Productive Explanation: A Framework for Evaluating Explanations in Psychological Science

van Dongen, N.; van Bork, R.; Finnemann, A.; Haslbeck, J.M.B.; van der Maas, H.L.J.; Robinaugh, D.J.; de Ron, J.; Sprenger, J.; Borsboom, D.

DOI
10.31234/osf.io/qd69g

Publication date
2024

Document Version
Final published version

Productive Explanation: A Framework for Evaluating Explanations in Psychological Science

Noah van Dongen[1], Riet van Bork[1,2], Adam Finnemann[1], Jonas M. B. Haslbeck[6,1], Han L. J. van der Maas[1], Donald J. Robinaugh[3,4], Jill de Ron[1], Jan Sprenger[5], and Denny Borsboom[1]

[1]Department of Psychology, University of Amsterdam
[2]Center for Philosophy of Science, University of Pittsburgh
[3]Department of Applied Psychology and Department of Art + Design, Northeastern University
[4]Department of Psychiatry, Massachusetts General Hospital
[5]Department of Philosophy and Education, University of Turin
[6]Department of Clinical Psychological Science, Maastricht University

Author Note

This manuscript is accepted for publication at *Psychological Review.*

Correspondence may be addressed to Noah van Dongen at nnnvandongen@gmail.com. We would like to thank Tessa Blanken for her helpful comments on an earlier version of this manuscript. We also thank Edouard Machery, Henry Chase, Dejan Makovec, Adam Koberinski, Hong Hui Choi, Lotem Elber, Raquel Krempel, and Tessa Blanken for helpful suggestions that improved the revision of this manuscript. The code to reproduce the analyses and the figures can be found at github.com/jmbh/explanationpaper. The work by NvD, RvB, AF, JMBH, JdR, and DB was supported by NWO Vici grant no. 181.029. DR's work on this paper was supported by funding from the National Institute for Mental Health, grant nr. K23 MH113805. JMBH was supported by the project "New Science of Mental Disorders" (www.nsmd.eu), supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.004.016). The content is solely the responsibility of the authors and does not necessarily represent the views of any funding agency.

Abstract

The explanation of psychological phenomena is a central aim of psychological science. However, the nature of explanation and the processes by which we evaluate whether a theory explains a phenomenon are often unclear. Consequently, it is often unknown whether a given psychological theory indeed explains a phenomenon. We address this shortcoming by proposing a productive account of explanation: a theory explains a phenomenon to some degree if and only if a formal model of the theory produces the statistical pattern representing the phenomenon. Using this account, we outline a workable methodology of explanation: (a) explicating a verbal theory into a formal model, (b) representing phenomena as statistical patterns in data, and (c) assessing whether the formal model produces these statistical patterns. In addition, we provide three major criteria for evaluating the goodness of an explanation (precision, robustness, and empirical relevance), and examine some cases of explanatory breakdowns. Finally, we situate our framework within existing theories of explanation from philosophy of science and discuss how our approach contributes to constructing and developing better psychological theories.

Keywords: scientific explanation; psychological methods; models and theories; theories and phenomena; theory appraisal; quality of explanations

1 Introduction

In the wake of the replication crisis in psychological science, many psychologists have adopted practices that bolster the robustness and transparency of the scientific process, including preregistration (Chambers, 2013), data sharing (Wicherts, Borsboom, et al., 2006), code sharing, and massive reproducibility studies (e.g., Aarts, Anderson, et al., 2015; Walters, 2020). This set of reforms is critical to improving the empirical part of the research process. However, the weaknesses in psychological science do not only concern the empirical material of the field. There is also a shortage of strong psychological theories with accurate and informative predictions, and high explanatory power[1] for phenomena (Fried, 2021; Meehl, 1967, 1978; Oberauer & Lewandowsky, 2019). Accordingly, we need to develop methods that can support and facilitate the construction of successful theories in psychology (Borsboom, van der Maas, et al., 2021; Guest & Martin, 2021; Haslbeck, Ryan, et al., 2021; Robinaugh, Haslbeck, et al., 2021; van Rooij & Baggio, 2021). This paper contributes to this agenda by developing a novel model of explanation in psychological science and spelling out criteria for high explanatory power of a theory with respect to empirical phenomena.

In current practice, psychological explanations typically present a narrative where a theory renders an empirical phenomenon intuitively likely. Throughout this paper, we use the regulatory resource theory of ego-depletion (Baumeister, Bratslavsky, et al., 1998) as an example, as it offers an explanation pattern that is representative of much of psychological science. As is the case for most psychological theories, and many other disciplines in social science, regulatory resource theory presents a verbal narrative (see, e.g., Braithwaite, 1960). The theory suggests that self-control works like a muscle, which can be depleted of energy by using it. Research on ego-depletion is typically organized around experiments that require participants to either perform a challenging task or an easier control task, followed by an evaluation of perseverance or performance on a subsequent task. Across multiple studies, researchers have reported that participants who have performed the challenging task will, on average, show less perseverance or a worse performance in the subsequent task (Vohs, Schmeichel, et al., 2021), a phenomenon referred to as the "ego-depletion effect," because of the theorized depletion of self-control posited to underlie the decrement in perseverance or performance.

[1] The concept of explanatory power is used in a technical sense in models of statistical explanation such as Schupbach and Sprenger (2011), but we use it in a generic way, as referring to the quality or goodness of an explanation.

However, the fact that the theory is purely verbal in character makes it hard to evaluate what exactly is implied by the theory. As a result, it is difficult to assess whether the theory indeed explains a certain phenomenon. Namely, the theory lacks the details to provide answers to important questions, such as: How challenging should the task be? For how long must one be engaged in? How much time can there be between tasks before ego recovers?

This example shows that the link might initially seem intuitive and sound, but looking closer reveals several problems. One is that the connection between theory and relevant experimental manipulations becomes a matter of opinion. For example, in a recent large-scale study of ego-depletion (Vohs, Schmeichel, et al., 2021), predictions did not follow from the theory. Instead, the theorists needed to be consulted and it was their opinion that informed the tests. The fact that this crucial link between theory and experiment had to be spelled out by polling the experts reveals an Achilles' heel in the theory: it is unclear what exactly the theory predicts on its own.

This state of affairs stands in stark contrast to other domains of science, and physics in particular, as already noted by Paul Meehl (1967, 1978). For example, nobody ever had to ask Einstein about the presumable starlight trajectory in the famous 1919 eclipse observed by Eddington (Dyson, 1917). Because Einstein's General Theory of Relativity implies a specific degree of starlight bending as a result of the sun's gravitational forces, nobody needed to consult him to make predictions with his theory. This is because General Theory of Relativity is implemented in a formal model and, consequently, every researcher will derive the same precise predictions from the theory.

Regulatory resource theory affords no such clarity. Here, the theory's predictions depend on how an individual researcher fills in the gaps and vagaries with their unstated personal assumptions and mental simulations. We do not want to argue that expert opinion and consensus is irrelevant; in the Einstein example, we also require consensus on auxiliary assumptions (e.g., concerning the adequacy of measurement procedures) that are not given by the theory itself. However, given that these auxiliary assumptions are fixed, the implications of the theory are clear. This is not the case for most psychological theories, because in many cases the details of the theory itself remain implicit.

This limitation has direct consequences for the evaluation of psychological theories. In the absence of clear explanatory links, it is exceedingly difficult to gauge whether the results from a statistical test count as evidence for the theory. As a consequence, it is currently equally unclear whether

Vohs, Schmeichel, et al. (2021) provide evidence against regulatory resource theory, or whether previous findings constituted evidence for the theory. In short, high explanatory power requires a precise and unambiguous specification of the theory.

The problem of vague explanations somewhat resembles the problem of questionable research practices (QRP): in both cases, there are various contributing factors at the level of education, research strategy and policy (e.g., lack of familiarity with simulation models, diverse options for interpreting results, misguided incentive schemes). A key premise of this paper is that, similar to new methodologies that address QRP and stimulate an enormous push towards transparent and reproducible science, methodological innovation can support the development of more precise explanations. We pursue this goal by means of developing an account of explanation that lends itself for the use of computational or mathematical models—e.g., for clarifying whether the relevant explanatory links between theories and phenomena exist, and for making explanations more precise. Our project is therefore in the company of further calls for formalization in psychology (Borsboom, van der Maas, et al., 2021; Fried, 2021; Guest & Martin, 2021; Smaldino, 2017), the development of new software tools that facilitate the construction of formal models (e.g., the *R*-package *Grind* or the *Insight Maker* platform; de Boer, 2023; Fortmann-Roe, 2014, respectively), innovative approaches to theory construction (Borsboom, van der Maas, et al., 2021; Haslbeck, Ryan, et al., 2021; van Rooij & Baggio, 2021), interest in abductive methods of inference (Haig, 2021), general considerations on psychological method (e.g., Cummins, 2000; Haig, 2005), and existing approaches to indirect inference in system dynamics (Haslbeck & Ryan, 2021; Hosseinichimeh, Rahmandad, et al., 2016).

Our ambition in this paper is to develop a methodology that can be used to evaluate psychological explanations. Specifically, we take the role of phenomena as a mediator between theory and data into account (Bogen & Woodward, 1988) and define the concept of productive explanation in a three-layered structure containing, data, phenomena and theory.

In our productive explanation framework, a theory T explains a phenomenon P (to some degree) if and only if a formal model of the theory T produces a statistical pattern representing the empirical phenomenon P. Here, production is not used as a synonym for causation, but rather in the sense of (re-)producing a statistical representation of the phenomenon. Specifically, a productive explanation requires (a) representing empirical phenomena as statistical patterns in data, (b) explicating a verbal theory

into a formal model, and (c) simulating data from this formal model to verify if the statistical pattern is indeed produced (Section 2). To show how productive explanation can operate, we apply it to the illustrative case of ego-depletion (Section 3).

In addition, we propose a number of criteria for evaluating the goodness of the explanatory relation between theory and empirical phenomenon, and illustrate the use of these criteria by applying them to our formalization of regulatory resource theory (Sections 4 and 5). However we do not intend to offer a prescriptive methodology with a single or best (set of) method(s) to conduct and evaluate each step. The productive explanation framework is particularly suited to investigate explanations that target statistical patterns in data (e.g., mean differences, statistical interactions, patterns in correlation coefficients). Also, statistical patterns are the most typical products of empirical research in psychology. This focuses our discussion on the core of psychological science. However, it should be noted that we do not think our approach is intrinsically limited to this type of phenomenon, and future work may develop extensions to deal with different types of phenomena (e.g., singular events or phenomena established through qualitative research).

The final part of the paper situates our framework in the philosophy of science literature and argues that our model of explanation, with its focus on connecting the theoretical, phenomenal and observational levels, contributes a major conceptual innovation. Moreover we argue that our framework is compatible with the plurality of explanations (e.g., causal, mechanistic, mathematical, dynamical) found in modern science (Reutlinger & Saatsi, 2018). Other models of explanation may explicate specific aspects of our framework, or capture explanations where our model does not apply (Section 6). In conclusion, we discuss how the productive explanation account can assist with creating better psychological theories (Section 7).

2 Clarifying how theories explain

The philosophical literature on scientific explanation typically contrasts an *explanandum* — a fact or observation to be explained—with an *explanans*—a theory or hypothesis that accounts for the explanandum. For example, according to Charles S. Peirce (1931, p. V.185), "theory T explains phenomenon P" means "if the world were as T says it is, P would follow as a matter of course." The seminal deductive-nomological (D-N) model by Hempel and Oppenheim (1948) is developed along the same lines: the explanandum must be a logical consequence of the explanans, a set of sentences which

must contain at least one law of nature as an essential premise (i.e., one could not deduce the explanandum if that premise were removed). Similarly, Hempel's account of statistical explanation (Hempel, 1965) and Bayesian models of explanatory power (e.g., Halpern & Pearl, 2005; Schupbach & Sprenger, 2011) conceptualize explanation as a direct relation between explanans and explanandum without any intermediaries.

All these conceptions of explanation are based on rationalizing the occurrence of the explanandum by means of the explanans—for example by specifying the processes or mechanisms that gave rise to the fact or observation of interest (Cummins, 2000). This idea aligns well with scientific practice. For example, if the theory of evolution were true, the diversity of the animal world would follow as a matter of course; if the Big Bang theory were true, wavelength shifts would follow as a matter of course; if the common cause theory of general intelligence were true, the positive manifold of IQ test scores would follow as a matter of course.

The binary nature of the explanation relation in these accounts is inspired by a logical positivist picture of science (Carnap, 1928): there is a sharp distinction between the theoretical and the observational level; theories explain and are confirmed by data, and data are collected by means of perception, or by extending our perception with the help of scientific instruments. That is, the explananans is in the theoretical realm while the explanandum is in the realm of the observed.

In a seminal paper, Bogen and Woodward (1988) reject this positivist picture and argue convincingly that we should not identify the phenomena explained by a theory with sets of observations or data. Phenomena are, according to Bogen and Woodward, established by means of evidence from observations, but they are not directly observable. In psychological science, we typically call them "effects". Take the recency bias or the base rate fallacy: they are undisputed psychological phenomena, but they transcend any particular set of structured or unstructured data. Phenomena thus occupy an intermediate position between the theoretical and the observational level. They are what theories attempt to explain, while datasets serve as evidence for them and help to establish their existence. Scientific explanation pertains to phenomena, not to data.

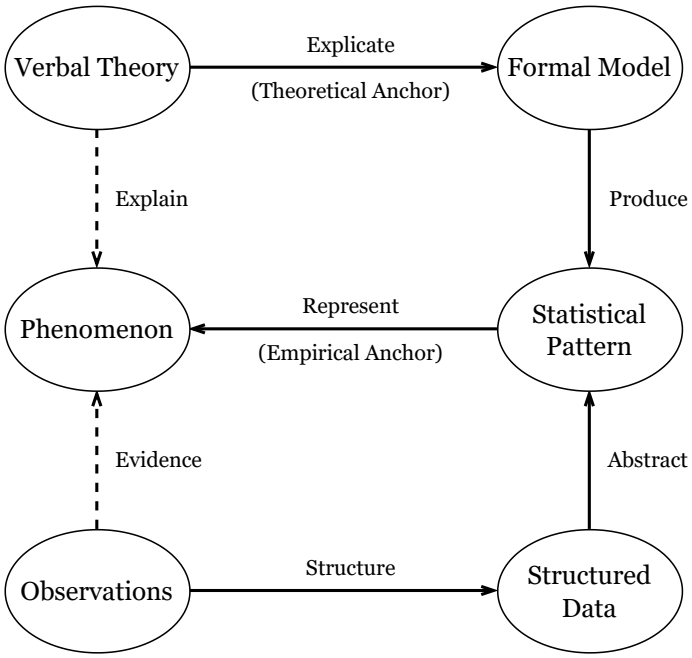
Almost no models of explanation take these conceptual distinctions into account (an exception is Ströing, 2018). We tackle this task and develop, in this section, a model of explanation that distinguishes clearly between the level of phenomena and the level of observations, and that is directly appli-

cable to psychological science. A detailed comparison to existing accounts of scientific explanation can be found in Section 6.

2.1 Data, phenomena, and theories

Figure 1 displays the components of our productive explanation framework, and the relation in which they stand to each other. The three vertical levels correspond to the theory, phenomenon, and observation, while the left and the right hand side correspond to the qualitative and the quantitative level, respectively.

Figure 1
The Productive Explanation Model.



Note. The left side, Verbal Theory–Phenomenon–Observation, can be interpreted as the qualitative side, with the right side, Formal Model–Statistical Pattern–Structured Data, as its quantitative counterpart. Vertically, the figure is structured in the theory level, the phenomenon level, and the observation level. The arrows should be read as actions from one side to the other. The arrow from "Verbal Theory" to "Formal Model" means that "the theory is explicated into a formal model." Another way of saying this is that "the formal model explicates the theory," the theory is the starting point. The "produce" connection between Formal Model and Statistical Pattern is described in Section 2.

Observations are unstructured recordings of sensory information, made with or without instrumental assistance. They are, for instance, what is seen

in a drop of blood under a microscope, a mark on a response sheet of a multiple-choice exam, or the readout of a scale in the doctor's examining room. Observations are distinct from structured data as they are immediate and not (yet) captured in a systematic and codified manner.

Data are structured recordings of observations. They are, for instance, measurement readings, test scores, or behavioral recordings that are entered in a dataset (e.g., a spreadsheet) by coding these observations into variables. Observations from qualitative research also become structured data once codified and stored in a systematic way (e.g., as variables). Variables are functions that arise from systematically assigning values to these observations, for instance by representing positive answers to a questionnaire item by a '1' and negative answers as a '0'. Thus, variables constitute mappings by which formal symbols are attached to particular observations, making these observations suitable for statistical analysis. Importantly, data are *particular*: any data point is a record of a specific observation made at a particular time and place. For the rest of this manuscript, when we mention 'data' we speak of 'structured data' as described here.

Statistical patterns are relations or patterns in data that transcend the particulars of individual datasets. These patterns can be represented by the methods and models of statistics, which hold across the individual datasets. Where data are particular and specific, statistical patterns always carry elements of generalization and abstraction. For instance, when we move from "Scores on IQ items A and B in test X are positively correlated in our data" to "scores on IQ items are positively correlated" we abstract away from the particulars of our data file by generalizing the patterns we see to other tests: e.g., from IQ items A and B of test X to IQ items in general, or from individuals in our sample to the entire population. Such generalizations go beyond the available data by asserting the generalization of the statistical patterns across other, possibly hypothetical, datasets.[2] the sign of a parameter (e.g., $r > 0$). For instance, the statistical pattern of correlations between IQ test scores could be specified to not only be positive, but to range from about 0.6 to 0.8 (Kamphaus, 2019).

Phenomena are stable features of nature that form the target of explanatory theories. Though they are not necessarily directly observable, they are evidenced by observations or measurements from particular instrument (Bogen & Woodward, 1988). In psychology, examples of phenomena are patterns in human cognitive development, the concurrence of certain men-

[2]Note that, though we give a single example of a correlation between continuous variables, the notion is not meant to preclude distributions of single variables, other types of variables, or other kinds of associations.

tal disorders, or reliable effects of interventions. These phenomena can be represented by patterns in data (i.e., forms of distributions or associations between variables). For example, the common concurrence of depression and anxiety (the phenomenon) is evidenced by a robust positive correlation between depression scores and anxiety (the statistical pattern). Like statistical patterns, phenomena are not particular, single observations, but general. They generalize across particular individuals, times, places, or modes of assessment. In other words, they possess at least one general aspect, such as generalization across time for a single individual, or generalization over a group of people at a particular location, or, possibly but not necessarily, a combination of generalizations across populations, times, locations, and modes of assessment. Phenomena accommodate a certain degree of non-empirical content, such as boundary conditions (e.g., humans only), methodological assumptions about measurement instruments (e.g., operation of MRI machines), and accepted theories (e.g., biological evolution, mammalian physiology). Phenomena can be evidenced qualitatively from observations (e.g., in the case of animal diversity). However, in psychological science, it is typically the case that evidence follows an indirect path from observations through data to a statistical pattern, which is considered evidence for the phenomenon.[3]

Formal Models are precise statements about components and relations among them that represent some system. For our framework, we consider formal models stated in mathematical equations or code in a programming language.[4] Well known examples are, for instance, climate models that are being used for the investigation and prediction of global warming, traffic models for prediction and improvement of traffic flow, and contagion and recovery models for predicting and intervening on epidemics. Formal models represent the components and relations of the subject matter that are deemed relevant, while surface details and minor variations are removed through abstraction and idealization (e.g., assuming frictionless motion, ignoring social dynamics in contagion processes). These formal models are capable of producing data that show distinct statistical patterns.

Verbal Theories are general conjectures about the world; assertions of the existence of particular entities, dimensions, structures, or relations—specifically dynamical and causal relations—with the goal to explain phe-

[3]We attempt to be inclusive with our characterization of phenomena and conform to the perspective of Bogen and Woodward (1988). However, we acknowledge the possibility that the characterization might exclude some that others consider to be phenomena.

[4]Other types of formal models and ways to formalize scientific theories may exist, but these approaches do not allow one to assess the theory's capacity to explain phenomena within our productive explanation framework.

nomena.[5] In essence, a theory consists of a set of theoretical assumptions about the world, which can serve as guidelines for building or imagining a hypothetical model in which the theory is true. Theories typically support a verbal narrative that serves to 'make sense' of phenomena.[6] In addition, theories usually carry implications on what would happen under various interventions, i.e., causal information (Pearl, 2009; Spirtes, Glymour, & Scheines, 2000; Woodward, 2003). For this reason, theories often support inferences about the "natural" course of systems, as well as how they would change upon causal manipulation. In short, well-developed scientific theories facilitate the classic scientific goals of understanding, prediction, and control. However, verbal theories are typically not specific enough to clarify if and how well they explain their phenomena. To address this limitation, we propose a circuitous route via formal models and statistical patterns.

2.2 Productive explanation

We propose a production-based account of explanation: a framework that can be used to assess if and how well a psychological theory explains a phenomenon.[7] Productive explanation provides a clear connection between a verbal theory and a phenomenon through the use of two intermediaries (a) a formal model that is an explication of the theory and (b) a statistical pattern that represents the phenomenon (top part of Figure 1). To achieve this connection between verbal theory and phenomenon, productive explanation involves three steps.

Step 1: Represent the phenomenon as a statistical pattern. First we clarify what the theory is supposed to explain. In our framework, the explanandum is a phenomenon of interest (mid-left of Figure 1), which is represented by a statistical pattern (mid-right in Figure 1), abstracted from multiple datasets (e.g., a mean difference between two groups in a particular direction). This requires the translation of the commonly qualitative representation of phenomena, in conjunction with relevant theoretical and methodological considerations, into the characteristic statistical patterns by which they are established. In the same way that the formal model provides a statistically tractable explication of the verbal theory, the statistical pattern provides a

[5] We intend empirical, scientific theories that are supposed to explain what we observe. Not all theories have explanatory goals (e.g., legal theories, or normative theories in philosophy), but they are outside the scope of this paper.

[6] There are examples of theories that imply phenomena without making sense of them, the best known example being quantum mechanics.

[7] This is not meant to be a general definition of explanation in science. We specifically call our perspective 'productive explanation' to acknowledge the existence of other approaches—see Section 6 for a comparison.

quantitative representation of the phenomenon (indeed, the phenomenon is often discovered by means of establishing the existence of such patterns, cf. Bogen & Woodward, 1988). This is something that psychological scientists do on a regular basis. For example, armed with theoretical assumptions on mental disorders and diagnostic tools, researchers have inferred from various datasets the phenomenon that depressed mood, anhedonia, low energy, and low self-worth commonly co-occur, which finds representation in a correlation matrix with four variables and positive correlations ranging from about 0.3 to 0.6. The distinction between statistical patterns and their link to the phenomena they represent is thus informed by empirical research. The stability of the phenomenon as a target of the explanation increases to the extent that the phenomenon is robustly evidenced by data and supported by adequate statistical models of data (cf. Suppes, 1960).

Step 2: Explicate the verbal theory as a formal model. In our framework, the explanans is a verbal theory (top-left in Figure 1): a narrative that contains general conjectures about relevant parts of the world, their structure and their (causal) relations to each other. Typically, such a theory would make the occurrence of a phenomenon *prima facie* plausible. To evaluate whether it is actually an explanation of the phenomenon, we *explicate* it into a formal model (top-right in Figure 1). This requires the representation of crucial attributes and relations postulated by the theory in a formal system. That is, we substitute the inexact concepts in the verbal theory by a set of exact terms that are more suitable for development in a mathematical model (Carnap, 1950, p. 3). As theories typically aim at explaining many different phenomena, it may not be necessary that the entire theory is explicated in a formal model. This formal system can take various forms; e.g., that of a probabilistic causal model, a dynamical systems model, or an agent-based model. What is important is that processes and relations encoded in the model are faithful to the relevant processes and relations postulated in the theory and that the formal system provides a means for analytic derivation or simulation from the model.

Step 3: Evaluate whether the formal model produces the statistical pattern. To complete the link between theory and phenomenon, we analytically derive or simulate the behavior that follows from the formal model and evaluate whether the formal model *produces* the statistical pattern representing the phenomenon of interest.[8] The production relation thus relates the two quan-

[8]Note, the productive relation is between the formal model and the statistical pattern. This is distinct from production of the phenomenon in Salmon's (1984) causal-mechanical theory of explanation, or production according to a well-defined mechanism (see Section 6 for discussion). In addition, the successful production of the data pattern should not be

titative counterparts of theory and phenomenon: the formal model and the statistical pattern. We speak of production when the results generated by the formal model of the theory are sufficiently similar to the phenomenon's data pattern, which could be assessed with standard statistical methods (e.g., equivalence tests).

All three steps can (and often will) contain pragmatic and value-driven elements. Specifying the phenomena and statistical patterns that represent them, choosing an adequate model for explicating the theory, and setting up assessment criteria for explanatory power all depend on the researchers' (scientific) goals and interests. This distinguishes our account from proposals inspired by a positivist picture, where explanation is essentially a logical relationship between sets of sentences.

For an adequate explanation, the link between the qualitative and quantitative counterparts needs to be tight. This is what we call *anchoring*: to the extent that the formal model is a good explication of the theory we say that the formal model is *theoretically anchored*, and to the extent that the statistical pattern adequately represents the phenomenon, we say that the statistical pattern is *empirically anchored* (see Figure 1). If the formal model and statistical pattern are appropriately anchored, and the formal model produces the statistical pattern, the theory can be said to explain the phenomenon. This does not mean that the explaining theory is true, or that explanatory power is a binary concept: rather, it comes in degrees and depends on the strength of the anchoring and production relations. In Section 4, we discuss a set of salient criteria for assessing the goodness of an explanation.

Finally, we comment on why we use the term 'production' instead of 'description', 'prediction', and 'causation'. While theories typically describe the properties of and relationships between elements of a target system, not all such descriptions have explanatory power with respect to the target phenomenon (see also *empty formalism* in Section 5).

Neither are all successful predictions explanatory. Two variables may be highly correlated and predictively informative (e.g., playing golf and driving an expensive car), but this does not mean that playing golf *explains* why people buy expensive cars, or vice versa (see also the *empirical relevance* criterion; Section 4.5). Thus, productive explanation cannot be identical to "successful prediction". Specifically, a formal model that "explains" car ownership patterns in terms of the owner's sports preferences would have poor theoretical anchoring: it is not based on a plausible general narrative of consumer behavior, and it leaves out economic resources as a crucial common cause.

equated with the confirmation of a theory in the hypothetico-deductive sense of empirical research.

Scientific theories, even if informal, typically provide causal information indicating the (directional) effects that we can expect from interventions on the target system. Such information acts as a safeguard against spurious "explanations" based on statistical associations only.

The car/golf example does suggest that productive explanation will usually amount to some form of causal explanation. In this case, the narrative conveyed by a sociological theory determines the causal dependencies between the relevant variables (i.e., both depend on income). Thus, we know which variables we need to manipulate in order to produce the statistical pattern that represents the phenomenon. However, many explanations in science involve a substantial non-causal component (Reutlinger & Saatsi, 2018). Consider the following case of explanation in biology, taken from Baker, 2005. The cicada's life cycle period of 17 years is said to be explained by the evolutionary principle of minimizing intersection with other (nearby/lower) periods as being advantageous, ecological constraints that limits the cicadas life cycle to periods between 14 and 18 years, *and* the number theoretic theorem that prime periods minimize intersection. Though this example contains causal relations (e.g., predating and other selection pressures), an essential feature, the primality of 17, is non-causal. Yet, to the extent that a formal model explicating these principles is anchored in a general biological theory, it fits the productive explanation model. Our framework can thus accommodate "mixed" explanations that combine a causal factor with mathematical, stochastic or other types of explanation.

In other words, the productive explanation model is not committed to a particular account of explanation in science, e.g., causal, statistical, mathematical or functional explanations. Rather, it provides a general account for explanatory relationships between theory and phenomena in which specific types of explanation may help to strengthen the theoretical anchoring, the empirical anchoring or the production of the phenomenon. Section 6 discusses these questions in greater detail and compares our account with standard theories of explanation in philosophy of science.

2.3 Summary

The three steps of productive explanation can be summarized in the following definition:

> Theory T explains phenomenon P (to some degree) if and only if there are a statistical pattern S and formal model M such that

1. the formal model M is an adequate explication of the relevant parts of theory T;
2. the statistical pattern S is an adequate representation of phenomenon P;
3. the formal model M produces—possibly jointly with auxiliary assumptions—the statistical pattern S.

As this definition makes clear, productive explanation is *not* defined via a direct link between theory and phenomena, such as the requirement that the phenomenon can be derived logically or mathematically from the theory (e.g., Hempel & Oppenheim, 1948). Such a strict criterion would be inadequate to the verbal and imprecise nature of many phenomena and theories in psychological science and miss out on many cases of valid psychological explanation. Instead, productive explanation is defined indirectly, via a relation of production that holds between formal models that explicate (part of) a theory, and statistical patterns. In this way, we give a precise and operational meaning to what it means that a psychological theory explains a phenomenon, taking into account that most scientific practice proceeds on the level of defining formal models and making inferences with them.

Of course, not all explanations along these lines are equally good. The quality of an explanation can vary according to the strength of the links between the elements of the model, and can yield explanations with higher or lower explanatory power. In Section 4, we introduce criteria for evaluating the quality of an explanation. But first, we show how our productive explanation works in practice.

3 Productive Explanation: A Case Study

We demonstrate in this section how one can use the productive explanation account to analyze the explanatory power of a psychology theory. We chose the regulatory resource theory of ego-depletion, because it is based on a well documented verbal theory and there is a large body of literature on the statistical pattern that the model needs to produce. In addition, the multilab replications by Hagger, Chatzisarantis, et al. (2016) and Vohs, Schmeichel, et al. (2021) allow us to discuss the consequences of choosing different statistical patterns for representing the phenomenon that the theory is supposed to explain, and the connection between explanatory power and overall evaluation of a theory.

The regulatory resource theory is about self-control. Generally, self-control refers to the cognitive processes that inhibit and suppress immedi-

ate thoughts, impulses, expression of emotions, and behaviors that deviate from standards or goals (e.g., Inzlicht, Schmeichel, & Macrae, 2014). The regulatory resource theory of ego-depletion postulates that one's ability for self-control depends on a finite resource. According to the theory, this resource can be interpreted as a stored supply of energy, for which some use the term *willpower* (Baumeister & Tierney, 2012). Every act of self-control reduces the amount of willpower, so that one's ability to inhibit or suppress the next unwanted impulse is diminished (Vohs & Baumeister, 2004).[9]

The regulatory resource theory is said to explain the phenomenon of ego-depletion. Based on the this theory, one expects people to perform worse on a task that requires self-control, right after having done something else that used this resource. This phenomenon is the state named *ego-depletion* (Baumeister, Bratslavsky, et al., 1998). It has been conjectured that all responses of mental control, impulse prevention, emotion regulation, and behavioral guidance dependent this resource. The ego-depletion phenomenon has been a source of considerable controversy in recent years and, as previously mentioned, there is good reason to question whether it is actually a phenomenon (Hagger, Chatzisarantis, et al., 2016). We will show that this controversy largely is due to the fact that the explanatory link between the regulatory resource theory and ego-depletion is relatively unclear.

3.1 Step 1: Representing the Phenomenon as a Statistical Pattern

In the first experiments on ego-depletion, people were asked to wait in a room with a plate of cookies and a plate of radishes. The control group was allowed to eat what they wanted, but the other group was only allowed to eat the radishes. In the consecutive task, both groups were asked to solve geometrical puzzles. Unbeknownst to the participants, these puzzles were unsolvable and the experimenters timed how long individuals persisted until giving up. The control group persisted more than twice as long as the radish-eating group (Baumeister, Bratslavsky, et al., 1998). This setup has become generally known as the *sequential-task experimental paradigm*. Typically, in the first task, one group gets the assignment to suppress some response (e.g., emotional expression when viewing a dramatic movie) while the other group is free to respond as desired. In the next task, some mental effort in the form of self-control is required of the participants (e.g., stifle laughter in response to skits by Robin Williams) and the performance or persistence is compared between groups. Specifically, the statistical pattern is a mean

[9]For simplicity, we here leave out an additional part of the theory in which one's available willpower can be increased through use or training (Vohs & Baumeister, 2016).

difference in the outcome of the second task between the control group and ego-depleted group (e.g., mean number of laughs when watching skits by Robin Williams) in favor of the control group, which is large enough to be detectable in a typical psychological experiment.

3.2 Step 2: Explication the Regulatory Resource Theory in a Formal Model

To examine whether the regulatory resource theory can explain ego-deletion, we developed preliminary formal models based on the verbal theory and tested if they produced the statistical pattern of the sequential-task paradigm. To showcase the process of explicating the verbal theory, we have separated the model development in three distinct models in order of complexity and completeness. The formal nature of these models allows us to describe the consequences of the regulatory resource theory in a precise manner. The code to reproduce the models, figures and analyses can be found at https://github.com/jmbh/explanationpaper.[10]

The key theoretical principles of the verbal theory are that (a) there is a domain-general "resource" called *willpower* (W), (b) tasks requiring willpower deplete this resource (e.g., a task that requires self-control depletes the resource, but a task that requires no self-control does not), and (c) the resource is replenished at some rate after the task ends. To these, we add the following auxiliary assumptions: willpower is a non-negative valued quantitative variable ($W \in \mathbb{R}^+$), willpower changes continuously in time, and willpower has a default value when at rest (i.e., when there is no task that requires willpower; W^r).

3.2.1 Model 1: Simple depletion

The core of ego-depletion theory is that willpower decreases when performing a task requiring willpower. To mathematically model this process, we can use a differential equation that defines how willpower changes over time. Stated in terms of change, the change of willpower (dW) over a small time interval (dt) is determined by the degree to which a task i demands (δ) willpower at time t: δ_t^i. The simplest way to implement the process of ego-depletion then is:

[10] We also developed two additional models that also included more complicated non-linear relations and the possibility to improve maximum willpower through training. However, as these elements are not necessary to produce the phenomenon, we decided not to include them. These models can also be found at https://github.com/jmbh/explanationpaper.

$$\frac{dW}{dt} = -\delta_t^i \tag{1}$$

This model holds that, as a person is completing task i at time t, that person's willpower linearly decreases at a rate of δ_t^i: the longer the task takes, the more willpower is depleted. Note that in order to simulate from this process we also need to specify an initial state of willpower (W_0), which for instance could be its default value at rest (W^r).

Note that the effect of task demand (δ) on willpower does not depend on the current level of willpower (W). Consequently, willpower can become negative. To prevent this, we can add a condition: if $W < 0$ then $W = 0$. This condition implies that the rate of depletion is constant until willpower is fully depleted. However, a linear decrease in willpower, as a function of sustained task demand, implies a drop to exactly zero and implies that individuals can *completely* deplete their willpower. A more plausible model, which avoids this implication, is a model of gradual depletion to which we turn next.

3.2.2 Model 2: Gradual depletion

To prevent the model from suddenly collapsing into a state in which willpower is completely depleted, the model can be adapted such that the rate of depletion slows down with diminishing willpower. This leads to a gradual decline of willpower as a function of sustained task demand. Such a model can for instance be implemented by multiplying the task demand at time t with the amount of willpower remaining at that time:

$$\frac{dW}{dt} = -W\delta_t^i \tag{2}$$

This has the effect that when W is small, also the rate of change dW/dt is small. In Model 2, depletion is the product of the demand and the amount of willpower still available and the equilibrium state is $W^* = 0$. This is because for $W = 0$ the rate of change dW/dt is equal to 0. Regulatory resource theory seems to be silent on this choice between Models 1 and 2, but, it has important implications for how we should expect willpower to behave over time.

3.2.3 Model 3: Recovery

Models 1 and 2 are not adequate explications of the verbal theory. They fail to incorporate a core principle of regulatory resource theory, namely that willpower replenishes. A model that implements this should include

a process that leads willpower to increase again after the task demand is lifted. There are multiple ways to implement this principle in the model. A straightforward way of including willpower recovery, building on the gradual depletion equation, is:

$$\frac{dW}{dt} = -W\delta_t^i + \rho(W^r - W) \tag{3}$$

Here we introduced W^r, which is the amount of willpower one returns to once it is fully replenished. We then add a term $\rho(W^r - W)$ that implements the recovery process, because it increases the rate of change as long as $W < W^r$. Similarly to depletion, this implies that replenishing is first strong but then slows down when approaching the maximum level of willpower W^r. For our purposes here, we set $\rho = 0.10$ and $W_0 = W^r = 1$.

Model 3 implements each of the theory's key principles and we therefore choose to evaluate this model. However, to explicitly connect the theory to experimental conditions, we should be able to simulate expected findings from the sequential-task paradigm, and this requires additional *auxiliary assumptions* regarding the precise operating conditions of the experimental procedure.

First, we must make assumptions about the level of task demand in each of three phases of the task. We will assume that in Phase 1, the demand of the control task is lower than that of the experimental task (0.05 versus 0.20). In Phase 2, which we consider to be more demanding, the tasks given to the experimental and control conditions are identical; hence, in Phase 2, the demand of the control task is set to be equal to that of the experimental task (0.40). In Phase 3, there is no task demand for either group.

Second, we must make assumptions about the relationship between willpower and task performance. This requires the specification of a measurement model. Here, we use a standard function from Item Response Theory to map resource-level to task-performance (e.g., Lord, 2012):

$$P(\text{Task}_t^i = \text{correct}) = \frac{1}{1 + e^{-\alpha(W_t - \beta^i)}}, \tag{4}$$

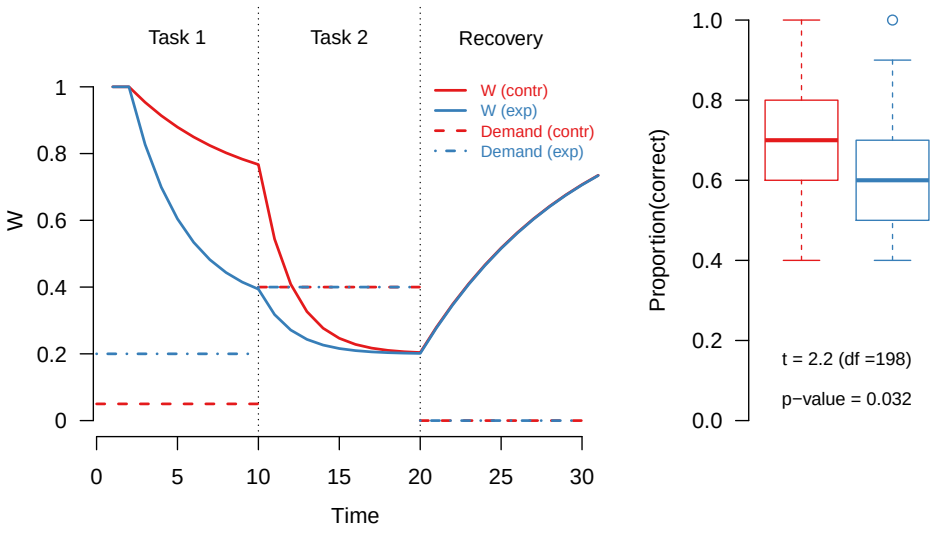
which specifies that the probability of a correct answer on a binary choice question[11] to the task at time t is a function of the available willpower at that time W_t and the difficulty of the task, weighted by parameter α. We set the parameter $\alpha = 2$ and $\beta = 0$. Higher levels of α would make the task

[11] We decided to use the scenario where participants can answer items on the task either wrongly or correctly, because we found this to be the most simple scenario and adding more response categories would not have changed the outcome.

performance depend more strongly on the available willpower, where when $\alpha = 0$ the task performance is independent of the resource level.

Together, the model and these formalized auxiliary assumptions allow us to determine what the theory predicts should occur in the sequential-task paradigm. The right panel of Figure 2 shows that expected task performance in the control condition (N=100) is higher than in the experimental condition (N=100).

Figure 2
Formal Model of Ego-Depletion



Note. The depletion of willpower across time in Model 3. The left side of the figure shows changes in willpower and task demand across time. In Task 1, the control group and experiment group differ in task demand. In Task 2, the task demand is increased and set equal for both groups. In Recovery, the task demand for both groups is set to zero. The right side of the figure shows the results of a t-test when 100 subjects per group are simulated from the model during their completion of Task 2.

3.3 Step 3: Does the Regulatory Resource Theory Explain Ego-Depletion?

Thus far, we have built a model from which we have simulated data. To evaluate whether a model produces the statistical pattern, one needs to check if the statistical patterns of the phenomenon matches the statistical pattern produced by the formal model explicating the theory. From any fully specified model one can either simulate data or derive its implied statistical pattern(s) directly. For analytically tractable models it is possible to derive the model-

implied patterns. Similarly, if the formal model is a structural equation model, one can obtain the model-implied covariance matrix, which could be used in cases where the statistical pattern concerns correlations or covariances (e.g., if the statistical pattern is that correlations between items are positive). However, in many cases one doesn't have analytical access to whether the statistical pattern is implied by the model. In those cases, the model's implications can be accessed and analyzed through simulations.

In the ego-depletion example, the statistical pattern, anchored in the phenomenon, is a mean difference between the two conditions in favor of the control condition that is sufficiently large to be detectable in most typical psychological experiments. The data that we simulated from the model can be analyzed in the same way as empirical data would be analyzed to find evidence for the phenomenon. So, we simulated from the model 200 observations (100 for each condition), which can be considered a typical sample size for a study on ego-depletion. As the phenomenon is represented as a difference in means, we conducted an independent samples t-test, which shows that there is a significant mean difference between the two conditions under certain parameter choices.[12] As such, the production of the statistical pattern that represents the phenomenon can be determined as a statistically significant difference between the samples.

For such simulated datasets one has the same uncertainty in inference that one would have for inferring the statistical pattern from empirical datasets that are used to evidence the phenomenon in the first place. That is, one can use statistical tests on the simulated data, and one can repeat the simulations to get more certainty on whether the presence of the pattern in the simulated data is coincidental or is characteristic for data simulated from the model.

In summary, Model 3 embodies all principles of regulatory resource theory and produces the statistical pattern that represents the phenomenon of ego-depletion. Thus, the regulatory resource theory explains ego-depletion to some degree as we have succeeded in establishing a productive explanation that connects regulatory resource theory to the phenomenon of interest, by means of a formal model that explicates the theory and produces the relevant statistical pattern.

Three interesting and perhaps unanticipated consequences of the theory come to light in Figure 2. First, the statistical pattern can only arise if the

[12] Alternatively, the phenomenon could be represented as a range or distribution of effect sizes. The test could then concern whether the effect resulting from simulated data falls within this range (e.g., an equivalence test). However, without a systematic review of the literature, we could not express the phenomenon in these terms.

demand of the first task outweighs the recovery rate of willpower. If this is not the case, then willpower will recover more quickly than the task can deplete it, and no differences between conditions will materialize. Second, the extent to which task demand needs to outweigh recovery rate depends on the duration of the task. For shorter tasks, this outweighing needs to be stronger, because the task has less opportunity to deplete willpower. Third, as Phase 2 progresses, the difference between the two groups diminishes. This means that if this phase of the experiment takes too long, results will not provide evidence in favor of the phenomenon, even though it does exist. As a result, representing the ego-depletion phenomenon as a mean difference, without relating that difference to task demand and duration could be considered a (partially) incorrect pattern (see Section 5).

These observations indicate that the opportunity to observe ego-depletion effects depends on a quite delicate balance between parameters that characterize the interplay between task demand and willpower. This has important ramifications for the interpretation of experimental results. For example, from the current perspective, not finding a mean difference, like in the multilab replication studies (Hagger, Chatzisarantis, et al., 2016; Vohs, Schmeichel, et al., 2021), counts as evidence against regulatory resource theory only if the experimental paradigm instantiates the conditions outlined above. This serves to emphasize the importance of this type of theoretical work: without a clear idea of whether the theory actually explains the experimental effects tested in empirical research, a failure to demonstrate such effects severely limits the extent to which the theory can actually be tested.

4 Evaluating the Quality of Explanations

If the formal model produces the desired statistical pattern, the theory is said to explain the phenomenon. However, this does not mean that the explanation has high explanatory power. There is a rich philosophy of science literature on what makes for a good explanation (see e.g., Ayer, 1925; Friedman, 1974; Glymour, 2020; Kitcher, 1981; Lipton, 2004; Mackonis, 2013; Thagard, 1988; Ylikoski & Kuorikoski, 2010). In what follows, we do not give an exhaustive list of criteria, but we discuss three criteria that are particularly relevant for the productive explanation approach in terms of explanatory power. The first criterion, *precision*, refers to the strength of the theoretical anchor between verbal theory and formal model (see Figure 1). If a verbal theory is precise, it specifies many aspects of a formal model. Therefore the

range of formal models that are consistent with the verbal theory is relatively small. The second criterion is *robustness*: it specifies to what extent the formal models consistent with the verbal theory produce the phenomenon. A key aspect of evaluating robustness is to assess the impact of auxiliary assumptions that are necessary to specify a formal model but are not specified by the theory. Finally, *empirical relevance* captures to what extent the theory is necessary for the explanation of the phenomenon, in the sense that the phenomenon should not be explainable from background knowledge alone. We would like to stress that these criteria are essentially about the strength of the links between theory, models and (representations of) phenomena in the productive explanation model: they do *not* state whether the theory in general, or the explanation in particular, is plausible, verisimilar or scientifically respectable in the first place.

4.1 Precision

One of the steps in the productive explanation framework involves anchoring the formal model in the theory (Section 2.2). This anchoring is stronger, if the theory is precise in the sense that it specifies in detail the components of the theory, and how they are related in the formal model. When creating a formal model of a precise theory, there are less additional arbitrary decisions one has to take.[13] In contrast, an imprecise theory requires many of such decisions. Precision can be seen as defining the range of formal models that are consistent with the theory in terms of, for instance, ranges of parameter values and discrete alternatives for parameters and functions. If precision is high, only a narrow range of formal models will be consistent with the theory. In contrast, if the theory is imprecise, it yields a relatively wide range of formal models and many additional decisions are necessary.[14]

To illustrate precision, consider the mutualism theory (van der Maas, Dolan, et al., 2006), which offers an explanation for the phenomenon that scores on different cognitive tasks are robustly positively correlated. Mutualism theory explains the 'positive manifold' between cognitive test scores by

[13]Note that such decisions can be informed by empirical results, for instance setting certain parameter values in the model. Of course, such specifications should not be considered part of the (precision) of the theory and thus be subjected to robustness assessment (see Section 4.3).

[14]Note that precision does not only pertain to quantitative relations. For instance, the theory of biological evolution can be precise in the sense of stating the existence of two sexes for sexual reproduction and the process of meiosis. However, this does not mean that the parameter values of the model that explicates this theory are also specified. For now, we are agnostic about the prospect of whether future psychological theories will be precise to the extent of specifying the parameter values of their formal model and remain indifferent to any normative claim that we should strive to reach such a level of precision.

proposing positive reinforcing relationships among basic cognitive processes (perceptual, memory, decision, etc.). According to the productive explanation account, the mutualism theory explains the positive manifold between cognitive test scores only if a formal model anchored in this theory *produces* a statistical pattern of positive correlations between cognitive processes. The formal network model presented by van der Maas, Dolan, et al. (2006) is anchored in this theory, because it implements its components in a way that is consistent with the theory. For example, the cognitive processes directly effect each other (implemented by edges between the variables) and the relation between cognitive processes is that of reinforcement (implemented by making the edges in the model positive). However, the theory does not specify the strength of the positive relations, or whether the reinforcement relations can differ in strength, and how these strengths are distributed. A more precise theory specifies more of such details and therefore fewer formal theories are consistent with it. While this is not necessarily the case, the smaller range of formal models associated with precise theories will frequently produce narrower statistical patterns. For example, specifying that all edges in the network are positive limits the statistical patterns that we can produce with such a model to positive correlations.

Sometimes the formal model also needs to incorporate *auxiliary assumptions* to enable the production of data. These auxiliary assumptions are additional theoretical and methodological assumptions that are independent of the theory. For instance, a model of a socio-psychological theory of how people interact and interpret each other's behavior might also require the specification of modeling choices in sense perception, perceptual processing, and relevant measurement instruments. Such assumptions are not part of the theory, but rather color in more of the background and together with the theory enable the prediction of data. As there might be more than one way to capture these auxiliary assumptions in the formal model, their alternative explications contribute to the range of formal models that are consistent with the theory, but do not impact the precision of the theory.

The arbitrariness of modeling-decisions and multi-explicability of auxiliary assumptions can be revealed in a many-modeler approach (e.g., van Dongen, Finnemann, et al., 2022), where multiple research teams construct a model from the same verbal theory independently of each other. For instance, a table that sums up the choices that are derived from the theory and the arbitrary modeling decisions could then serve as a proxy of precision, thus allowing us to gain a more detailed and transparent insight into psychological theories.

Finally, we take precision as the opposite of vagueness rather than of generality. Formalizing a theory forces one to be precise about what the theory pertains to. If a theory pertains to more objects/contexts/choices of measurement, the theory is more general (this is similar to what Popper, 1959, p. 106, calls *degree of universality*). A highly general theory can therefore also be precise. For example, as we saw in the ego-depletion case study (Section 3.3), the production of the statistical pattern depends on the duration of the task. If one adds to the theory that the ego-depletion effect should only hold for a small range of task durations, this would make the theory less general whereas if the theory states that the ego-depletion effect is present for all task durations (that is, the explanation is not conditional on a certain task duration), this would be a more general theory. In both cases, adding such information to the theory makes the theory more precise compared to a theory that does not mention the role of the duration of the task at all. Precision and generality are thus not opposites of the same dimension.

4.2 Precision in the Case Study

In our case study in Section 3, we were forced to make several ad-hoc choices in specifying the formal model. While these decisions are necessary to derive predictions from the theory, they are not determined by that theory. In effect, the theory provides willpower, task difficulty, recovery, and only some minor restrictions on limits and relations. For instance, the simple depletion process of Model 1 includes a linear decline of willpower, which has some implausible consequences. Thus, in moving from Model 1 to Model 2 (Equation (1) to Equation (2)), we added the assumption that the current rate of depletion depends on the current amount of willpower. This was not specified by the theory, and so one could also choose to leave it out, resulting in a different shape of the depletion of willpower over time. In addition, one could choose various other models that would lead to different shapes of the depletion function.

In a next step, recovery was added to the model. The recovery of willpower is motivated by the theory, but it does not contain an indication of how fast it recovers (the value of r in Equation (3)). The theory also does not specify whether willpower recovers only if $W < W^r$, only if a person is not doing a task, only if task demand is low, or in any combination of these conditions. While the model we built incorporates the first of the above criteria, one can change the model to incorporate the other conditions without being inconsistent with regulatory resource theory.

As a result, we consider the regulatory resource theory to be low on precision, in that it allows for a broad range of different models that potentially show different behavior. Thus, we can conclude that many of the modeling decisions made in Section 3.2 were not theoretically anchored (e.g., the inclusion of additional terms in functional form of model relations, or the choice of parameter values).

4.3 Robustness

A precise theory specifies in great detail the components of a theory and how they are related. This leaves us with fewer arbitrary modeling decisions and thus with a relatively narrow range of formal models that are consistent with the theory at hand. Yet, even for the most precise theory, there will be auxiliary assumptions that will extend the range of consistent formal models. Now, one can assess if the statistical pattern is not only produced by the initial model, but also by other models consistent with the theory, which would indicate that the production does not depend on such auxiliary assumptions or arbitrary modeling decisions that are not part of the theory. Robustness captures the extent to which the phenomenon is produced across the formal models consistent with the theory. If one were to be able to actually enumerate the range of formal models consistent with a theory, one could define robustness as the proportion of formal models consistent with the theory that produce the phenomenon of interest. Robustness in this sense is related to *explanatory generalization*, described by Hitchcock and Woodward (2003), as invariance with respect to, for instance, parameter values (arbitrary modeling decisions) and background conditions (auxiliary assumptions). We illustrate robustness with two examples.

Cognitive-behavioral theories of panic propose that panic attacks arise when the bodily sensations associated with autonomic arousal are perceived as a threat, leading to a 'vicious cycle' between arousal and perceived threat (Clark, 1986; McNally, 1990). Implicit in these theories is the assumption that arousal (and, thus, arousal-related bodily sensations) can rise and fall independent of any effects from perceived threat, presumably reflecting natural fluctuations in arousal arising from internal or external perturbations. To formalize this theory, it is necessary to make these assumptions explicit and posit precisely how arousal fluctuates over time. For example, in one recent formalization, researchers used a correlated noise function with a Gaussian distribution to model fluctuations in arousal, thereby assuming random perturbations with a probability distribution that is unimodal and a quickly decreasing density away from the mode (Robinaugh, Haslbeck, et al., 2019).

Because these aspects of the model are unspecified by the theories that anchor it, the model's ability to produce the phenomena of interest should ideally not be affected by the selection of other plausible implementations of arousal fluctuations. In this case, the production of the phenomenon (and thus the explanation of the phenomenon) is robust against alternative implementations. By contrast, if alternative implementations of aspects of the model that are unspecified by the theory qualitatively change the statistical patterns produced by the model, the explanation is less robust.

Another example is Schelling's (2006) model for segregation in urban areas. The model is built on the theoretical principle that individuals have a preference for at least a certain percentage of neighbors being similar to themselves. If their preference is met, they stay; if not, they move to another random available place in the city. This model has become famous, because it shows that a modest preference of wanting only 30% of one's neighbors to be similar to oneself already leads to high levels of segregation. The implementation of the model, however, requires several choices not specified by the theory. For example, we need to define a grid which serves as the abstraction of a city. If the qualitative behavior produced by the model strongly depended on grid size, the explanation would be less robust. This would not mean that the theory should be discarded, but it would limit the scope of the explanations it could provide: for example, the model might suggest that segregation only emerges from similarity preference in cities of particular size. Accordingly, testing for robustness[15] provides an opportunity to identify the theories' boundaries and limitations and, thereby, to further develop the theory.

4.4 Robustness in the case study

We discussed above that regulatory resource theory is relatively imprecise, because we had to take many arbitrary modeling decisions to implement a formal model. For instance, the theory is silent on how to connect willpower to task performance. We chose to use the IRT function, though we could have chosen many other functions. In a robustness analysis we evaluate to what extent these arbitrary modeling choices affect the evaluation of the phenomenon. This can be done by utilizing methods of sensitivity analysis (Saltelli, Tarantola, & Campolongo, 2000; Smith, 2013), which involve run-

[15] As robustness is defined as the production of a statistical pattern across models, sensitivity analyses and other methods for the analysis of statistical robustness might be applicable here (see Huber, 1981, for a comprehensive account of robustness analysis).

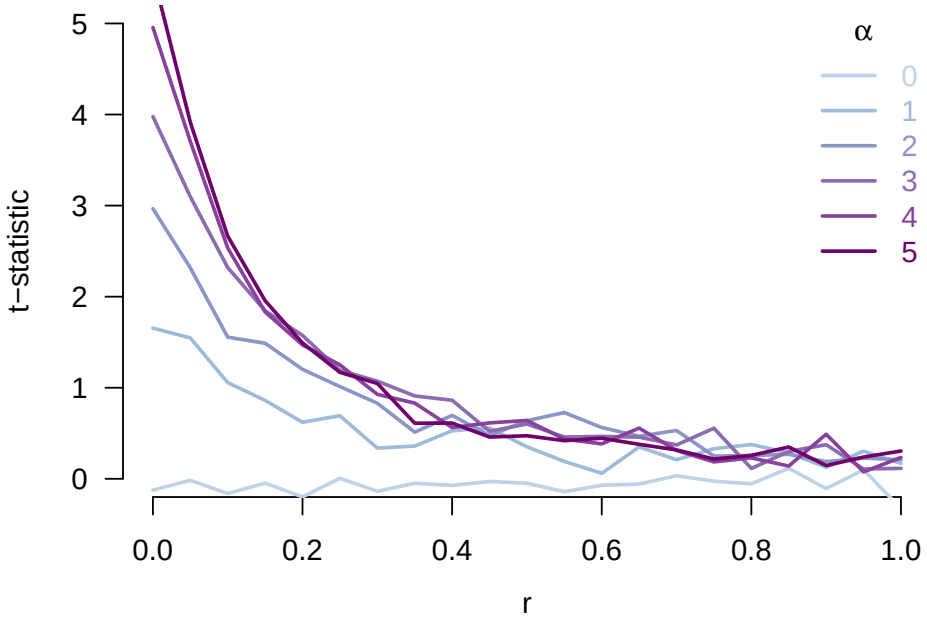
ning many simulations in turn and evaluating their consequences for the patterns in the data.

In the interest of brevity, we only assess robustness across values of two parameters in Model 3, namely the recovery rate r, and the weight α that determines how strongly the resource level affects task performance. For a robustness analysis regarding the parameterization of the model, we need to specify a range for plausible parameter values and subsequently run the model for all combinations of parameter values. We let α range from 0 to 5, with a step size of one and we vary r from 0 to 1 with a step size of 0.1. For every combination of parameter values, we determine if the targeted phenomenon (i.e., a mean difference between experimental conditions) is present. We do so by simulating 100 subjects under the experimental condition setting and 100 subjects under the control setting, a scenario typical for ego-depletion studies. We then perform t-test on the generated data to see if there is a statistically significant difference between the experimental group and the control group ($t > 1.65$ for a one-sided test with 198 degrees of freedom).[16]

Because there is some inherent uncertainty in the model, as the mapping from resource level to correctly performing the task is probabilistic (i.e., Equation 4), we perform 100 iterations per combination of parameter values and average the results. We summarize our results regarding effect size in Figure 3.

[16]propose to perform tests like these because the statistical patterns are not just a matter of a parameter's sign (e.g., mean difference in a particular direction). Alternatively, one could conduct this simulation with very large sample sizes and use a statistical function for an effect size, like Cohen's d, to express the results. Any effect size that would be detectable with "realistic sample sizes" could then count as production of the statistical pattern and as such be used to evaluate robustness.

Figure 3
Robustness of the Productive Explanation



Note. For a range of values of recovery rate r and measurement parameter α, we performed a t-test to evaluate the difference between the average proportion of correct responses in Phase 2 for the control group (n=100) and the experimental group (n=100). The figure represents the values of the average t-statistic over a 100 iterations as a function of r, plotted for different values of α

The results provide insight into the extent to which the theory is robust and the conditions on which robustness depends. We conclude that regulatory resource theory is low on robustness with respect to the two parameters investigated and the typical sample sizes that we assumed in our simulation, because the t-statistic depends strongly on both r and α and only for a narrow range of their parameter values is the statistical significance threshold of $t(198) = 1.65$ reached. As expected, when α is zero, i.e., the task performance does not depend on the resource level, there is no difference between the conditions. The higher α the stronger the effect becomes. When α is above zero, the recovery rate becomes essential. For moderate to high recovery rates (> 0.3) the task is not depleting the resource sufficiently to show significant differences in performance between the experimental and the control group. Notice that, because α is a parameter in a measurement model, while r is a parameter in a substantive theoretical model, the results

highlight the fact that measurement issues and substantive issues can interact in determining whether an effect is indeed present in the data. This underscores the need for adequate psychometric modeling in tandem with theory formation.

Before moving on to the last quality criterion, we would like to reflect on the connection between precision and robustness. In principle, precision and robustness are orthogonal: we can have a theory with high precision, which implies a relatively narrow range of formal models consistent with a theory. If most of those few formal models are producing the phenomenon of interest, this precise theory is robust; if a sizable proportion of the few formal models do not produce the phenomenon, it is not. On the other hand, we could have a very imprecise theory, which implies a wide range of formal models that are consistent with it. Again, we could have the situation in which most of those theories produce the phenomenon of interest, thereby rendering the imprecise theory a robust theory; but it could also be that a large proportion of the many models are not producing the phenomenon, which would give us a theory that is imprecise and not robust. Note that, though precision and robustness might logically not imply one another, in practice it is quite plausible that less precise theories also tend to be less robust.

4.5 Empirical Relevance

The empirical relevance of a theory is the extent to which the particular components of the theory, captured by the model, are necessary for the production of a phenomenon's statistical pattern (Ayer, 1925). If the phenomenon is likely to arise from auxiliary assumptions, alone—e.g., a robust correlation arises due to similarity in the methods used to measure the constructs of interest—then the model is lacking in empirical relevance. This criterion is already present in Hempel and Oppenheim's (1948) D-N model of explanation and also finds expression in Hitchcock and Woodward (2003) as invariance under background conditions. In our account, it is another expression of the "Theoretical Anchor" connection in Figure 1, as the modeling assumptions that embody the theory are irrelevant for the production of the statistical pattern if empirical relevance is lacking. In contrast, if the phenomenon is unlikely to arise given a set of auxiliary assumptions alone, yet is produced by the theoretical model, then the theory has high empirical relevance.

As an example, Krueger (1999) offered a putative explanation for the finding that comorbidity of mental disorders organizes along two main di-

mensions; namely, that all symptoms are directly or indirectly produced by two distinct underlying psychological processes (internalizing and externalizing). In support for this theory, Krueger (1999) showed that a model that represents these processes as two higher-order latent variables explains the patterns of correlations (i.e., statistical pattern that represents the empirical phenomena). However, Borsboom (2002) argued that the patterns can also be explained by the fact that diagnoses of the different mental disorders utilize overlapping symptoms. Thus, in this view, disorders like Generalized Anxiety Disorder and Major Depressive Disorder show higher comorbidity because their diagnoses rest partly on assessment of the same symptoms (e.g., insomnia, fatigue; see also Bogenschutz and Nurnberg (2000). In this case, the theory posited by Krueger (1999) is a plausible explanation, but also an explanation with relatively low empirical relevance, as the empirical patterns found at the level of mental disorders can be produced by the fact that the DSM-IV (American Psychiatric Association [APA], 1994) was used for diagnosing these disorders.

Assessing the empirical relevance of a model cannot be done by just inspecting the model and its predictions. Rather, its about trying to find an alternative explanation by using background knowledge or a subset of theoretical model's assumptions. If these alternative explanatory principles (e.g., measurement choices and experimental designs) are by themselves sufficient to produce the statistical pattern, then this undercuts the empirical relevance of the theoretical model.

The concept of empirical relevance is thus associated with the standard research practice of identifying *validity threats* and ruling out *alternative explanations* of phenomena. In fact, standard examples of internal validity threats in quasi-experimental research, such as confounding, maturation, history, and selection bias (Cook & Campbell, 1979) can all be seen as verbal blueprints for building models that could produce an empirical phenomenon (e.g., a mean difference between conditions) without requiring the core explanatory principles that constitute the theoretical model.

4.6 Empirical relevance in the case study

To assess the empirical relevance of the regulatory resource theory, we evaluate whether the theory is necessary for the production of the phenomenon's statistical pattern. In some contexts, it might be possible to build a baseline model as a stand-in for the background to which the theory should be compared, analogous to how in the context of machine learning simple guessing baseline algorithms (so-called 'featureless learners' that constantly predict

the mean of the target variable in training data) are used as a threshold to which other models are being compared (see e.g., Pargent & Albert-von der Gönna, 2018; Schoedel, Au, et al., 2018).[17] For example, in structural equation modeling, theoretical assumptions can sometimes be expressed in constraints, in which case the unconstrained model can function as a baseline model.[18] Though for the ego-depletion case study, we did not see a way to formalize the regulatory resource theory as a structural equation model and we were thus not able to build such a formal baseline model. We therefore instead approach this question in the following two ways.

First, we can ask whether there are aspects of the theory that do not appear in the model and, thus, are not necessary for the production of the phenomenon. For example, proponents of regulatory resource theory posit a finite resource that is specific to self-regulation. This aspect of the theory is not necessary to produce the specific phenomena we have focused on here. Rather, all that is required is that there is some resource supporting performance that is depleted more by some tasks than others. Accordingly, the theory that a regulation-specific resource is responsible for the sequential-task phenomenon has low empirical relevance, as this resource could just as readily be, for example, one's energy level.

A second lens through which we may examine the empirical relevance of the theory is whether there are auxiliary assumptions or other elements of background knowledge that are not included in the theory. For example, we may hypothesize that participants enter sequential-task studies with an expectation of how much effort is appropriate to dedicate to the study, and that some tasks may expend this effort more readily than others. Similarly, we may expect that being given something of value (a cookie) will increase the effort participants are willing to give, while being denied something of value (being presented with cookies, but told to eat radishes) may decrease the effort participants feel is appropriate to give to the study. If we incorporate this "effort" component into our model, we can evaluate whether the sequential-task phenomenon follows from effort even in the absence of ego-depletion. If so, this would further undercut the empirical relevance of the theory, as the phenomenon would follow from our auxiliary assump-

[17] We thank one of the reviewers for this suggestion.

[18] In such a case, the baseline model, by its very nature, is capable of producing the phenomenon. However, this happens only under particular parameter settings, one of which being our theoretical model. In this case, empirical relevance can be assessed in combination with a robustness analysis of the baseline model. For instance, the theoretical model can be seen as empirically relevant if the baseline model only produces the statistical pattern under constraints similar to the theoretical model.

tions and background knowledge about the world, regardless of whether the theory is true.

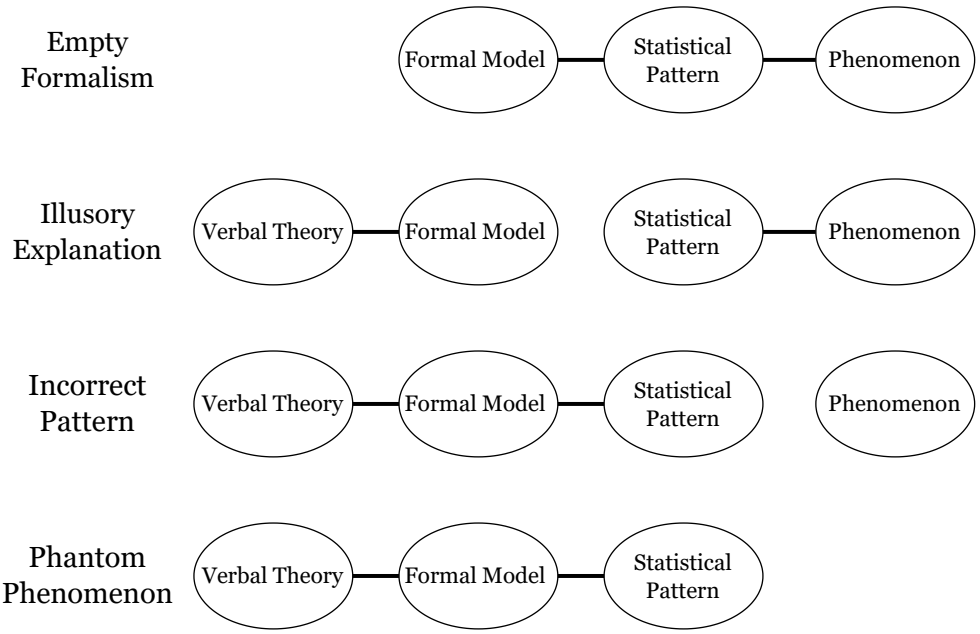
The goal of these evaluations of the case study was not to give a quantitative measure of the goodness of the explanation, but to explicate how the quality of an explanation depends on how the theory is tied to the formal model, and on the conditions where the formal model produces the relevant statistical pattern. Trading off these criteria is a delicate and context-sensitive task beyond the scope of this paper, but they provide a rough overall guideline for when explanations are scientifically valid, and/or superior to competitors.

5 Explanatory Breakdowns

In psychological science, it may often be the case that one or more of the links of the productive explanation model are missing. It is instructive to consider common ways in which explanations can fail, as these failures serve to highlight the ways in which we can improve psychology's explanatory systems using productive explanation. The framework depicted in Figure 1 implies four important cases of explanatory breakdown that require attention, which we have named: *empty formalism, illusory explanation, incorrect pattern, and phantom phenomenon*. Each of these cases is depicted in Figure 4.

Empty formalism denotes either the complete absence of a verbal theory or the absence of any meaningful interpretation of the components of the formal model (e.g., parameters, variables, functions). In this case, we may find a sophisticated mathematical formalism that however is not connected to any theory of interest. An example of such a case is the *critical positivity ratio* (Fredrickson & Losada, 2005): an extensive mathematical model that implies that there is a critical ratio of positive-to-negative affect of 2.9, which purportedly would be useful in explaining human flourishing. However, as has been pointed out in the literature (Brown, Sokal, & Friedman, 2013), the model used is based on properties of turbulence in liquids that have no discernible connection to psychological theory. In this case, the modeling is disconnected from any explanation based in substantive theory. We assert that verbal theories are a necessity. Not only for the interpretations of symbols and relations in the models, but also to highlight connections between separate phenomena in order to help integrate science. Otherwise, we might end up with a disjointed collection of models that purport to explain individual phenomena.

Figure 4
Four Cases of Explanatory Breakdown



Note. This figures represents the patterns of solid arrows from Figure 1. In each of the cases a specific part of the productive explanation chain is inadequate or missing.

Illusory explanation denotes a case in which we are able to anchor a formal model theoretically, and we also have an adequate representation of the phenomena in the form of a statistical pattern, but the model simply does not produce that pattern. Anybody who has ever constructed formal models is familiar with this situation. The human intellect is not capable of assessing the implications of theories beyond a certain complexity. It is easy to believe that an explanation is adequate if the story conveys a sense of understanding; however, this reasoning may be fed by cognitive biases while actually, the model fails to produce the statistical patterns of interest (see also Trout, 2002). In fact, much of the rapidly growing field of complex systems science studies models implementing relatively simple rules that give rise to extremely complex and counter-intuitive behavior. This shows why formal modeling in general is so useful: it complements the limited capacity of the human intellect with a tool that allows us to evaluate whether our stories actually play out as intended.

Incorrect pattern denotes the case in which there is a mismatch between the phenomenon and the statistical pattern that is used to represent it. As an example, one may consider a cyclic pattern as is typically used to model the phenomena associated with bipolar disorder. Such a pattern might be repre-

sented as a strict phase transition in which all positive (negative) symptoms suddenly flip from being present (absent) to being absent (present), whereas the actual phenomena involve a much more complicated and nuanced pattern, in which positive and negative symptoms may be active at the same time. A similar situation occurs when variables change at different time scales, or if the population of interest displays considerable heterogeneity while statistical patterns are represented as uniform; for instance, Cramer, Van Borkulo, et al. (2016) model transitions between depressive and healthy states as uniform, while depression is a likely heterogeneous construct in which transitions arise in different symptom sets for different people.

Finally, *phantom phenomenon* denotes a situation in which we try to explain a phenomenon that does not exist in the first place. Examples of such phantom phenomena unfortunately seem to be prevalent in psychology, as the replication crisis shows. For instance, experimental results previously taken as evidence for facial feedback and behavioral priming are now in doubt as a result of replication failures, calling into doubt the phenomena they are purported to represent. The multilab replications of ego-depletion (Hagger, Chatzisarantis, et al., 2016; Vohs, Schmeichel, et al., 2021) could be interpreted as evidence for the thesis that ego-depletion is in fact a phantom phenomenon. In total 59 groups tested close to 5000 participants, but found no difference in performance between control group and those who were supposedly ego-depleted. That is, apart from whether regulatory resource theory would produce the effect that is referred to as 'ego-depletion', it is doubtful whether this effect exists. Such cases are expensive, as they invite efforts by the scientific community to construct explanatory frameworks for phenomena that turn out to be nonexistent. For this reason, adequate empirical work is essential, and this may involve a reorientation of psychological research from testing theories to establishing phenomena and the conditions under which they occur. While the former rewards successful predictions, especially if these are counter-intuitive, the latter rewards precise documentation of cases in which statistical patterns are or are not exhibited.

Importantly, explicating the links between theories and phenomena is crucial to the interpretation of failed replications. As we illustrated in Section 3.3, specification of the statistical pattern that characterizes the phenomenon, contextual factors, and experimental setup can affect whether or not one will observe the ego-depletion phenomenon in a particular study.

How well do current psychological theories fare with respect to the above cases of explanatory breakdown? This is difficult to assess, because the great majority of psychological explanations has not been formalized. Hence, the

proper stance with respect to this issue is one of agnosticism. However, this agnosticism is itself a cause for considerable concern. First, it is evident that, upon formalization, theories can turn out to lack the explanatory qualities that were assigned to them in their verbal form and thus constitute illusory explanations (e.g., Harris, 1976). Second, the fact that phenomena that have long been taken for granted in psychological science may be represented by incorrect patterns (e.g., exaggerated effect sizes due to questionable research practices) or even turn out to be pure phantom phenomena (e.g., Aarts, Anderson, et al., 2015; Vohs, Schmeichel, et al., 2021) is of course food for thought, as it suggests that a subset of psychological explanations has targeted phenomena that may not exist.

6 Comparison with Theories of Scientific Explanation

Philosophy of science has an extensive literature on scientific explanation. Before concluding, we consider it illuminating to situate our framework within this literature—classical accounts of explanation as well as contemporary proposals. In this section, we compare our productive explanation framework to philosophical theories of explanations to showcase some of its limitations, strengths, and open questions.

Perhaps the most well-known theory of scientific explanation, even if it is somewhat dated, is Hempel and Oppenheim's (1948) *deductive-nomological or D-N model*. Explanans and explanandum are conceived of as sets of sentences, and for an explanation to succeed, the explanans must entail the explanandum (possibly drawing on auxiliary assumptions) and contain at least one true law of nature that is necessary for deriving the explanandum. This model had enormous impact on shaping the discussion on scientific explanation. Productive explanation is consistent with some of its aspects, such as the empirical relevance of the explanans for producing the explanandum, but where we deviate is more important.

First, the explaining theory and the explained phenomenon need not stand in a direct logical relationship. Rather, both are verbal descriptions which need to be *explicated* into formal models, or to be represented by statistical patterns. Second, we dispense with the concept of a law of nature, because they are notoriously hard to get by in the social sciences (Giere, 1999).

Third, we believe that logical and mathematical derivations are too narrow to account for the explanatory connection between the theory and phenomena, and for scientific practice in psychology. Instead, we consider it suf-

ficient that *some* model anchored in the theory produces the pattern that represents the phenomenon. When other models (e.g., different parametrizations) anchored in the theory may fail to produce the statistical pattern, this does not invalidate the explanation completely, but diminishes its robustness and quality.

Fourth, we allow for productive explanations without requiring that the explanans is true, or truthfully captures all relevant mechanisms. Explanatory models and theories are often based on idealizations and abstractions (Batterman, 2002; Weisberg, 2007), and we would be ill-advised to rule out such explanations. Fully correct and verisimilar explanations are also hard to imagine in complex areas of psychological sciences. Just like an argument can be valid (i.e., the conclusions follow from the premises) while it is not sound (i.e., the premises are false), explanatory theories need not be literally true for promoting scientific progress (cf. Meehl, 2002).

The limitations of the D-N model, especially in the field of social sciences and life sciences, have prompted the development of various competing models. Some of them extend the basic rationale of the D-N model—the rationalization of the explanandum in the light of the explanans—toward statistical explanation (Crupi & Tentori, 2012; Hempel, 1965, 1968; Schupbach & Sprenger, 2011), or stress how explanatory facts account for the statistical association between different attributes (Salmon, 1971). However, such attempts neither acknowledge the role of phenomena as the typical target of explanations (with the exception of Ströing, 2018), nor do they distinguish between theories and models at the level of explanantia. While these accounts can provide useful formalizations of specific aspects of the productive explanation model (e.g., how well a formal model accounts for a statistical pattern), they are, in our opinion, too limited and too specialized to figure as a full-scale model of scientific explanation.

More recently, a large number of research articles on scientific explanation tie explanation to causation and stress the causal character of scientific explanation (this tradition goes back to Dowe, 2000; Salmon, 1984). Such attempts are typically part of one of the following three research programs: mechanistic explanations (e.g., Glennan, 2017; Machamer, Darden, & Craver, 2000), interventionist or counterfactual explanations (e.g. Hitchcock & Woodward, 2003; Woodward, 2003; Woodward & Hitchcock, 2003), and kairetic explanation (Strevens, 2004, 2008). We briefly discuss the first two of them and relate them to the productive model of explanation, as these have

a broader basis of support. For further discussion, including a treatment of Strevens's kairetic account, see Ross and Woodward (2023).[19]

According to the mechanistic approach, very popular in philosophy of the life sciences and philosophy of neuroscience, successful explanations describe a *mechanism*, i.e., an organized activity that produces regular changes and has well-defined finish or termination conditions. Mechanistic explanations appeal to the causal mechanism that produces a phenomenon of interest, and in particular, they describe in detail how the parts of a system interact in producing a (typically higher-level) phenomenon.

While mechanistic explanations can without any doubt increase our understanding of a target system and promote scientific progress, they do not provide an answer to the question of how psychological *theories* can explain specific phenomena, or how statistical regularities in data patterns can be accounted for. They are most useful when we wish to explain a high-level phenomenon by means of low-level interactions. Nonetheless, mechanistic explanations can be integrated into the productive explanation picture. For example, mechanistic models are useful for identifying mediators and moderators in formal models of causal inference. More generally, the in-depth description of a mechanism helps us to identify a formal model which in turn produces the statistical pattern that represents the phenomenon.

Interventionist or counterfactual explanations, by contrast, can be abstractly characterized as answers to "what-if-things-had-been-different questions" (Woodward, 2003). This type of explanation is based on causal models, e.g., structural equation models or directed acyclical graphs, and the explanandum typically consists in the behavior of one or more variables in this model. Hypothetical interventions on the cause, represented as a variable in such a graph, show whether the explanandum is crucially dependent on the value of the cause, and whether it co-varies with the cause under a wide variety of interventions and values of the boundary conditions.

Productive explanations need not be counterfactual in this sense, yet causal models are often good explications of a verbal narrative conveyed by the theory. Similarly, they are a good tool for analysing whether a certain data pattern is reliably produced. his is especially true for models where uncertainty is an essential part of the picture, such as causal Bayes nets (e.g., Pearl, 2009; Spirtes, Glymour, & Scheines, 2000). Productive explanations that rely on causal information will often look similar to counterfactual

[19]There are also various non-causal theories of scientific explanation, e.g., a thriving literature on mathematical explanation (Baker, 2009; Lyon & Colyvan, 2008; Pincock, 2012, e.g.) and explanations by abstraction and removal of detail (e.g. Batterman, 2002; Ross, 2015). It would go beyond the aim of the paper to discuss all of them in detail, especially many of them are motivated by phenomena in physics and/or the life sciences.

causal explanations, but our model also provides a general story about how theories, models and statistical patterns are connected in successful explanations. Whether our account of productive explanation always requires a causal component is left open for future research: certainly it matches well with structural equations and other causal models, and all of our examples involve a causal narrative, but it is not clear, and not implied by the concepts used in the definition, that productive explanation must always be based on causal information. To the extent, however, that substantive psychological theories make causal claims, it is likely that at least in the context of psychological science the answer will be yes.

Summing up, the accounts of explanation discussed in current philosophy of science typically answer more detailed questions and presuppose specific contexts for asking explanation-seeking questions. To the extent that these contexts occur in behavioral and cognitive science, they can elucidate some aspects of the productive explanation model—especially in places where we still have to fill in the details (i.e., how formal models are anchored in theory, and how statistical patterns are produced from a formal model). The productive explanation model does not declare these accounts mistaken; rather it tries to adopt a bird's-eye view on explanation in psychological science that applies regardless of whether causal-interventionist, mathematical or mechanistic relations are considered to be of primary importance. On the other hand, specific accounts of explanation can help to evaluate whether or not the proposed explanation is of high quality, e.g., by means of analyzing the theoretical anchoring of the formal model or the actual production relationship.

7 Discussion

The progress of science depends upon cycles of iterative theoretical and empirical development. The replication crisis has illustrated that we cannot be certain as to which empirical phenomena should be considered robust—in fact, the phenomenon modeled in the present paper may well be a phantom phenomenon (Vohs, Schmeichel, et al., 2021). Analogously, several papers on the theory crisis (e.g., Oberauer & Lewandowsky, 2019) have made clear that purely verbal theories make it hard to pin down a particular generating model, which in turn makes it difficult to see what the theory actually implies (Robinaugh, Haslbeck, et al., 2021; Scheel, 2022).

The fact that psychology faces both a theory crisis and a reproducibility crisis at the same time means that we struggle with imprecise theories that

relate to uncertain phenomena in unclear ways—a difficult situation, to say the least. However, the current surge towards improvement of both the robustness of phenomena and the theories that are supposed to explain them have suggested numerous methodological improvements in both empirical and theoretical areas (Borsboom, van der Maas, et al., 2021; Guest & Martin, 2021; Haslbeck, Ryan, et al., 2021; Smaldino, 2020; van Rooij & Baggio, 2021). We see our paper as a contribution to these efforts. In this section, we explore how our work on explanation in psychological science facilitates the process of theory construction and theory development, and we identify open questions and lines for future research.

To recapitulate, the productive model of explanation offers a general account of how theories explain empirical phenomena by means of formal models and statistical representations of said phenomena. It renders the connection between theory and empirical phenomena explicit, transparent, and, thus, available for evaluation. Specifically, a theory explains a phenomenon to some degree if a formal model, adequately explicating the theory (i.e., theoretical anchoring), produces the pertinent statistical pattern, adequately representing the empirical phenomenon (i.e., empirical anchoring). The process of constructing a productive explanation can reveal explanatory breakdowns that arise because one of the elements in the explanatory chain is weak or missing. This motivates the need for quality criteria. We have offered three such criteria for evaluating how well a theory explains its phenomena: *precision*, *robustness*, and *empirical relevance*.

The productive-explanation account strengthens the quality of the theoretical cycle mainly in two dimensions: (1) allowing for a rigorous evaluation of the explanatory power of psychological theories, and (2) encouraging researchers to explicate verbal theories into formal models. The act of building formal models helps to make theories more precise and transparent, and the resulting theories will more easily enable building new or improved models.

Improving explanations in psychological science will also crucially depend on the modeling skills that we teach our researchers. We think that, in this respect, some optimism is appropriate. Formal modeling is simply very interesting and the climate for teaching and developing modeling skills is steadily improving with the increased availability of computational tools. We note that, even though our implementation of regulatory resource theory is highly limited, it already brings out unexpected implications of the theory. For example, the model implies that, even if a manipulation is successful, the difference between experimental and control groups will gradually vanish over time as Task 2 continues. Importantly, this indicates that current fail-

ures to replicate core phenomena (Vohs, Schmeichel, et al., 2021) may not constitute evidence against the theory, simply because it is unclear whether the theory actually implies that these phenomena should be expected in the first place. This type of unexpected implication, in our experience, is quite common in formal modeling and this provides interesting handles for substantive psychologists to engage with, for instance in developing new empirical implications and experimental paradigms.

Even though we have proposed a framework of explanation that can help psychological scientists to construct more explicit and precise explanations, we do not intend a strictly prescriptive reading of the methodology. It is possible that different models of explanation (e.g., mechanistic or causal-interventionist models) provide similar resources to those developed in the present paper; in addition, some phenomena may not lend themselves to being expressed as statistical patterns in data (e.g., some phenomena may be qualitative in nature or refer to general capacities; van Rooij & Baggio, 2021). Thus, while the approach we deliver in this paper should be seen as a comprehensive methodological toolbox that articulates a framework in which psychologists can work, it is not to be read as a normative paper that outlines how psychologists invariably should work. Methodology should be attuned to the substantive context, and psychology may encompass contexts where our proposal is not applicable in its current form.

Finally, we envision several extensions of our production-based account of scientific explanation that may be pursued in future research. First, as theories are typically not evaluated in relation to a single phenomenon, the production based approach needs to be scaled-up to include an arbitrary number of phenomena, which may be explained to different degrees. This raises questions such as: how well does a theory explain if it covers a set of relevant phenomena, but only some are robustly produced by the model? Second, we need to be able to compare the explanatory power of competing theories with respect to a phenomenon. Such an extension could address questions like: how does a theory that explains many phenomena with a low level of robustness compare to a theory that explains a proper subset of these phenomena with a high level of robustness? In other words, future research will have to aim at developing comparative criteria for the explanatory power of scientific theories. It is especially interesting to connect our production-based account of explanation to Thagard's 1988 *explanatory coherence*. Notably, an implementation of explanatory coherence as an Ising model has recently been developed (Maier, van Dongen, & Borsboom, 2021). This approach allows researchers to compare theories using a process

that can take relations between multiple theories across multiple phenomena into account. By equipping us to more rigorously evaluate the connection between theories and phenomena (i.e., the links/edges in the Ising model of explanatory coherence), the production-based account described here (including efforts to evaluate the quality of that connection) may be fruitfully combined with this work in order to rigorously compare competing theories across a range of phenomena.

Last but not least, we hope that the productive-explanation framework will prove useful in the further development of psychological theories, and the advancement of more robust psychological science.

Contribution Statement. The idea was primarily conceived by NvD and DB. It was further developed by RvB, AF JMBH, HvdM, DM JdR and JS. The formal models were developed and assessed by JMBH, HvdM, AF, and JdR. JS assisted with the philosophical embedding of the idea. The original draft was written by NvD, DB, RvB and JS. Revising and editing was done by the entire team. The project was managed by NvD.

Acknowledgements. JMBH has been supported by the gravitation project "New Science of Mental Disorders" (www.nsmd.eu), supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.004.016).

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A., … Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV* (4th ed.). Autor.

Ayer, A. J. (1925). *Language, truth and logic*. Dover.

Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind*, *114*(454), 223–238.

Baker, A. (2009). Mathematical explanation in science. *The British Journal for the Philosophy of Science*.

Batterman, R. W. (2002). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1965.

Baumeister, R. F., & Tierney, J. M. (2012). *Willpower: Rediscovering the greatest human strength*. Penguin.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, *47*(3), 303–352.

Bogenschutz, M. P., & Nurnberg, H. G. (2000). Theoretical and methodological issues in psychiatric comorbidity. *Harvard Review of Psychiatry*, *8*(1), 18–24.

Borsboom, D. (2002). The structure of the DSM. *Archive of Genneral Psychiatry*, *59*(6), 561–569.

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766.

Braithwaite, R. B. (1960). Models in the empirical sciences. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science. proceedings of the 1960 international congress* (pp. 224–231). Stanford University Press.

Brown, N. J., Sokal, A. D., & Friedman, H. L. (2013). The complex dynamics of wishful thinking: The critical positivity ratio. *American Psychologist*.

Carnap, R. (1928). *Der logische aufbau der welt*. Weltkreis.

Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, *49*(3), 609–610.

Clark, D. M. (1986). A cognitive approach to panic. *Behaviour research and therapy*, *24*(4), 461–470.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin.

Cramer, A. O., Van Borkulo, C. D., Giltay, E. J., van der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS One*, *11*(12), e0167490.

Crupi, V., & Tentori, K. (2012). A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems). *Philosophy of Science*, *79*, 365–385.

Cummins, R. (2000). "How does it work?" vs. "What are the laws?" Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117–145). MIT Press.

de Boer, R. J. (2023, October 3). Grind[computer software].

Dowe, P. (2000). *Physical causation*. Cambridge University Press.

Dyson, F. W. (1917). On the opportunity afforded by the eclipse of 1919 may 29 of verifying einstein's theory of gravitation. *Monthly Notices of the Royal Astronomical Society*, *77*, 445–447.

Fortmann-Roe, S. (2014). Insight maker: A general-purpose tool for web-based modeling & simulation. *Simulation Modelling Practice and Theory*, *47*, 28–45.

Fredrickson, B. L., & Losada, M. F. (2005). Positive affect and the complex dynamics of human flourishing. *American Psychologist*, *60*(7), 678–686.

Fried, E. I. (2021). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, *71*(1), 5–19.

Giere, R. N. (1999). *Science without laws*. University of Chicago Press.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Glymour, C. (2020). Probability and the explanatory virtues. *The British Journal for the Philosophy of Science*.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*, 371–388.

Haig, B. D. (2021). Abductive research methods in psychological science. *PsyArXiv*.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part ii: Explanations. *British Journal for the Philosophy of Science*, *56*, 889–911.

Harris, R. J. (1976). The uncertain connection between verbal theories and research hypotheses in social psychology. *Journal of Experimental Social Psychology*, *12*(2), 210–219.

Haslbeck, J. M. B., & Ryan, O. (2021). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, 1–32.

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*.

Hempel, C. G. (1965). *Aspects of scientific explanation*. Free Press.

Hempel, C. G. (1968). Maximal specificity and lawlikeness in probabilistic explanation. *Philosophy of Science*, *35*(2), 116–133.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part ii: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.

Hosseinichimeh, N., Rahmandad, H., Jalali, M. S., & Wittenborn, A. K. (2016). Estimating the parameters of system dynamics models using indirect inference. *System Dynamics Review*, *32*(2), 156–180.

Huber, P. (1981). *Robust statistics*. John wiley and sons.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*(3), 127–133.

Kamphaus, R. W. (2019). *Clinical assessment of child and adolescent intelligence*. Springer.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.

Krueger, R. (1999). The structure of common mental disorders. *Archive of General Psychiatry*, *56*(10), 921–926.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). NY: Routledge.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Lyon, A., & Colyvan, M. (2008). The explanatory power of phase spaces. *Philosophia Mathematica*, *16*(2), 227–243.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*(6), 975–995.

Maier, M., van Dongen, N., & Borsboom, D. (2021). Comparing theories with the ising model of explanatory coherence (imec). *PsyArXiv*.

McNally, R. J. (1990). Psychological approaches to panic disorder: A review. *Psychological Bulletin*, *108*(3), 403.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(1), 806–834.

Meehl, P. E. (2002). Cliometric metatheory: II. criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports*, *91*(2), 339–404.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin and Review*, *26*(5), 1596–1618.

Pargent, F., & Albert-von der Gönna, J. (2018). Predictive modeling with psychological panel data. *Zeitschrift für Psychologie*, *226*(4), 246–258.

Pearl, J. (2009). *Causality* (2nd). Cambridge University Press.

Peirce, C. S. (1931). *The collected papers of Charles Sanders Peirce* (C. Hartshorne & P. Weiss, Eds.; Vol. I–VI). Harvard University Press.

Pincock, C. (2012). *Mathematics and scientific representation*. Oxford University Press USA.

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Reutlinger, A., & Saatsi, J. (Eds.). (2018). *Explanation beyond causation: Philosophical perspectives on non-causal explanations*. Oxford University Press.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743.

Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L., Kossakowski, J., Fried, E. I., Millner, A., McNally, R. J., van Nes, E. H., Scheffer, M., Kendler, K. S., et al. (2019). Advancing the network theory of mental disorders: A computational model of panic disorder. *PsyArXiv*.

Ross, L. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science*, *82*(1), 32–54.

Ross, L., & Woodward, J. (2023). Causal Approaches to Scientific Explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University.

Salmon, W. C. (1971). *Statistical explanation & statistical relevance*. Pittsburgh, PA, USA: University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Saltelli, A., Tarantola, S., & Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling. *Statistical Science*, *15*(4), 377–395.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295.

Schelling, T. C. (2006). *Micromotives and macrobehavior*. WW Norton & Company.

Schoedel, R., Au, Q., Völkel, S. T., Lehmann, F., Becker, D., Bühner, M., Bischl, B., Hussmann, H., & Stachl, C. (2018). Digital footprints of sensation seeking. *Zeitschrift für Psychologie*, *226*(4), 232–245.

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, *78*, 105–127.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. *Computational Social Psychology*, 311–331.

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*(4), 207–218.

Smith, R. C. (2013). *Uncertainty quantification: Theory implementation and applications*. SIAM.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search* (2nd). Springer.

Strevens, M. (2004). The causal and unification approaches to explanation unified–causally. *Noûs*, *38*(1), 154–176.

Strevens, M. (2008). *Depth*. Harvard University Press.

Ströing, P. (2018). Data, evidence, and explanatory power. *Philosophy of Science*, *85*(3), 422–441.

Suppes, P. (1960). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, Methodology and Philosophy of Science. Proceedings of the 1960 International Congress* (pp. 232–239). Stanford University Press.

Thagard, P. (1988). *Computational philosophy of science*. MIT Press.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*(2), 212–233.

van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861.

van Dongen, N., Finnemann, A., de Ron, J., Tiokhin, L., Wang, S., Algermissen, J., Altmann, E. C., Chuang, L.-C., Dumbravă, A., Bahník, Š., et al. (2022). Many modelers.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697.

Vohs, K. D., & Baumeister, R. F. (2004). Self-control. In *Encyclopedia of applied psychology* (pp. 369–373). Elsevier.

Vohs, K. D., & Baumeister, R. F. (2016). *Handbook of self-regulation: Research, theory, and applications*. Guilford Publications.

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L., Christensen, W. J., Clay, S. L., Curtis, J., ... Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, *32*(10), 1566–1581.

Walters, W. P. (2020). Code sharing in the open science era. *Journal of Chemical Information and Modeling*, *60*(10), 4417–4420
doi: 10.1021/acs.jcim.0c01000.

Weisberg, M. (2007). Who is a Modeler? *British Journal for the Philosophy of Science*, *58*, 207–233.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford university press.

Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part i: A counterfactual account. *Noûs*, *37*(1), 1–24.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*(2), 201–219.