, such 'weak' theories also hinder our field's theoretical progress – for example by hindering our ability to create robust interventions due to our inability to identify the conditions in which they ought (not) to work (Borsboom et al., 2021).

Though some authors writing about psychology's problems with theory generally appear to agree with Oberauer and Lewandowsky's (2019) framing (e.g., Fried, 2020a; Robinaugh et al., 2021), others diverge significantly. Van Rooij and Baggio (2020, 2021) for example explicitly reject both (the priority of) the aforementioned testing problem and its framing as a 'crisis'. Instead, they identify an obsession with effects over understanding as a problem in psychology, and focus more on the need for theory to lead to a priori plausible explanations than on theory's relevance for testing. Other authors largely forego specifying problems with psychological theory, and primarily posit methods for

improvement (e.g., Guest and Martin, 2021; Smaldino, 2017, 2020), calling into question what they take to be the problem. Despite this lack of agreement and other dissimilarities, most authors in this new literature nevertheless converge on the same solution for the discipline's theoretical problems: psychologists need to adopt formal and/or computational modelling, be it as a useful method in general (Fried, 2020a, 2020b; Guest & Martin, 2021; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Robinaugh et al., 2021; Scheel et al., 2021; Smaldino, 2019, 2020) or as a necessary step in a proposed theory construction method (e.g., Borsboom et al., 2021; Haslbeck et al., 2022).

Importantly, computational and formal modelling are not identical. By formal modelling, proponents generally refer to the translation of verbal, narrative theories into formal models (sometimes also called 'formal theories'; e.g. Haslbeck et al., 2022) – be they mathematical, computational or logical in form. A good example of this is the formalization of existing theory on panic disorder by Robinaugh et al. (2024). In turn, computational modelling refers either strictly to the formalization of a theory through a programming language (Guest & Martin, 2021; Smaldino, 2017, 2020), or to combining the former with a paradigmatic computationalist approach to cognition in which the locus of psychological investigation is the decomposition of human capacities (i.e., abilities) into algorithms and their implementations in the brain (Marr, 1982; Van Rooij and Baggio, 2021; also see Piccinini, 2009 for an overview). Notably, the latter interpretation of computational modelling thus not only consists of a method or process but also requires taking highly specific paradigmatic assumptions on board, and thus is better defined as an approach.

This divergence calls for some preliminary definitions. By 'formal modelling' I strictly refer to the formalization of verbal theories into formal models (and the iterative development of such models) as a method or process in itself – broader approaches to or interpretations of modelling thus are not to be included. Here, 'theory' refers to a logically coherent set of propositions aimed at the explanation of one or several phenomena. In turn, 'formal model' refers to any mathematical, logical or other formalization of a theory's structure and 'phenomenon' refers to a feature of the world that scientists seek to explain, be it an ability, process, event type, data pattern or other 'explanandum'. The former type of computational modelling – recasting theory into a programming language – then is a sub-category of formal modelling, for which reason I will use the latter as an umbrella term covering both instances in the remainder of this paper.

Generally, I agree with the aforementioned authors that formal and/or computational modelling in principle would be a valuable addition to any psychologist's methodological toolbox; both as an aid for thinking and as a non-experimental method for theory evaluation (e.g., Farrell and Lewandowsky, 2010; Smaldino, 2017). Yet, the argument for formal modelling is sometimes accompanied by more or less explicit rejections of verbal theories and theorizing (i.e., theory construction and evaluation on the basis of natural language properties) as if these are in competition with 'formal theorizing', such as by labeling the former as 'proto-theories' (Borsboom et al., 2021) or arguing that natural language is inherently ambiguous and thus best replaced with formal accounts (Fried, 2020a, 2020b; Haslbeck et al., 2022; Robinaugh et al., 2021). Some even go so far as to suggest that adoption of formal modelling is the only way to advance our field (Smaldino, 2017, 2020).

In contrast to such claims, I do not believe that formal modelling or formalization are the – and perhaps not even a – way forward for many psychological sub-fields in their current state, nor that (informal) verbal theorizing and theories are inherently problematic. Eronen and colleagues for example have already pointed out that formalization may be premature for many fields in psychology due to a lack of well-defined constructs, robust phenomena and valid measurement techniques, which according to them leads to the necessity to first engage in good (verbal) conceptualization in order to kickstart a process of epistemic iteration of good coordinative definitions and functions (e.g., Eronen and Bringmann, 2021; Eronen and Romeijn, 2020; see also Chang, 2004; Kellen et al., 2021). Similarly, some proponents of formal modelling tentatively argue that it is not applicable to all psychological sub-fields (Borsboom et al., 2021), especially since formal modelling is primarily applicable to isolated 'toy worlds' (Navarro, 2021). Critical evaluation of both the readiness of psychological fields for formalization and the general applicability of formalization to these fields thus is warranted.

My point however goes further than these claims: I believe that formal modelling is neither a necessary, nor a sufficient solution to the 'theory crisis' (i.e., the previously outlined problems with testability). Furthermore, I argue that adopting formal modelling can worsen this crisis if (and only if) not accompanied or preceded by strong verbal theorizing, as this is necessary to design informative experimental tests. Formal modelling should thus not be uncritically adopted or applied by psychologists as a replacement or even 'remedy' for verbal theorizing, if their goal is to improve the testability

of their theories and the evidential value they derive from these tests.

In the remainder of this paper I will support these claims. I will do so by first identifying what problem with theory the currently discussed 'theory crisis' exactly refers to, which I will do by introducing underdetermination of theory by data as an explanatory framework for the evidential import of findings for theories, and deriving the concept of degrees of underdetermination from its application. Using this framework, I show that the current interpretation of the 'theory crisis' is grounded in the observation of high degrees of contrastive and holistic underdetermination of psychological theories by our data, after which I tentatively identify the cause of this crisis as well as a possible path for its resolution. Then, I argue that given this analysis the adoption of formal modelling cannot solve this particular crisis in isolation of complementary advances in 'informal' verbal theory.

Underdetermination and the theory crisis

Even though I stated above that the 'theory crisis' refers to the often weak inferential link between psychological theories and hypotheses (see Fried, 2020a; Oberauer and Lewandowsky, 2019; Robinaugh et al., 2021), there is no single agreed upon formulation of this so-called 'crisis'. Furthermore, it remains unclear what particular (missing) property of psychological theories causes such weak inferential links. Proponents of formal modelling as a solution to this 'theory crisis' (i.e., Fried, 2020a, 2020b; Oberauer and Lewandowsky, 2019; Robinaugh et al., 2021) so far only specify the effects of weak theory (e.g., insufficiently informing hypotheses, no derivable boundary conditions), but do not clearly converge on a cause for such problems in terms of theoretical issues. The perceived lack of formalization cannot be a theoretical problem in itself either, since a formalization of a weak theory is not inherently stronger (although we may strengthen the theory in the formalization process). We are thus faced with a crisis declaration of psychological theories failing to do what they ought to as well as a proposal for resolving this situation, without clarity on what the exact failure is and which properties of theories cause it. Subsequently, the proposed remedy – formal or computational modelling – might not be a cure at all.

Yet, the authors that explicitly frame formal modelling as a solution to psychology's weak inferential links between hypotheses and theories (i.e., Fried, 2020a, 2020b; Oberauer and Lewandowsky, 2019; Robinaugh et al., 2021) do offer several supporting arguments for their view from which the core problem of this 'theory crisis' can be derived. Instead of discussing their arguments one by one, I will describe their views after introducing the problem of underdetermination of theory by data (see Stanford, 2017 for an overview) and my analysis of the theory crisis in its terms, since many of their arguments are either directly related to or even derivable from it.

Notably, several authors have discussed underdetermination in relation to the replication crisis and psychology before (e.g., Earp and Trafimow, 2015; Meehl, 1978, 1990a, 1990b; Trafimow and Earp, 2016; Tunç and Tunç, 2023). My account here is not an attempt to improve on their specific analyses, but it does expand upon them to draw attention to aspects of underdetermination that are relevant for this 'crisis of theory'. Furthermore, my goal is not to argue in favor of any particular interpretation or framing of underdetermination, but simply to use it as an organizing framework to think about the evidential import of research findings for theories and hypotheses.

Underdetermination of theory by data

When we speak of 'underdetermination of theory by data', we refer to the fact that our research results do not completely constrain what we should believe: upon being confronted with a particular research finding, there is never a straight-forward answer to what (theoretical) beliefs we should hold in response to it (Stanford, 2017). This underdetermination stems from two sources. The first is that a (set of) successfully predicted finding(s) does not simply entail that a theory is true. After all, it is always possible to formulate an alternative theory that predicts the same (set of) finding(s) – be it a wholly different theory, a mere variant, or simply a more precise or bounded form of the theory under test. Due to the persistent presence and potentially large size of this set of theoretical alternatives, a successfully predicted result cannot directly justify our choice for the theory it was derived from. This is called contrastive underdetermination (e.g., Stanford, 2017): any successful experiment is insufficient to justify our choice for one specific theory, as there always will be (unidentified) others that predict the same outcome. Whilst this type of underdetermination in principle always applies, the relative size of the set of potential viable alternatives differs between hypotheses. For example, compare only predicting any positive correlation between two variables, which only excludes all possible alternative theories that predict no or a negative relation from the set of alternatives upon success, to a successful highly precise prediction, which can only be matched by a comparatively much smaller subset of all possible theories.

The second source of underdetermination stems from

the fact that we never test a theory-derived hypothesis H in isolation: to relate the theory-derived hypothesis H (i.e., 'X should be the case') to the empirical context in which we do research, we always need a set of further assumptions – also known as auxiliary assumptions – that guarantee that the experiment and subsequent analyses constitute a valid test of H. These include, but are not limited to, assumptions about the validity of all used measurement techniques, the successful execution of all elements of the experimental design and its appropriateness as a test of the theory (see also Kellen et al., 2021), the absence of influence from any unexpected and/or unknown but nevertheless causally effective factors that can change the experimental outcome in a test-relevant way (also known as the 'ceteris paribus clause'; C_p), the non-violation of assumptions of used (statistical) analyses, and the representativeness and sufficient size of our sample (see e.g., Meehl, 1990a, 1990b). In turn, any experiment is not a direct test of the theory-derived hypothesis or theory itself, but a test of the conjunction of the hypothesis with all these assumptions: the test tests not just whether the hypothesis is true but whether the hypothesis and all auxiliary assumptions are true. Only the truth of all auxiliary assumptions and the hypothesis logically entails the prediction. This issue is also known as the problem of holistic underdetermination (Stanford, 2017).

Why these auxiliary assumptions are necessary becomes explicit when confronted with a failed prediction or a non-replication. Instead of showing that the hypothesis H (and by extension the theory T) is false, such a failure to find the expected results shows that the conjunction of H with all auxiliary assumptions is false. This means that this failure might be caused by the falsity of one or more auxiliary assumptions even if the hypothesis and/or theory are true, and it is therefore logically possible to blame such failure on any or multiple of the auxiliaries instead of on the hypothesis or theory. Such cases are also not hard to imagine: a research assistant instructing the participants might for example have deviated heavily from the study protocol and biased the participants in the wrong (or 'right') direction, one of the measurement techniques might be invalid, or the used analysis might be inappropriate for the data type and therefore lead to false results, or the new research context may have contained a new, unknown confound, and so on (see also Earp and Trafimow, 2015; Iso-Ahola, 2017; Meehl, 1990a; Robinaugh et al., 2021; Stroebe and Strack, 2014).

If any of these auxiliary assumptions is violated, the findings simply are uninformative for our evaluation of H or T – regardless of whether our prediction is confirmed or not. This is also not a matter of degree: if, for example, our measurement technique is in fact invalid, our findings should not have any bearing on our evaluation of H or T. After all, if I attempt to measure the brightness of stars in the sky by looking at the Christmas lights hanging from my ceiling, this finding simply should not be taken into account by astronomers due to the blatant violation of measurement validity assumptions (i.e., my visual inspection of Christmas lights is not a valid measurement of star brightness) and design assumptions (i.e., my Christmas lights are not part of the population I want to investigate) even if I somehow end up with the correct answer (see also Trafimow, 2017). Combining the latter with the fact that there are often many such assumptions involved in an experiment, it means that upon a failed prediction it is generally possible to blame this on the violation of an auxiliary assumption instead of falsifying the hypothesis or theory itself (a.k.a. a 'Lakatosian defense'; Meehl, 1990a; but also see Earp and Trafimow, 2015).

An important consequence of this is that while it is not always possible to direct blame towards the known auxiliary assumptions (e.g., measurement techniques are validated, manipulation checks were successful, we controlled for known confounds), it is always possible to reject the ceteris paribus clause (C_p). After all, we can always posit that a (previously) unknown confound or 'hidden moderator' has confounded the results (see e.g., Van Bavel et al., 2016), such as an unknown contextual factor or a difference in experimenter skills. Holistic underdetermination can thus seriously, though legitimately, complicate the interpretation of both original findings and (non-)replications[1].

I cannot stress enough here that the aforementioned methodological and statistical auxiliary assumptions are not all assumptions that are relevant to holistic underdetermination, even though they receive the most attention in psychological practice. In principle, all of our background knowledge and tacit assumptions can play a role in the evaluation and construction of our theories and experiments due to either being indirectly in-

[1]This also identifies an issue in the current replication crisis literature, where some authors propose testing such alternative explanations (e.g., Nelson et al., 2018; Stroebe and Strack, 2014; Tunç and Tunç, 2023). Since in psychology replication attempts can never be completely identical, any possible difference might be identified as 'relevant' post-hoc. Yet, notably this also goes for comparisons between replications themselves. Any testing strategy for combatting holistic underdetermination can thus easily spiral out of control into a so-called experimenter's regress (Collins, 1985; also see Morawski, 2019): a possibly never-ending spiral of replications and auxiliary assumption-validating experiments aimed at verifying the adequacy of previous experiments, and replications of these assumption-testing experiments themselves.

corporated in them, or possibly being contradicted by them (Quine, 1951). Theory itself and possibly atheoretical methodological choices (e.g., using chocolate as a reward, assuming that everyone likes chocolate) also can introduce further auxiliary assumptions into our experiments. For example, our choices of theory and disciplinary matrix (Kuhn, 1962) – or our 'theoretical framework' (Guest & Martin, 2021; Muthukrishna & Henrich, 2019) – as well as our conceptualization of key variables will impose far more (tacit) assumptions on experiments than the general methodology of psychology can. Meanwhile, whether we pay attention to it or not, the aforementioned 'background theory' also limits the possible auxiliaries we can assume for reasons of coherence.

The most important example of the latter claim of limitations on auxiliaries stemming from 'paradigmatic' influence are our (implicit) metaphysical positions (Hochstein, 2019), which I prefer to call our ontological commitments (i.e., commitments about the structure of the world). Ontological commitments relevant to psychology range from our fundamental ideas about the nature of human cognition to what we take to be the real-world referents or realizers of our theory's concepts, constructs or implied processes. Whilst not directly testable themselves, such commitments form the basis for any of our investigations – be it explicitly or implicitly – by informing not only what we might consider the things we investigate to be, but also how we ought to study them (Hochstein, 2019).

A simple example can be found in the definition of 'stress'. Whether we define 'stress' as a particular biological state (i.e., parasympathic nervous system activation), a particular mental state (i.e., subjective experience of stress), or a combination of these has direct bearing on the to-be assumed reliability and validity of, say, cortisol levels as a measurement of stress. After all, it is a direct measurement on the biological account, but an indirect measurement at best under the mental definition (if even a proper measurement at all). This definition also indirectly suggests mechanisms for intervention, and for confounding: direct cortisol injections could be an unconfounded intervention on stress on a biological account, but might be confounded by strong alternative competing mental states on a mental account of stress. In turn, this also partially sets how we should design our experiments such that they are free from any confounding influences, and thus how we would interpret experimental results: those supporting a mental account of stress might find cortisol injections an unsatisfactory induction of stress, as they require elimination of alternative mental states too.

Though often distal to our research questions, our base ontological assumptions about the nature of human cognition can also deeply influence our experiments. A most basic, common example is that we take the brain to be 'the seat' of the mind in some way, shape or form – but that without this commitment, neuroscience as a whole would cease to make sense, as well as reference to neuroscientific findings. Furthermore, allegiance to a paradigmatic position such as (radical) embodied cognition, radical behaviorism or computationalism can directly constrain the type of entities and mental processes we could use in psychological theory to begin with, by for example prohibiting talk of 'representations' (embodied cognition; see e.g. Shapiro and Spaulding, 2021), or requiring all processes to be specifiable as input-output algorithms (computationalism; see e.g. Piccinini, 2009) or as forms of behavior (radical behaviorism; see e.g., Skinner, 1953). Similarly, the complex systems approach's assumptions that the causal structure of cognition is interaction-dominant instead of component-dominant (Van Geert, 2019; Wallot & Kelty-Stephen, 2018) directly problematizes most existing psychological theories by effectively excluding the possibility of simple causal structures and isolatable entities or mechanisms. Such fundamental ontological commitments in turn also constrain how we conceive of psychological phenomena such as thought; e.g., as capacities that can be described as input-output algorithms (computationalist; Van Rooij and Baggio, 2020, 2021) or as processes instantiated in behavior that is inherently intertwined with its context (complex systems approach; Van Geert, 2019).

From here, these commitments also 'trickle down' into the mechanisms we propose due to for example assuming more or less context-dependence for processes, or requiring a particular type of mechanism. In turn, these too influence *which* confounds we would identify whilst remaining coherent with our background ideas – and resultantly *could*. Notably, they can also directly influence which (statistical) analysis methods are applicable; whether we assume that the phenomenon we study can be taken to be homogeneous across individuals or not – which varies between ontologies – greatly influences the usability of common research techniques, such as population aggregate statistics (see e.g., Molenaar, 2004; Richters, 2021). Ontological commitments thus provide constraints on nearly all other auxiliary assumptions, even basic methodological ones, by specifying the properties and boundaries of the phenomena we study, and the resultant requirements for studying them intelligibly.

Another relevant source of auxiliary assumptions in psychology was recently identified by Brenninkmeijer et al. (2019), who showed that experimental psychol-

ogists can significantly diverge in how they design their experimental materials, conduct themselves with their research subjects and try to motivate them. The psychologists in question concluded this might also affect their research results, which in fact has been observed by Landy et al. (2020). The differing approaches of experimental psychologists thus may form a key type of tacit knowledge involved in psychological experimental design. This can indirectly bring an individual's own 'tacit assumptions' into the mix of relevant auxiliaries: assumptions about the manner in which an experiment has been (or should be) conducted which are not directly methodologically or otherwise describable. Another example of such 'tacit assumptions' can be found in face validity-informed construction/evaluation of experimental setups and measurement instruments, for which researchers often simulate their own participation and interpretations of stimuli/questions. In turn, fit with 'personal simulations' or 'techniques' can also enter as a tacit auxiliary assumption about experimental design, and thus in principle can be used to dismiss certain research designs and findings by individuals.

Two types of 'theory'

Two implications of the above analysis need to be made explicit here. First, the above examples show that given different ontological commitments, tacit knowledge or other elements of our background theory, we might draw different conclusions about the evidential relevance of the same results for the same theories. Notably, changing any of these elements also can force changes in our evaluation of all results we previously believed to constitute evidence for or against our views.

Second, this analysis shows that the generally used definition of theory in this literature – that is, 'a logically coherent set of propositions aimed at the explanation of one or several phenomena' – does not cover many relevant and even highly 'theoretical' assumptions involved in our experimental practice (e.g., ontological commitments, statistical assumptions, measurement theory). Instead, the analysis so far shows us that any such theory aimed at the explanation of a single phenomenon can only be tested given a set of auxiliary assumptions drawn from many other theories (see also Teo, 2020 for a different angle leading to the same conclusion) as well as paradigmatic, personal and non-theoretical background knowledge or beliefs (Kuhn, 1962; Quine, 1951). This latter set is not exactly the same as the set of theory and all auxiliary assumptions directly relevant to a single phenomenon and its measurement, as certain auxiliary assumptions themselves may be derived from further background theory or theories, or yet further sources (e.g., theories about space-time).

A definitional segue therefore must be made here in order to distinguish the theory under test from the larger background of 'theory' in which it and any experiment testing its derived hypotheses are embedded, and from which its evaluation partially occurs. From this point on, I will use 'theory$_1$' when talking about theory in the former, narrow sense, as I have been doing so far: specific theories aimed at the explanation of one or more phenomena, such as 'attachment theory' or 'ego depletion theory', which can be more-or-less easily distinguished from auxiliary assumptions about, say, measurement validity. I will instead use 'theory$_2$' when talking about theory in the latter, broad sense of theory implied by holistic underdetermination: the entire set of theories and (tacit) assumptions directly or indirectly involved in the formulation and derivation of a hypothesis from a theory$_1$, and the design of an experiment capable of testing it[2]. Notably, this means theory$_2$ can include other theories$_1$ that are not under test. I will reserve 'theory' for cases in which either interpretation fits.

Theory$_2$ encompasses a large variety of possibly relevant 'background beliefs', which might vary per experiment and experimenter. This makes it hard to pin down. For most psychological experiments, it will range from more basic statistical and methodological assumptions to the aforementioned ontological commitments we hold. It also will likely cover our (personal or disciplinary) interpretations of key scientific concepts such as causality, correlation and validity. It also covers any indirectly or directly relevant knowledge and theories from other fields (e.g., about the nature of space-time), and most likely several 'tacit' assumptions derived from previous experimental experience and our assumptions about a participant's engagement with experimental tasks.

Degrees of underdetermination and how to decrease them

Let us now return to the relevance of our discussion of underdetermination for the psychological 'theory crisis'. How is underdetermination related to the weak link between psychological theories and experimental hypotheses? In order to show this, I first need to introduce a new concept: degrees of underdetermination. On the basis of the previous two sections, I argue that findings can underdetermine theories more or less. The higher the absolute number of theories – including unknown

[2]Notably, theory$_2$ is closer to Quine's conception of the role of knowledge in science (Quine, 1951), as well as encompassing the background knowledge on which we draw in order to abduce theories$_1$ (e.g. Haslbeck et al., 2022).

ones – that is able to account for the same (successful) finding to the same degree of accuracy, the higher the degree of contrastive underdetermination. The more auxiliary assumptions we could legitimately falsify instead of falsifying a hypothesis after an unsuccessful finding, the higher the degree of holistic underdetermination. These two definitions can be more concretely translated as the absolute amount of viable alternative explanations we could respectively formulate for successful and unsuccessful results given complete awareness of all alternatives: the more possibilities we have in either case, the more underdetermined our theory will be by any eventual finding (see also Fiedler et al., 2012). In practice, the degree of holistic underdetermination also can be considered the inverse of the degree to which all types of study-element related validity (e.g., construct, internal) have been established (see also Fabrigar et al., 2020), though it must be stressed here that the degree to which we can in fact establish such validity is dependent on the specificity of our theory_1 and (background) theory_2 – more on which will follow below.

From these definitions, we can also directly derive methods to decrease underdetermination. To decrease the degree of contrastive underdetermination we run into, we can try to make predictions that only would follow from a smaller subset of possible theories[3]. The most common, and perhaps easiest way to do this is to formulate predictions that are very precise or otherwise very unlikely without the theory, which is also known as a risky testing approach (see Meehl, 1978; Popper, 1959). Upon (a series of) predictive success(es), such a theory would only have to compete with a much smaller set of possible alternative explanations of the same finding(s) due to the fact that few alternative theories can predict the same exact outcome(s) (see also Meehl, 1990a). In turn, this means that such a successful unlikely or risky prediction is both more informative than a general prediction (i.e., we learn more) and provides more corroboration for the investigated theory relative to the set of all alternative theories than a 'safe' prediction that is compatible with many theories would.

Similarly, we can decrease the degree of holistic underdetermination by ensuring the validity of all of our auxiliary assumptions, and thereby reducing the chance we discover post-hoc that they are violated. In the philosophical literature, the approach closest to this technique is best formulated in Deborah Mayo's severity principle: a research finding only provides evidence for a hypothesis H if – and only if – this finding results from a test procedure which taken as a whole would have uncovered the falsity of or discrepancies from H if H were false (Mayo and Spanos, 2009, p. 32; see also Mayo, 2018).

The ability of a test procedure to in fact identify flaws in H depends not just on whether we (un)intendedly skew the data and statistical analyses in our theory's favor (e.g., through forgery or QRPs; see Mayo, 2018, p. 40-41), but also on the validity of our auxiliary assumptions. If we are uncertain about some of our auxiliary assumptions, the test procedure might not be able to uncover whether H is false, after all: if any auxiliary assumption is violated, the finding may simply be irrelevant for H. We can thus reduce the degree of holistic underdetermination by for example validating our measurement instruments in advance, using manipulation checks, ensuring that our design is carried out correctly and does not violate our theoretical assumptions, and ensuring that we use appropriate statistical techniques and have an adequate sample size (Mayo, 2018, p. 75-115). To further enhance severity in this particular form we can specify our theory to such a degree that it identifies which factors can possibly influence the phenomenon we study and, more importantly, which cannot. This not only allows us to design studies that control for the known confounds, but also to treat any remaining violation of the C_p-clause as a violation of our theory – because if there was another 'unknown' confound or 'hidden moderator', our theory is apparently incomplete. Although this level of specification is difficult to achieve, it in effect allows us to a priori specify the boundary conditions implied by our theory, whilst simultaneously decreasing the degree of contrastive underdetermination by excluding alternative theories that imply different boundary conditions for the similar predictions. Through this method, we can also indirectly test our ontological commitments, given that different ontological commitments would imply different boundary conditions down the line.

[3] A reviewer remarked that contrastive underdetermination in principle could be reduced by examining the explanatory coherence of a theory (cf. Maier et al., 2024). This is true if we assess how well a set of known theories explains or predicts a particular set of data, as this allows us to identify which theories are incompatible with the full range of gathered data and thus can be discounted from consideration. Yet, knowing that any current theory has good or even the best possible explanatory coherence with the data does not reduce the possibility that there are unconceived alternatives with the same or better coherence. After all, note that explanatory coherence is bounded by the set of known (to be) relevant data. Since novel theories can help identify novel measuring techniques, relevant phenomena and data points, our current best theory might turn out less and less coherent over time due to failing to explain the novel evidence. Explanatory coherence thus at the very best is an unreliable guide to contrastive underdetermination, if not potentially irrelevant to it.

Furthermore, we can develop our theory$_2$-based background assumptions and make these explicit, ensuring that there is no incoherence of these with our theory and/or experiment, as well as use these to further inform the strength of auxiliaries. Such severe testing then minimizes the amount of auxiliary assumptions that can function as possible 'scapegoats' when a prediction fails, and thereby allows us to more easily direct the blame towards the hypothesis, and by extension the theory it was derived from. Taken together, underdetermination of theory by data as a whole then can be relatively deflated through engaging in both risky and severe testing.

What is the theory crisis, and what is its cause?

It is now time to return our attention to the theory crisis: how does underdetermination relate to this? As mentioned in the previous section, underdetermination covers many of the problems that are argued to be at stake there, which becomes clear when we take a closer look at the relevant literature. In this section I will therefore first analyze the arguments in the theory crisis literature in relation to underdetermination, and argue that these indicate that the theory crisis refers to the currently high degree of underdetermination of our theories by our data. Then, I switch attention to the cause of this high degree of underdetermination in psychology.

Oberauer and Lewandowsky's (2019) view of the theory crisis is explicit throughout the opening of their paper (p. 1596-98): in order to deductively derive a hypothesis (defined by them as 'the assumption that an empirical generalization holds', p. 1597) from our theories, we currently often need to (tacitly) assume many different propositions that are not implied by our theories. Furthermore, if our hypothesis is not confirmed by our data, in the case of 'weak theory' there are many possible explanations for this failure that protect our theory from being affected. This occurs either because the theory does not strictly require the hypothesis to obtain, or by falsifying one of these auxiliary assumptions instead of our theory (p. 1598). This problem is echoed by Fried (2020a) when he states that the weaker the theory, "the more auxiliary scape goats can be blamed if the theory does not explain or predict well" (Fried, 2020a, p. 4). In other words, Fried (2020a) and Oberauer and Lewandowsky (2019) are worried about the high degree of holistic underdetermination that occurs in the testing of our current theories: we have many auxiliary 'scapegoats' due to their lacking specification and validation, and therefore many alternative explanations can be given for failed predictions.

Interestingly, Oberauer and Lewandowsky (2019) also indirectly mention the problem of contrastive underdetermination when they state that an empirical hypothesis derived from a theory needs to be specific to this theory in order to be diagnostic for it (i.e., the probability of the finding given that the theory is false, $P(X|\neg T)$, must be low), which echoes the risky testing approach described above. This claim is not an explicit part of their critique, but forms a key component of the amount of evidence for the theory we can draw from the experiment upon a successful finding in their analysis. Notably, $P(X|\neg T)$ can also be lowered by formal modelling due to the possibility to draw precise point predictions from formal mathematical models, as for example pointed out by Borsboom et al. (2021), Fried (2020a), and Robinaugh et al. (2021).

Fried (2020a) and Robinaugh et al. (2021) in turn explicitly derive both their critique of our current theories and their suggestion to adopt formalization from the work of Paul Meehl (1978; but also see 1990a, 1990b). Meehl in fact discussed underdetermination in his work, thereby intending to show that the failure of our theories and research practice is multiple. He for example argued that our current testing strategy (i.e., testing against null effects and/or an opposite directional effect) is inadequate to actually provide evidence for theories due to the high $P(X|\neg T)$ associated with predictions that 'there is a (directional) effect': a successful finding corroborates the entire set of theories that predict such a (directional) effect (Meehl, 1990a, p.123), which we have already identified as a case of high degrees of contrastive underdetermination. Besides this, he also argued that our auxiliary assumptions are often completely uncorroborated or even missing, and that we often mistake statistical hypotheses for substantive theories (Meehl, 1990a, p. 110-2). In other words: Meehl argued that, respectively, the degrees of contrastive and holistic underdetermination are very high in psychology, heavily limiting what we can learn from our studies.

Robinaugh et al. (2021) primarily discuss Meehl's former argument about contrastive underdetermination as well as our mistaken usage of statistical hypotheses as theories (also covered by Fried, 2020a), as well as Meehl's preferred solution to it: risky testing through point predictions derived from formal models and consistency testing (see Meehl, 1978, 1990a, 1990b). This method relies on the low odds of a highly precise prediction actually obtaining given any failure of an auxiliary assumption, causing a successful finding to provide support to the entire conjunction of H and all auxiliary assumptions whilst also eliminating contrastive underdetermination by being highly precise. Yet, they also indirectly touch upon the issues with holistic underdetermination by pointing out the necessity of strong aux-

iliary assumptions, especially for measurement. They later however delimit this to formal auxiliary assumptions about the hypothesized statistical and mathematical properties of the phenomenon under study and as well as interpretation of its measurements, thereby implicitly leaving out a large swathe of auxiliary assumptions. Nevertheless, Robinaugh et al.'s (2021) critique of psychological theories can be argued to boil down to the high degree of contrastive and holistic underdetermination we encounter in their testing.

Fried (2020a) is the most specific about what he believes to be wrong with psychological theories. In his critique he largely follows Meehl's lead, due to which he also (in)directly defines 'weak' theories as theories suffering from a large degree of both variants of underdetermination when tested. Besides his aforementioned critique grounded in holistic underdetermination, Fried for example states that weak theories "predict observations also predicted by other theories", which he explicitly identifies as a problem with a high degree of contrastive underdetermination (Fried, 2020a, p. 24). He also adds further elements that point in the direction of underdetermination, such as the suggestion to write 'falsification paragraphs' describing under which (boundary) conditions the hypothesis is supposed to hold and under which it ought not. Such falsification paragraphs would amount to an explicit expansion of our known auxiliary assumptions about experimental design, as well as a relative reduction in the amount of circumstances that could lead to falsification of the C_p-clause. Whilst doing so, Fried (2020a) importantly also points out the double standard inherent in referring to contextual sensitivity of effects in the case of failed predictions (i.e., violations of the C_p-clause or design-related auxiliary assumptions), whilst being unable to properly formulate the conditions under which a 'real' non-replication would take in advance of any testing. This shows that such theories either are incomplete or not entirely explicitly formulated or specified. Fried thus also appears to be referring to high degrees of underdetermination resulting from psychology's current research practices.

To conclude, it appears the theory crisis – as it is currently discussed – can be interpreted as the often high degrees of both contrastive and holistic underdetermination of our theories by our data. In order to offer solutions to this 'theory crisis', we however also require a clear specification of the cause of this situation. What causes the currently high degrees of underdetermination in psychological science? The analysis so far can offer us a preliminary answer, which is that current psychological theory does not inform our research designs, instruments, and other methodological choices enough to create conditions for either risky or severe testing, or to otherwise reduce the degrees of underdetermination. Yet, this seems a restatement of the problem. Why can current psychological theory not do so?

In my view, the above can be best explained by our theories$_1$ not being specific (see also Hagger et al., 2017; Klein, 2014; Oberauer and Lewandowsky, 2019), which I mean to use here in all its senses: precision, completeness, explicitness and clarity. A specific theory$_1$ is explicit about 'what' it explains by providing clear definitions of the explanandum and the constructs, components and concepts it invokes in both defining and explaining it. This requires ontological commitments concerning the realization of the explanandum within that reality (i.e., what it is materially, what counts as an instantiation of the explanandum and what does not) and other theoretical elements that are assumed to refer to something in the world. It is also explicit about the 'how' of its explanation by giving a complete account of not just the mechanism(s) or other relationships underlying the explanandum (see also Smaldino, 2017; Van Rooij and Baggio, 2021), but also the possible extraneous influences on these (e.g., confounds). This requires not only verbal specificity (e.g., precise reference, clear definition), but also can be supplemented with schematic representations (e.g., mechanistic models such as schematic images of a synaptic cleft; Bechtel and Abrahamsen, 2005) and, indeed, formal modelling. In other words, a specific theory$_1$ implies and informs many of the auxiliary assumptions required for its testing, and thereby also the conditions required for it to be tested (Fried, 2020a; Klein, 2014). A specific theory$_1$ thereby also can inform what Van Rooij and Baggio (2020) label strong tests of qualitative structure: "tests that directly tap into the workings of a system as it exercises the capacity of interest" or, as I would prefer to define this, exhibits the phenomenon of interest.

Such specific theory$_1$ always depends, both by definition and in practice, on explicit theory$_2$ for its development and, as shown previously, testing. As we saw previously, our ontological commitments – part of theory$_2$ – do not only inform our measurement practices, but also constrain what type of entities we can invoke and how we even conceive of psychological phenomena. More basic types of auxiliary assumptions can also be deduced from or implied by the theory$_1$ under test in combination with our theory$_2$, such as auxiliaries about possible confounds as long as a mechanism is specified by our theory$_1$. Given sufficient background knowledge of the system in which this mechanism is embedded (theory$_2$), we can after all directly abduce mechanism-external influences on the mechanism's functioning. If a theory$_1$ offers very little information about this, we can hardly

delimit the space of possible confounding factors for our hypothesized relationships, which in turn causes lacking auxiliary assumptions about these in terms of experimental design. This compromises our ability to design internally valid experiments whilst also inflating the likelihood of the C_p-clause being violated – i.e., the prevalence of 'hidden moderators' is unknown, and thus possibly large.

Theory$_1$ is also relevant to most other auxiliary assumptions related to experimental design: only our theory$_1$ can specify which aspects of the design matter to finding the effect or not (see also Klein, 2014). This is very clear when determining construct validity of measurement techniques: not only has construct validity always been strongly tied to theories from a psychometric view (e.g., Kane, 2001), every single method for establishing construct validity also relies on some type of theoretical reasoning about the to-be-measured construct, such as a specified nomological network of all this construct's relationships (Cronbach & Meehl, 1955), the specification of a causal connection between construct and measurement technique (Borsboom et al., 2004), an argument to the effect that our proposed interpretation of the measurement technique is in fact valid for our intended use (Kane, 2001) or the minimal claim that other constructs are to be positively, negatively or not associated with it in a certain pattern (Campbell & Fiske, 1959). Since whatever construct is (validly) measured by the technique must also be the one the theory$_1$ discusses, this implies that not just the validity of our assumptions about construct validity but also the degree to which we can in fact investigate this are largely theory$_1$-dependent (see also Borsboom et al., 2004; Kane, 2001). Also, note that in the case the measurement is independent from theory$_1$ (e.g., measuring temperature when testing for disease presence), its validity will depend on theory$_2$ nonetheless – as has been shown above.

Specificity thus is a necessity if we are to decrease the degrees of underdetermination we are confronted with. Yet, there is often no such specificity in psychological theory, because our theories currently often offer no clear, ontologically specific conceptions of the phenomena, constructs or processes we study: i.e., it is unclear how exactly they are realized in the world, or we are not committed to any particular realization. This is the case for our theories in both the narrow and broad sense: there is often no clear background paradigm or set of ontological commitments (Kellen et al., 2021; Muthukrishna & Henrich, 2019). The absence of specific theory in terms of ontological commitments about the nature of human cognition and resultant definitions of constructs (e.g., what is a belief, trait or mental process precisely?) deeply worsens contrastive underdetermination by allowing for many alternative underlying mechanisms, whilst worsening holistic underdetermination due to our resulting inability to specify confounds and good measurement techniques. However, currently lacking validation of measurement techniques (see e.g. Chester and Lasko, 2021; Flake and Fried, 2020) and often lacking consensus on and conceptualization of key constructs even in non-ontological terms (Eronen & Bringmann, 2021; Feest, 2019) by themselves already lead to severe problems with the degree of holistic underdetermination we encounter. In turn, we are in the dark when we are faced with our results: do these support or disconfirm our theory, or are they wholly irrelevant?

More concisely put, the lack of specificity means that the results of many of our studies are currently uninformative; they do not constrain what beliefs we should hold in response to them, neither in regard to which theories we should hold nor in regard to whether the findings are actually relevant for the evaluation of these theories. And importantly, we then cannot reliably use our previous findings to inform or help evaluate future research or findings. This underdetermination then is not just a problem for theory$_1$ evaluation, but also caused by a lack of theoretical specificity in the first place[4].

The 'theory crisis' then refers to the high degree of underdetermination we encounter when testing our theories$_1$, due to their lack of specification as well as lacking specification of the background assumptions involved in the creation of hypotheses and their testing (i.e., theory$_2$). Even though this may not be the only problem with psychological theory and theorizing (see e.g., Van Rooij and Baggio, 2021), I want to stress that this nonetheless is a serious problem as highly underdetermined research is an exercise in futility. If we are uncertain about not just the validity of our auxiliary assumptions but also which assumptions we actually hold (e.g., which factors could be considered a possible confound and which not, whether our measurement techniques are actually valid), even successful predictions

[4]Note that this is at odds with a previous analysis by Trafimow and Earp (2016), who, in response to an argument by Klein (2014) similar to mine, argued that not badly specified theories but auxiliary assumptions are the cause of the replication crisis. Although I do not claim that theoretical problems underlie the replication crisis here, I would contend that auxiliary assumptions are not always theory-independent; say, whether a measurement technique can be argued to be valid depends on what we conceive its measurand to be. Without such a conception within the theory, auxiliary assumptions about construct validity will always remain weakly motivated. Well-specified theories in turn will strictly imply the auxiliary assumptions under which they can be tested.

can hardly be argued to provide evidence for the tested theory due to the possibility of assumption violations – if not here, then in the future. Similarly, replication and validation attempts are likely to become both goose chases and money pits due to being similarly underdetermined by theory. And if we engage in tests of hypotheses that could be true under (comparatively) many theories, we do not actually learn much from our experiments. Then, if much of our research indeed is as underdetermined as I and the aforementioned authors argue, the two forms of underdetermination also entail that the path towards a cumulative psychological science currently is blocked, as we are unable to derive much (if any) evidence for our theories from our experimental research. In such a case, the evidential relevance of any highly underdetermined study after all is either low to begin with (high contrastive underdetermination) and/or extremely uncertain (high holistic underdetermination) – meaning that we cannot rely on any findings to inform our theories or knowledge. In other words, this suggests that without strong theories we build research programs on the evidential equivalent of quicksand.

Does formal modelling solve the theory crisis?

Assuming that the 'theory crisis' currently is a serious problem for (parts of) psychology and refers to the high degrees of underdetermination present in the testing of psychological theories$_1$, we can now ask the question whether formal modelling is capable of solving this problem, and if so, to which extent. I will approach this question by first identifying the main arguments in favor of formal modelling, after which I respectively discuss its effects for contrastive and holistic underdetermination. During this discussion, I argue that formal modelling is neither sufficient nor necessary for the reduction of either degree of underdetermination in psychological science, as well as show that it is not inherently superior to 'verbal' or otherwise 'informal' theory regarding this end (e.g., verbal descriptions, mechanistic models).

Let us begin by identifying the main arguments in favor of formal modelling in the current literature. I here include arguments of all proponents of formal modelling, not just those claiming to address the 'theory crisis', to make the case for formal modelling as strong as possible. In order of their overall acceptance amongst proponents of modelling, the expected benefits of formal modelling are:

1. Deductive precision: It is more straight-forward to deduce hypotheses from formalized theory (i.e., a theory formalized into a formal model) than from unformalized verbal theory, due to the possibility to directly apply the rules of formal logic or mathematics to formalized theories or models as well as the concomitant elimination of any ambiguity inherent to natural language. This allows us to make hypotheses that must be the case according to our theories, instead of merely being possible (Borsboom et al., 2021; Fried, 2020a, 2020b; Guest & Martin, 2021; Haslbeck et al., 2022; Oberauer & Lewandowsky, 2019; Robinaugh et al., 2021; Smaldino, 2017, 2019, 2020; Van Rooij & Baggio, 2021).

2. Evaluability and simulation: Formalization also allows us to formally evaluate whether a (tested) hypothesis actually follows/is deducible from a (formalized) theory. This allows us to test the viability of our theories as an explanation for their explanandum, as well as check for any constraint violations or unlikely implications, in advance of empirical testing (Borsboom et al., 2021; Fried, 2020a; Guest & Martin, 2021; Oberauer & Lewandowsky, 2019; Robinaugh et al., 2021; Scheel et al., 2021; Smaldino, 2020; Van Rooij & Baggio, 2020).

3. Explicitness: The creation of a formal model forces you to make core variables and auxiliary assumptions about e.g. variable interrelationships, moderator variables and boundary conditions explicit (Oberauer and Lewandowsky, 2019; Smaldino, 2020, p. 1614; Borsboom et al., 2021; Guest and Martin, 2021; Robinaugh et al., 2021; Scheel et al., 2021; Van Rooij and Baggio, 2021).

4. Risky testing: Following Meehl (1990a, 1990b) and Popper (1959), formal models can supply precise mathematical predictions, which increase the testability and falsifiability of theories – i.e., they allow theories to be riskily tested. This is possible by, for example, comparing theory-implied simulated data with actual empirical data (Borsboom et al., 2021; Fried, 2020a; Robinaugh et al., 2021; Scheel et al., 2021).

5. Open theorizing: Formal theories are 'open theories', which – in contrast to unformalized, verbal theories – can be easily interpreted, used and modified by those who did not conceive them, as well as being transferrable across domains. Furthermore, their (annotated) code can be shared openly (Fried, 2020a; Guest & Martin, 2021; Robinaugh et al., 2021; Smaldino, 2020).

6. Emergence from simulations: Simulations based on formal models can lead to new hypotheses

as well as the prediction of emergent phenomena that would not be identified without formal modelling, due to the intractability of reasoning through all possible variations of values/states in complex verbal models (Robinaugh et al., 2021; but also partially present in Oberauer and Lewandowsky, 2019).

Though not necessarily intended as such by some, most of these arguments can be brought to bear on issues with underdetermination. Let us begin with the reduction of contrastive underdetermination, for which formal modelling appears to be an especially fruitful approach. To formalize a verbal theory$_1$, we after all need to (3) specify the exact mathematical or logical relationships between all variables (e.g., Fried, 2020a; Guest and Martin, 2021), making any such model a far more specific instantiation of the verbal theory it is derived from – especially given the fact that this translation can often occur in multiple mutually exclusive ways (Robinaugh et al., 2021). Such specification in turn allows (1) direct derivation of highly precise hypotheses and thus allows us to (4) engage in risky testing by formulating unlikely hypotheses, especially if the formal model's variable values are directly translatable into real-life measurement values (Haslbeck et al., 2022). An added benefit here is that (2) we can evaluate the fit of hypotheses to theory$_1$ quickly through simulation, decreasing the chance we test an irrelevant or impossible hypothesis. Lastly, the possibility that the formal model leads to (6) unexpected, 'emergent' predictions under some parameter values, as pointed out by Robinaugh et al. (2021), further increases the value of formal modelling if we wish to engage in risky testing.

As Robinaugh et al. (2021) already recognized, the possibility to in fact engage in such risky testing based on point predictions depends on our already having highly precise, valid and accurate measurement instruments – if not, any failed prediction can be simply blamed on measurement problems such as invalidity, low reliability or random error. As mentioned previously, it has however recently been shown that both validation practices and such measurement instruments are often missing (Chester & Lasko, 2021; Feest, 2019; Flake & Fried, 2020) and that the development of measurement techniques is coupled to advances in concepts (i.e., part of verbal theory) through a process of epistemic iteration (Chang, 2004; Eronen & Bringmann, 2021; Eronen & Romeijn, 2020). Furthermore, as I argued in the previous sections, the validity of measurement techniques and the degree to which we can establish this validity rely largely on the specificity of our theory$_1$ and the details of our theory$_2$. Previous validity claims can after all be easily obviated by changes in the structure of either of the two, such as changes in the hypothesized real-world referents of core concepts. Robinaugh et al. (2021) do not offer a direct solution to these problems, but instead argue that possible measurement problems will become apparent by also formalizing our measurement theories. In turn, this formalization will make our hidden (auxiliary) assumptions about our measurement techniques explicit, such as what our measurements actually indicate. They go on to state that this newfound explicitness will improve our measurement techniques, without describing how this will actually come to pass.

Whilst useful, specifying the (previously tacit) assumptions required for risky tests with certain measurement techniques however does not yet make these instruments valid or precise – the only way to achieve the latter would be to eliminate sources of error in our existing measurements or to conceive better measurement techniques. Neither of these latter solutions is dependent on or supported by formal modelling, nor by other forms of formalization. Formal modelling also does not fulfill any role for the specification of what our measurement techniques in fact (are supposed to) measure. Instead, more specific verbal theorizing is required for such improvements to measurement instruments and their interpretations, as these require knowledge about the real-world referent of the measured construct or quantity and its embedding in a real-world system.

We should, for example, first determine what the to-be-measured thing in fact is in the world before we can even think about measuring it. Given such knowledge, we could formulate an ontologically specific, mechanistic model (Bechtel & Abrahamsen, 2005) that covers the quantity and its interactions with the environment in order to identify new approaches to its measurement, allowing us to validate measurements using triangulation or knowledge of the causal connections between measurement technique and construct (Borsboom et al., 2004). Such a mechanistic model is not formal per se; in most cases, mechanistic models are verbal and schematic descriptions of the components involved in generating a phenomenon (Bechtel & Abrahamsen, 2005). Another option would be to create a complete overview of all the (confounding) factors and processes that contribute to the measurement process and its error, in order to see if these can be eliminated or controlled for in our predictions (i.e., a form of nomological validation; Cronbach and Meehl, 1955; Hagger et al., 2017). In either case, we need to engage in a form of specific theorizing: we need to specify the ontological referent of our to-be measured quantity or offer an otherwise clear conceptualization of it, specify not only which causal relations it enters into but also how

these occur, and derive possible moderating variables from these assumptions.

While this increases specificity of our theory$_1$, neither of these approaches has anything to do with formalizing a theory$_1$ per se – they could easily occur independently of formalization. Furthermore, such specification also requires, for example, a complementary paradigmatic view of human cognition implying ontological commitments that can support and guide such specific interpretations. In fact, if our theories$_1$ and experiments are not deeply informed by such commitments, it is unlikely we can exclude other theories$_1$ or interpretations derived from other paradigms (see also Oude Maatman, 2020 for an example of this regarding interpretations of 'mental states'). This also goes for formal models of such theories$_1$; without a commitment to a paradigmatic background theory, there are extremely few constraints that could eliminate any particular formalization (see also Van Rooij and Baggio, 2021). This entails that specification of theory$_2$ also is an important, if not necessary, tool for reduction of contrastive underdetermination in general.

The aforementioned points also show that a premature move to risky testing based on formal models – that is, without good measurement techniques or an embedding in a broader paradigmatic view on human cognition – could make the degree of holistic underdetermination worse. Without some conception about what we measure as well as possible sources of measurement variance and error, it is difficult to support the many assumptions about the target system, reliability, validity and accuracy required to trust precise measurements, let alone even create such precise measurement techniques. Even if such measurement techniques could be created in absence of advances in (informal) verbal theory, a predictive success could be heavily criticized as possibly irrelevant to the theory due to invalid measurement. This mirrors the calls for deeper consideration of the importance of conceptualization and measurement by Eronen and colleagues (2021, 2020) and others (Feest, 2019; Morawski, 2019).

The decrease in contrastive underdetermination offered by formal modelling then appears to be counterbalanced with the inability to properly capitalize on it due to an increase in its holistic sibling. That is, unless serious advances in measurement are made that are dependent on more specific theorizing and increased visibility of our deeper theoretical and ontological commitments (i.e., theory$_2$) – and thus not on formal theorizing or modelling alone. The actual contribution of Robinaugh and colleagues' (2021) formalization strategy then would be to make our problems with holistic underdetermination more explicit: can we be certain of the measurement interpretations we require? This is very useful in its own right, but not a solution to the contrastive horn of the theory crisis, entailing that formal modelling by itself cannot be sufficient as a solution to high degrees of underdetermination.

Notably, we also do not need formal modelling to decrease contrastive underdetermination, even though it is very helpful for this. It is, after all, more than possible to derive specific, risky predictions from theories without actually engaging in formalization. Consider, for example, positing not a single effect but a larger pattern of directional effects across pre-specified variations in experimental circumstances that are strictly implied by the (verbal) theory$_1$ (see also Tunç and Tunç, 2023) – which is already far more risky than gambling on single effect directions. Another example is positing that a hypothesized mechanism or construct exists as hypothesized by the theory$_1$, which is essentially a hypothesis about existence and form rather than effects. A good example of this would be the prediction and later discovery of the double helix structure of DNA by Watson, Crick, Franklin and Wilkins. Ontologically specific descriptions of mechanisms also imply their own confounding variables, and therefore can be riskily tested by investigating whether these confounds indeed are 'confounding' as long as these confounds are overall unlikely in other theories.

Though offering less reduction of contrastive underdetermination than mathematical point predictions, the above examples show that such 'informal', verbal predictions can nonetheless be risky, especially if they are stacked in a broader theory-testing program. And notably, they achieve this without the necessity to first create highly precise measurement techniques, offering a way to empirically establish warrant for a theory$_1$ before such steps are taken. From a purely experimental perspective, it would then be advisable to specify before we formalize. To conclude, formal modelling is also not necessary to reduce degrees of contrastive underdetermination – though it remains a highly useful method for doing so.

Let us now turn to the question whether formal modelling can help decrease holistic underdetermination. Formal modelling is argued to have direct bearing on the degree of holistic underdetermination through the (3) specification of theoretical assumptions it requires and, through this requirement, is argued to forcibly lead to (Oberauer and Lewandowsky, 2019, p. 1614; Robinaugh et al., 2021). For example, Oberauer and Lewandowsky (2019, p. 1614) argue that the formulation of a Bayesian network model (e.g., a DAG) would involve the identification and incorporation of "assumed moderator variables, boundary conditions and other

auxiliary assumptions". Similarly, Guest and Martin (2021) argue that computational modeling forces "scientists to explicitly document an instance of what their theory assumes, if not what their theory is", which in their view notably is possible not just in mathematics, but also in natural language.

Guest and Martin (2021) thereby also touch upon the importance of distinguishing formalization from specification, although I do not use these words in the same way as they do. Formalization of a theory_1 is a form of specification in my sense of the word: it is the specification of the logical or mathematical relationships between variables or entities in the theory_1 and thereby the overall dependency structure amongst these. This can also be applied to concrete instances of the theory_1 (i.e., a specific context or task to which the theory ought to apply) if adapted to real-life measurement techniques, and offers the benefit of computable simulation of our theory's implications. As we have seen, this can lead to a welcome decrease in contrastive underdetermination under the right circumstances. Formalization in itself however does not help us identify, say, the underlying mechanism of the phenomena we study, even though it might help evaluate our hypothesized mechanism in regard to prior theoretical_2 constraints (Van Rooij & Baggio, 2021). Similarly, it is not formalization or formal modelling that would allow us to identify the relevant moderator variables, boundary conditions and other auxiliary assumptions Oberauer and Lewandowsky (2019) are speaking of.

Guest and Martin's (2021) account instead shows that general specificity is necessary for our creation of (useful) formal models to begin with: without preceding (verbal) theoretical specificity, our formal models are underconstrained by reality, and ill-adapted to comparison with or prediction of real systems and environments. Though they can potentially inform experimental practice by providing precise or emergent hypotheses, formal models and the process of formalization do not offer further relevant information for this experimental practice itself, nor for its evaluation. Although all formalized aspects of theories_1 are necessarily specific (e.g., variable interrelationships), few things that ought to be specific to successfully reduce holistic underdetermination can be explicitly formalized, after all, such as ontological commitments, definitions, referents, and auxiliaries about experimental design and measurement. Nor will these automatically become specific due to engaging in formalization: we cannot, for example, identify real-life confounds for causal connections between variables on the basis of (generating) the mathematical or logical relationships that might represent these causal connections after formalization, no matter how precise they are.

Contra Oberauer and Lewandowsky (2019), I would therefore argue that "adopting formal modelling 'forces' specificity" is a claim that puts the cart before the horse: theoretical specificity is not a concomitant benefit of formal modelling as suggested but instead a prerequisite for maximizing its utility, and thus should not be conflated with it. The possibility to improve on these matters is wholly independent of the process of formalization, unless we conflate formalization with carefully thinking about what we assume and what our theory implies (which we can also do non-formally). If the counter-argument then is that one might or must start thinking about these matters whilst or due to engaging in formal modelling – i.e., formal modelling forces theorizing (Guest & Martin, 2021) – I would counter we could just as well proscribe a method where this is not a side-effect but the 'main dish', such as the application of conceptual analysis, reading the history and philosophy of psychology, or the simple method of carefully thinking about what we assume and what our theory implies.

Nevertheless, formal modelling is an incredibly useful tool for instances where the theory is too complex to reason through unaided (for examples; see Smaldino, 2017). Yet, even then it is only likely to highlight and improve a specific subset of auxiliary assumptions, which differs markedly from those we are drawn to reason about unaidedly. Formal models are merely a mathematical or otherwise formal decontextualized redescription of the theory, from which all such information necessary for experimental design and other implementations has been purposefully eliminated in order to achieve the clarity required for more formal evaluation. Notably, this is not a bug, but a feature of formalization. It is quite literally the goal of formalization to eliminate semantic noise (i.e., meaning, reference) and keep only the more easily evaluable syntactic structure. This move is very useful – but not for a reduction of holistic underdetermination. There, this feature possibly becomes a bug due to indirectly directing our attention away from real-world implementability and the aforementioned practical issues with meaning and reference psychology is often plagued with, and towards mathematical, logical and structural intricacies of the model. Robinaugh et al. (2021, 2024) themselves already show such an effect when they focus on the auxiliary assumptions about mathematical relationships between measurements and quantities that formalization uncovers, instead of the preceding, experimentally more pressing question whether these measurements actually measure what they ought to measure – let alone whether the theoretical entities involved are well-defined enough.

Due to its inherent decontextualization, formal modelling thus appears to potentially move in exactly the opposite direction from what is required to reduce holistic underdetermination in practical experimentation; it directs attention away from experimental practice and implementation, where the problems with holistic underdetermination however are most present.

Even given the above, one could still argue that formal modelling might still serve a general role in resolving the theory crisis as a method that enforces specific theorizing upon the researcher, and thus forms an indirect countermeasure to underdetermination. Whilst neither sufficient nor necessary, it thus still confers a possible benefit if adopted by psychologists, and could be unproblematically combined with other tools for reasoning.

It however is important to recognize that as a method, formal modelling is not and cannot be theory-neutral (Danziger, 1985), which partially counters this last, middle-of-the-road perspective. The applicability of formal modelling after all also relies on certain theoretical and auxiliary assumptions of its own, such as the general viability of quantifying (all) variables and relationships in the theory if the theory is to be translated into a mathematical formula or if point predictions are to be achieved, or of the possibility to coherently apply a form of logic or computational language without problematically 'deforming' whatever is described.

Assumptions about the quantifiability of the psychological as well as the viability of our current attempts at quantification however have been seriously criticized (e.g., Michell, 2021; Tafreshi et al., 2016). Non-mathematical approaches to formal modelling may be exempt from the latter critiques, but are also not wholly theory-neutral. This is the case because formal modelling as a method can only be applied to theories that are amenable to it. It therefore imposes constraints on what type of theories and research can be pursued if it is to be adopted as a new standard in the field (i.e., it enforces a methodological imperative; Danziger, 1985). As Navarro (2021) points out, formal modelling is not well-suited to tasks and theories that are themselves too complex, thereby indirectly forcing researchers into the direction of simplified theory and highly constrained research designs, which also delimits the type of research questions we might ask. Though I cannot delve deeply into this debate here, I therefore want to stress that the assumptions underlying the viability of formal modelling itself as well as its desirability for a psychological research area should also be considered in any future push towards its adoption – like several proponents also argue (e.g., Borsboom et al., 2021).

Formal modelling or formalization itself then has little bearing on holistic underdetermination, and therefore is neither sufficient nor necessary for the resolution of this horn of the theory crisis either. In order to be able to reduce holistic underdetermination, we instead must engage in several non-formal types of theorizing, such as clear conceptualization, specifying real-life mechanisms and referents, or, like Van Rooij and Baggio's (2021) use of computationalism, explicitly adopting the assumptions of an existing well-developed paradigmatic theoretical framework (i.e., one important element of our theory_2). In other words, we must create specific theories_1, which also requires awareness and explicitness of theory_2. Only then can formal modelling be applied fruitfully.

Given the inability of formal modelling to address holistic underdetermination, we can conclude it is neither sufficient nor necessary to resolve the current 'theory crisis', drawing my argument to a close. Nevertheless, a few loose ends remain after this conclusion.

First, there is the possible counter-argument that all of the above is implicit in the current literature. Yet, the necessary types of specification outlined above remain unmentioned in current proposals for theoretical improvement in psychology; though closest to what I outline – whilst occupied with a wholly different problem – even Van Rooij and Baggio (2020, 2021) do not draw too much attention to the fact that their proposed method relies on many (disputable) theoretical background assumptions about the structure of human cognition, such as the deeply theoretical claim that computable capacities are the core explananda of psychology. The recent formalization-based theory construction methods instead focus on the intricacies of the formalization process whilst defining theory purely as theory_1 (e.g., Borsboom et al., 2021; Haslbeck et al., 2022; Van Rooij and Baggio, 2020), largely ignoring necessary specifications in and awareness of theory_2 that are necessary for creation and testing of a specific theory_1 (e.g., paradigmatic assumptions, tacit knowledge, which theory of causality is adopted – though see Van Rooij and Baggio, 2021 for a strong example that does do this implicitly by adopting computationalism). In doing so, the current literature thereby appears to primarily aim at providing methods for formalizing and evaluating existing theories_1 instead of in fact telling us something about how to construct specific theories_1 from scratch, including the creation of well-defined concepts and the process of demarcating phenomena to begin with.

Second, I have not truly dealt with the introduction's pejorative treatments of verbal theory yet. After all, it could be argued that any verbal theorizing in the end could or should be formalized, meaning that verbal theory is still some form of 'prototheory' (Borsboom

et al., 2021). In response to this, I argue that many of the theories and assumptions in theory$_2$ are inherently 'informal', due to being verbal or schematic (e.g., a schematic image of a synaptic cleft). That these are not formal does not invalidate them, nor does it make them imprecise by definition: many such assumptions are either simply not formalizable, or their formalization would lead to information loss. Similarly, I have not explicitly shown that verbal theory is not necessarily ambiguous or imprecise (Fried, 2020a; Robinaugh et al., 2021). To this I would say that if a construct is vague or ambiguous, this may not be the result of the imprecision of language, but of the imprecise usage of language. If it is difficult to define core theoretical terms such as 'tendency' (Navarro, 2021), the answer therefore does not necessarily have to be formalization if we wish to make our theory testable. Instead, we should make explicit what it is we mean, as this is necessary to connect our theory to experimental practice for all theories, even formal ones. By doing so, we after all gain the ability to reason about and reduce the holistic underdetermination associated with our measurement techniques and experimental designs. Even though such conceptual specification may be complicated and perhaps more suited to philosophers than to experimental psychologists, it will be a necessary step if we want to derive warrant for our theories from our results (see also Mayo, 2018, p. 99-101).

If one agrees with me up to this point, the remaining solution is to not only improve our 'informal' theories, but also engage in theorizing in a much broader sense than is implied by the current debate. Strong theory is not derived from, nor dependent on, the form of theorizing (i.e., verbal or formal), but only stems from the process of theorizing itself – that is, formulating a coherent (and in time well-supported) set of assumptions aimed at the explanation of a single or multiple phenomena (i.e., theory$_1$) that is consistent with a broader, explicit set of assumptions and practical knowledge that informs this theory$_1$ and is required to connect it to real-life practice (i.e., theory$_2$). Formal modelling can potentially assist in such theorizing in areas that have sufficiently specific verbal theory and can shoulder the assumptions of formal models, whilst specific verbal theorizing will be necessary everywhere.

Conclusion

In this paper, I have argued that formal modelling is not a solution to the 'theory crisis', since it does not engage with the problem that lies at its core: the high degree of both contrastive and holistic underdetermination of psychological theories by our experiments and results, which is caused by a lack of specificity of our theories. In doing so I also outlined a different approach to the resolution of this crisis: an 'informal' route, which involves a deeper engagement in and appreciation of specific non-formal theorizing. Whilst I recognize the added value of formalization and formal modelling in psychological research as a whole, I hope to have shown that we cannot relegate (good) verbal theorizing to the bargain bin, and that it should not be considered necessarily imprecise (Fried, 2020a; cf. Robinaugh et al., 2021) or a mere stepping stone towards actual, formal theory; 'prototheory' (Borsboom et al., 2021). Instead, specific verbal theorizing is a necessary component of any attempt at solving the theory crisis, and a prerequisite before we should even start to think of formalization.

After all; the quality of a theory is not only a matter of form, but primarily one of content. As mentioned throughout the article, formal modelling can doubtlessly assist in making this content more explicit and evaluable, but as we have seen, its usefulness remains heavily limited by the preceding verbal basis it is applied to – a theory$_1$ that is limited in its specificity by the theory$_2$ drawn on to support it. To strengthen this basis, I believe we must not just create specific verbal theories but also look beyond our disciplinary borders to fields engaging with the foundational assumptions underlying our work. In order to be explicit about theory$_2$, to evaluate its quality and coherence with our theory$_1$, we after all will also need to be aware of it. We therefore must engage with fundamental philosophical and theoretical discussions about these very background assumptions – which luckily have continued in the shadow of mainstream empirical psychology. Fields like theoretical and critical psychology, science and technology studies (STS), philosophy of mind and philosophy of science can be drawn from to inform and criticize not just theories$_1$ but also to become aware of and evaluate our theory$_2$: our tacit knowledge (e.g., Morawski, 2019), ontological commitments and metaphysical positions (e.g., Hochstein, 2019), and other implicit assumptions involved in our research practice.

Author Contact

Correspondence should be addressed to Freek J.W. Oude Maatman at freek.oudemaatman@ru.nl.

ORCID - Freek Oude Maatman: 0000-0002-4795-435X

Conflict of Interest and Funding

The author declared no potential conflicts of interest.

Author Contributions

Freek J.W. Oude Maatman: Conceptualization, Writing - original draft, Writing - review & editing

Open Science Practices

This article is conceptual and not eligible for Open Science badges. The entire editorial process, including the open reviews, is published in the online supplement.

Acknowledgements

I want to thank Jan Bransen, Sander Bisselink, Markus Eronen and the Radboud Theoretical Psychology group for their feedback on early versions of this manuscript. I also want to thank Iris van Rooij, Olivia Guest, and Paul Smaldino for their comments and feedback on previous preprinted versions of the manuscript.

References

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity [Place: US Publisher: American Psychological Association]. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Brenninkmeijer, J., Derksen, M., & Rietzschel, E. (2019). Informal Laboratory Practices in Psychology (S. Vazire & M. Nuijten, Eds.). *Collabra: Psychology*, *5*(1), 45. https://doi.org/10.1525/collabra.221

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Chester, D. S., & Lasko, E. N. (2021). Construct Validation of Experimental Manipulations in Social Psychology: Current Practices and Recommendations for the Future [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(2), 377–395. https://doi.org/10.1177/1745691620950684

Collins, H. (1985). *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Danziger, K. (1985). The Methodological Imperative in Psychology [Publisher: SAGE Publications Inc]. *Philosophy of the Social Sciences*, *15*(1), 1–13. https://doi.org/10.1177/004839318501500101

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology [Publisher: Frontiers]. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *30*(6), 786–799. https://doi.org/10.1177/0959354320969876

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A Validity-Based Framework for Understanding Replication in Psychology [Publisher: SAGE Publications Inc]. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Farrell, S., & Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology [Publisher: SAGE Publications Inc]. *Current Directions in Psychological Science*, *19*(5),

329–335. https://doi.org/10.1177/0963721410386677

Feest, U. (2019). Why Replication Is Overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The Long Way From -Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *7*(6), 661–669. https://doi.org/10.1177/1745691612462587

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them [Publisher: SAGE Publications Inc]. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fried, E. I. (2020a). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I. (2020b). Theories and Models: What They Are, What They Are for, and What They Are About. *Psychological Inquiry*, *31*(4), 336–344. https://doi.org/10.1080/1047840X.2020.1854011

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Hagger, M. S., Gucciardi, D. F., & Chatzisarantis, N. L. D. (2017). On Nomological Validity and Auxiliary Assumptions: The Importance of Simultaneously Testing Effects in Social Cognitive Theories Applied to Health Behavior and Some Guidelines [Publisher: Frontiers]. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01933

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories [Place: US Publisher: American Psychological Association]. *Psychological Methods*, *27*(6), 930–957. https://doi.org/10.1037/met0000303

Hochstein, E. (2019). How metaphysical commitments shape the study of psychological mechanisms [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *29*(5), 579–600. https://doi.org/10.1177/0959354319860591

Iso-Ahola, S. E. (2017). Reproducibility in Psychological Science: When Do Psychological Phenomena Exist? [Publisher: Frontiers]. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00879

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The Problem of Coordination and the Pursuit of Structural Constraints in Psychology [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 767–778. https://doi.org/10.1177/1745691620974771

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *24*(3), 326–338. https://doi.org/10.1177/0959354314529616

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago.

Landy, J. F., Jia, M. (, Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., Van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Maier, M., Van Dongen, N., & Borsboom, D. (2024). Comparing theories with the Ising model of explanatory coherence [Place: US Publisher: American Psychological Association]. *Psychological Methods*, *29*(3), 519–536. https://doi.org/10.1037/met0000543

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

Mayo, D. G., & Spanos, A. (Eds.). (2009). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Ra-*

tionality of Science. Cambridge University Press. https://doi.org/10.1017/CBO9780511657528

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology [Place: US Publisher: American Psychological Association]. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990a). Why summaries of research on psychological theories are often uninterpretable [Place: US Publisher: Psychological Reports]. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/PR0.66.1.195-244

Meehl, P. E. (1990b). 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

Michell, J. (2021). Representational measurement theory: Is its number up? [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *31*(1), 3–23. https://doi.org/10.1177/0959354320930817

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? [Place: US Publisher: Educational Publishing Foundation]. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 218–238. https://doi.org/10.1037/teo0000129

Muthukrishna, M., & Henrich, J. (2019). A problem in theory [Publisher: Nature Publishing Group]. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Navarro, D. J. (2021). If Mathematical Psychology Did Not Exist We Might Need to Invent It: A Comment on Theory Building in Psychology [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 707–716. https://doi.org/10.1177/1745691620974769

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance [Publisher: Annual Reviews]. *Annual Review of Psychology*, *69*(Volume 69, 2018), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

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

Oude Maatman, F. (2020). Folk psychology and network theory: Fact or gamble? A reply to Kalis and Borsboom [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *30*(5), 729–734. https://doi.org/10.1177/0959354320952863

Piccinini, G. (2009). Computationalism in the Philosophy of Mind. *Philosophy Compass*, *4*(3), 515–532. https://doi.org/10.1111/j.1747-9991.2009.00215.x

Popper, K. R. (1959). *The logic of scientific discovery* [Pages: 480]. Basic Books.

Proulx, T., & Morey, R. D. (2021). Beyond Statistical Ritual: Theory in Psychological Science [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 671–681. https://doi.org/10.1177/17456916211017098

Quine, W. V. (1951). Main Trends in Recent Philosophy: Two Dogmas of Empiricism [Publisher: [Duke University Press, Philosophical Review]]. *The Philosophical Review*, *60*(1), 20–43. https://doi.org/10.2307/2181906

Richters, J. E. (2021). Incredible Utility: The Lost Causes and Causal Debris of Psychological Science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

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 [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., McNally, R. J., Ryan, O., de Ron, J., Van der Maas, H. L. J., Van Nes, E. H., Scheffer, M., Kendler, K. S., & Borsboom, D. (2024). Advancing the network theory of mental disorders: A computational model of panic disorder [Place: US Publisher: American Psychological Association]. *Psychological Review*, *131*(6), 1482–1508. https://doi.org/10.1037/rev0000515

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Shapiro, L., & Spaulding, S. (2021). Embodied Cognition. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford

University. Retrieved December 10, 2025, from https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis [Publisher: Annual Reviews]. *Annual Review of Psychology*, *69*(Volume 69, 2018), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Skinner, B. F. (1953). *Science and human behavior* [Pages: x, 461]. Macmillan.

Smaldino, P. (2017). Models Are Stupid, and We Need More of Them [Num Pages: 21]. In *Computational Social Psychology*. Routledge.

Smaldino, P. (2019). Better methods can't make up for mediocre theory [Bandiera_abtest: a Cg_type: World View Publisher: Nature Publishing Group Subject_term: Research management, Careers]. *Nature*, *575*(7781), 9–9. https://doi.org/10.1038/d41586-019-03350-5

Smaldino, P. (2020). How to Build a Strong Theoretical Foundation. *Psychological Inquiry*, *31*(4), 297–301. https://doi.org/10.1080/1047840X.2020.1853463

Stanford, K. (2017). Underdetermination of Scientific Theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. Retrieved December 10, 2025, from https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Tafreshi, D., Slaney, K. L., & Neufeld, S. D. (2016). Quantification in psychology: Critical analysis of an unreflective practice [Place: US Publisher: Educational Publishing Foundation]. *Journal of Theoretical and Philosophical Psychology*, *36*(4), 233–249. https://doi.org/10.1037/teo0000048

Teo, T. (2020). Theorizing in psychology: From the critique of a hyper-science to conceptualizing subjectivity [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *30*(6), 759–767. https://doi.org/10.1177/0959354320930271

Trafimow, D. (2017). Implications of an initial empirical victory for the truth of the theory and additional empirical victories. *Philosophical Psychology*, *30*(4), 415–437. https://doi.org/10.1080/09515089.2016.1274023

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein [Publisher: SAGE Publications Ltd]. *Theory & Psychology*, *26*(4), 540–548. https://doi.org/10.1177/0959354316637136

Tunç, D. U., & 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 Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Van Geert, P. L. (2019). Dynamic Systems, Process and Development. *Human Development*, *63*(3-4), 153–179. https://doi.org/10.1159/000503825

Van Rooij, I., & Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, *31*(4), 321–325. https://doi.org/10.1080/1047840X.2020.1853477

Van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

Wallot, S., & Kelty-Stephen, D. G. (2018). Interaction-Dominant Causation in Mind and Brain, and Its Implication for Questions of Generalization and Replication. *Minds and Machines*, *28*(2), 353–374. https://doi.org/10.1007/s11023-017-9455-0