



The Theory Crisis in Psychology: How to Move Forward

Markus I. Eronen¹  and Laura F. Bringmann² 

¹Department of Theoretical Philosophy, and ²Department of Psychometrics and Statistics, University of Groningen

Abstract

Meehl argued in 1978 that theories in psychology come and go, with little cumulative progress. We believe that this assessment still holds, as also evidenced by increasingly common claims that psychology is facing a “theory crisis” and that psychologists should invest more in theory building. In this article, we argue that the root cause of the theory crisis is that developing good psychological theories is extremely difficult and that understanding the reasons why it is so difficult is crucial for moving forward in the theory crisis. We discuss three key reasons based on philosophy of science for why developing good psychological theories is so hard: the relative lack of robust phenomena that impose constraints on possible theories, problems of validity of psychological constructs, and obstacles to discovering causal relationships between psychological variables. We conclude with recommendations on how to move past the theory crisis.

Keywords

theory, phenomena, robustness, validity, causation

In recent years, more and more authors have called attention to the fact that the theoretical foundations of psychology are shaky (e.g., Fiedler, 2017; Gigerenzer, 2010; Klein, 2014; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Reber, 2016; Robinaugh et al., 2020; van Rooij, 2019). The claim is that psychological theories are in general of poor quality and that the focus in psychology should shift more toward developing better theories instead of (just) improving statistical techniques and practices and performing more replication studies. In other words, we are facing a “theory crisis” that is more fundamental than the replication crisis that has received far more attention (Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Reber, 2016).

This point is of course not new but notably was also emphasized by Paul Meehl throughout his career (e.g., Meehl, 1967, 1978, 1990). Meehl pointed out that psychological scientists are fond of developing new theories, but instead of resulting in cumulative theoretical progress, these theories tend to just come and go: Theories are neither decisively refuted nor accepted as part of established knowledge; they simply hang around until they are abandoned or forgotten. He mentions as examples theories of “level of aspiration” and “risky shift,” which were received with much enthusiasm in the 1930s and 1960s, respectively, but are now largely forgotten.

In the 40 years that have passed since Meehl’s (1978) classic article, the role of theories in psychology has not changed much. For example, the book *ABC of Behavior Change Theories* lists 83 theories in the field of behavior change alone, ranging from self-regulation and self-efficacy theories to ecological models (Michie et al., 2014).¹ It is safe to assume that none of these theories is universally accepted or decisively refuted. As a more specific example, consider ego-depletion theory (Baumeister et al., 1998, 2000). After a period of great enthusiasm, this theory has been heavily criticized in recent years, and currently there is no conclusive evidence either for or against it (Friese et al., 2019).

An explanation for the lack of theoretical progress in psychology is that psychological theories tend to be formulated so vaguely or abstractly that it is difficult to falsify or test them (Meehl, 1978, 1990). Moreover, even when a theory is found to be deficient and unable to explain some phenomena, psychological scientists often continue to use it, focusing on its past successes (e.g., the Rescorla-Wagner model of classical conditioning;

Corresponding Author:

Markus I. Eronen, Department of Theoretical Philosophy, University of Groningen

E-mail: m.i.eronen@rug.nl

Miller et al., 1995). These factors result in a plethora of coexisting and overlapping psychological theories that are known to be deficient but have not been decisively falsified (Meehl, 1990). Therefore, a common theme in the recent literature on the theory crisis is that psychological theories should be improved by making them more formal and precise or by teaching psychologists how to build better theories (e.g., Gigerenzer, 2010; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; van Rooij & Baggio, 2021).

We find these efforts important and laudable. However, in this article, we take a different approach. We argue that the core of the problem is that developing good psychological theories is extremely difficult and that understanding the reasons why it is so difficult is a crucial first step in making progress in the theory crisis. In other words, the problem is not (just) that psychological scientists do not put enough effort into developing theories or do not know how to build theories but that there are great obstacles to building good psychological theories because of the nature of the subject matter. To explain and analyze these obstacles, we draw from recent philosophy of science.

With this approach, we follow in Meehl's footsteps: In the article that is the focal point of this special issue (Meehl, 1978), he provided a list of difficulties that make human psychology hard to study scientifically. However, Meehl was naturally relying on the philosophy of science of the day, and since then there have been many developments that are highly relevant for the theory crisis, especially in understanding the nature of data, theories, and causality. We draw from these developments in philosophy of science and discuss three key reasons for why developing good psychological theories is so hard: the lack of constraints on theories by robust phenomena, problems of validity of psychological constructs, and obstacles to discovering causal relationships between psychological variables.

Phenomena as Constraints for Theories

In this section, we argue that phenomena constrain theory development in science, but that in psychological science, there is not enough knowledge of robust phenomena to impose sufficient constraints. To start with, in philosophy of science, it is common to distinguish among data, phenomena, and theories (Bogen & Woodward, 1988; Haig, 2013; Woodward, 1989). Data are the raw observations based on experiments or data collection: In the case of psychological science, they can be, for example, responses to questionnaires or observations of behavior. Data serve as evidence for phenomena, which are relatively stable features of the world: For example, the data from different Stroop task experiments provide

evidence for the Stroop effect. If we then want to explain the phenomena, we need theories that describe how they come about.²

This framework is well established and has been applied to psychological science (Borsboom et al., 2021; Haig, 2013). However, the relationships between theories and phenomena are usually discussed only as "one-way traffic": A theory is formulated to explain phenomena, and therefore it should be possible to derive or predict the relevant phenomena from the theory. For example, a central (and, in our view, valid) argument in the theory debate in psychological science is that psychological theories are so vaguely formulated that they do not make precise predictions regarding phenomena (e.g., Oberauer & Lewandowsky, 2019). What has received far less attention in this debate is that this relationship is bidirectional: Phenomena also impose constraints on the possible theories (Bechtel & Richardson, 1993; Craver & Darden, 2013). In other words, a theory has to be consistent with all the relevant phenomena of the field, which narrows the space of possible theories.

Let us illustrate this with an example. Before introducing the theory of evolution, Charles Darwin had gathered an immense amount of descriptive evidence (Browne, 2006; Darwin, 1859; Rozin, 2001). During his famous voyage on the H.M.S. *Beagle* (lasting nearly 5 years), he made numerous observations and wrote them down in his notebooks, which in the framework described above correspond to data. From these data he derived interesting patterns, such as the distribution of different but very similar bird species on the islands of the Galapagos. Over the years after his return, Darwin intensively studied a broad range of topics, including selective breeding, the fossil record, and the samples he had collected during the voyage. In all of these areas, he found phenomena suggesting that species have common ancestors and are selected by nature in a manner analogous to selective breeding. He wrote the *Origin of Species*, a large part of which consists of detailed descriptions of the various lines of evidence, on the basis of these findings (Browne, 2006; Darwin, 1859).

Importantly, this evidence was not only diverse but also highly robust: The phenomena were verifiable and detectable in several independent ways and not dependent on a specific theoretical framework or observation method (Eronen, 2015, 2019; Kuorikoski & Marchionni, 2016; Munafò & Smith, 2018; Wimsatt, 2007). For example, the patterns of the evolution of traits could be observed in the selective breeding of pigeons, cattle, and dogs, and any other researcher could in principle confirm these patterns. These phenomena were therefore generally agreed on in the scientific community and

imposed very strong constraints on the space of possible theories. A theory of evolution had to fit with not just one or two of these robust patterns but with all of them.

The history of astronomy provides an even more striking example of the constraints that phenomena impose on theories. In this case, the relevant phenomena were the patterns in the movement of celestial objects (most importantly the moon and the planets). These patterns were based on centuries of observations and highly robust; the problem was coming up with a theory that satisfied the stringent constraint imposed by the phenomena (Hoskin, 1997). Ptolemy's geocentric model, according to which planets followed complex epicycle-based trajectories, survived for centuries partly because it was extremely difficult to come up with a theory that would have fit the phenomena better or equally well (Hoskin, 1997). Thus, when Copernicus and Galileo developed their heliocentric theories, the space of possible theories was very strongly constrained by the phenomena. The constraints on contemporary theoretical physics are even more extreme: There is a vast body of robust and undisputed patterns ranging from particle physics to astronomy, and any new theory of physics needs to be consistent with all of these patterns.

The situation in psychological science is very different. To see this, let us recall the distinction between data and phenomena. In psychological science, there is an increasing amount of data available from questionnaires, wearable devices, Internet behavior, and so on. However, these data are often of questionable quality (see the next section), and many areas of psychological science still have no large body of robust phenomena comparable to that of biology or physics.

As an example, consider the ego-depletion effect (Baumeister et al., 1998, 2000): the phenomenon that people perform worse on a task requiring self-control (e.g., solving a difficult puzzle) after having previously engaged in a task requiring self-control (e.g., resisting the temptation to eat cookies). The original and highly influential theory explaining this phenomenon is the strength (or muscle, or resource) model of self-control, according to which self-control is a limited and domain-general resource that is used by any tasks that require self-control and can be depleted (Baumeister et al., 1998, 2000).

Hundreds of studies that seem to support this theory have been published (Inzlicht & Friese, 2019). However, in recent years, both the ego-depletion effect itself and the theory behind it have been called into question (Friese et al., 2019). In a multilab preregistered replication study (Hagger et al., 2016), little evidence for ego depletion was found: The overall effect size was small ($d = 0.04$), and for most of the participating laboratories,

the 95% confidence intervals of the effect size included zero. The authors concluded that "if there is any effect, it is close to zero" (p. 558). Moreover, it has been pointed out that even if the effect is real, the available evidence is compatible with other theories in addition to the strength model of self-control (Inzlicht & Friese, 2019). For example, in the process model proposed by Inzlicht and Schmeichel (2012), the ego-depletion effect is explained by reduced motivation and shifts in attention instead of a generic resource that is depleted.

Importantly, this is not an isolated example. The numerous replication failures of findings in psychology, even phenomena that were thought to be well established (e.g., stereotype threat, neonatal imitation, various priming effects; Bird, 2018), suggest that the situation is similar in other areas of psychology (Inzlicht & Friese, 2019). In other words, in many areas of psychology, there is no broad range of robust phenomena that would impose strong constraints on theories. This means that the possible theories are *underdetermined* by evidence: The available evidence (i.e., the relevant phenomena) is not sufficient to determine which theory we should believe to be true (Stanford, 2017).³ In this light, it is not surprising that little theoretical progress has been made in areas of psychology in which relatively few robust phenomena have been established.

Psychological Constructs and Epistemic Iteration

Another important factor explaining why there are so few good theories in psychology is the lack of attention on improving and validating psychological constructs. In the psychological literature, we find a large and increasing number of psychological constructs. New constructs and corresponding scales are constantly introduced, new terms are invented for what seem to be old constructs, the same term is used for apparently different constructs, and so on (Hagger, 2014). For example, in her review of constructs in the psychological literature on control, Ellen Skinner (1996) found more than 30 constructs related to perceived control alone, and since then many more have been introduced (Hagger, 2014).

In principle, to be acceptable scientific constructs, all of these psychological constructs should have *construct validity*. The notion of construct validity was introduced by Cronbach and Meehl (1955), and its meaning has greatly evolved and ramified in the decades that followed (Newton & Shaw, 2013). Some of the core ideas are that the construct should be embedded in a theoretical framework (or a "nomological network" as originally phrased by Cronbach & Meehl, 1955) and that measurements of the construct should be valid in the sense that

they measure what they are intended to measure (Borsboom et al., 2004).

The problem is that although it is widely agreed that construct validity is crucially important, in practice psychological scientists give it very little attention compared with measures such as reliability. For example, Flake et al. (2017) reviewed a random sample of articles published in *Journal of Personality and Social Psychology* and found that most of the articles reviewed reported no validity evidence whatsoever for the constructs used. When evidence was reported, it typically consisted only of a citation to another article. Likewise, the articles collected in Zumbo and Chan (2014) show that psychological scientists tend to report relatively little validity evidence and focus much more on other psychometric properties, most importantly reliability. The simplest explanation for this is that providing reliability evidence is relatively easy, whereas providing validity evidence is very hard. For the former, there are well-established and quantified measures, such as Cronbach's α . For the latter, there is no simple quantitative measure, and there is not even agreement on what construct validity is or what validity evidence should amount to (Newton & Shaw, 2013). If construct validity is understood in terms of the phrase "the test should measure what it is intended to measure," which often appears in textbooks and guidelines, then establishing validity requires showing that variation in the attribute of interest is actually causing the variation in the test scores (Bringmann & Eronen, 2016; Borsboom et al., 2004). As construct validation of this kind is hardly ever done, the result is that psychological science is permeated by numerous psychological constructs of unknown validity (Flake et al., 2017; Fried & Flake, 2018).

Ego-depletion research is a prime example of this. As Lurquin and Miyake (2017) point out, the key concept "self-control" has never been clearly defined or operationalized. It is often used very broadly to refer to any kind of (inhibitory) control over thoughts, emotions, or actions without further specifying the nature of this control (Lurquin & Miyake, 2017). Moreover, the setups that are used to measure or manipulate self-control in ego-depletion studies have never been validated (Inzlicht & Friese, 2019). In a recent study, Wimmer et al. (2019) systematically tested one of the most widely used tasks to induce ego depletion, the letter-cancellation task, in which participants have to cross off letters following complex rules. They did not find any evidence that this task would affect self-control or inhibitory control (Wimmer et al. 2019).

As an example from clinical psychology, consider major depressive disorder (MDD). The definition of MDD stems from the 1970s and has not essentially

changed since then, although it is increasingly clear that the validity of the construct is problematic (De Jonge et al., 2015; Fried, 2017). For example, because there is great heterogeneity in different cases of MDD (e.g., two individuals can have MDD without sharing a single symptom), it is doubtful that MDD in itself is a well-defined category (Fried, 2017). In addition, the numerous scales that are used to measure MDD often have little content overlap, making it unclear whether they are really measuring one and the same construct (Fried, 2017; Fried & Flake, 2018).

It is illuminating to contrast these examples with the natural sciences. Concepts or classifications in the natural sciences are constantly refined through further experiments and observations and by improving the theoretical framework in which they are embedded. A concept that is initially rough and poorly defined (e.g., the commonsense notion "fish") is refined and reconceptualized (e.g., into the concept "Pisces" in the traditional Linnaean taxonomy of species, defined roughly as finned animals perpetually living in water), and then the new version is again tested and adjusted on the basis of new theories and evidence (e.g., "Pisces" is no longer considered a scientific category but has been divided into several distinct classes on the basis of evolutionary relationships).

Examples of this are abundant in the sciences: For example, the concept "electron" was introduced to physics in the 1890s, and it initially meant an elementary unit of electric charge, but since then its meaning has evolved through experiments and theoretical advances such as the quantum theory, and now "electron" refers to an elementary particle that is a fermion, has a charge of -1 , spin of $1/2$, and so on. Chang (2004, 2016) calls this process "epistemic iteration" and characterizes it as "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (Chang, 2004, p. 224)

In contrast, in psychology, this kind of iteration is not the norm, although official guidelines emphasize the importance of validation and how it should be seen as an ongoing process (Flake et al., 2017). There are, however, some positive exceptions (see also Kendler, 2012). For example, when Ebbinghaus pioneered the scientific study of memory in the 1880s, he was treating "memory" as a monolithic commonsense notion and did not distinguish between different kinds of memory (Tulving, 2007). In subsequent research, especially starting from the 1950s, many different kinds of memory have been introduced, such as nondeclarative and declarative memory, the latter of which can be further divided into episodic and semantic memory (Michaelian

& Sutton, 2017). The different categories and kinds of memory are not fixed but are still refined and debated in light of new evidence and arguments (Tulving, 2007).

One practical reason why psychological constructs are often so resistant to change is “generative entrenchment,” a concept coined and developed by William Wimsatt (1986, 2007). Once a concept has many other concepts, theories, or practices depending on it, it becomes “entrenched” and will be very difficult to change, even if it is known to be deficient or problematic. This is because changing the concept could collapse the structures depending on it, leading either to a disaster or a revolution (Wimsatt, 2007, p. 140). Psychological constructs (especially in clinical psychology) often become deeply entrenched over time, as they have applications not only in other theories and models but also in society at large. For example, constructs such as MDD play an important role in diagnosing patients or in making decisions about health insurance.

However, epistemic iteration and validation of psychological constructs is crucially important for finding a way out of the theory crisis. As we argued in the previous section, the basis for good theories is robust phenomena. Phenomena, in turn, are inferred from data, and if the data are based on constructs and measurements that for the most part have not been well understood or validated, the phenomena that are inferred are unlikely to be robust. In other words, one source for the lack of robust phenomena in psychology is the lack of emphasis on the process of construct validation.

Psychological Theories and the Problem of Finding Causes

The third reason why there are so few good theories in psychology is that finding psychological causes is extremely challenging. It is widely agreed that a key feature of good theories is that they should, in one way or another, track causal relationships (e.g., Craver, 2007; Pearl, 2000; Woodward, 2003). For example, Darwin’s theory of evolution described the causes of evolution (natural selection), and the DNA theory describes the causal mechanism of inheritance. In this light, it is reasonable to require that psychological theories, insofar as they aim to explain how the mind works, should also reflect the causal mechanisms of the mind (Bechtel, 2008; Thomas & Sharp, 2019). In other words, they should capture causal relationships between psychological variables.

The problem, however, is that discovering causal relationships between psychological variables is often extremely difficult or impossible, as extensively argued in Eronen (2020). To explain why, we rely on the framework

of the interventionist theory of causation (Woodward, 2003, 2015; see also Pearl, 2000, 2009), which lays out the conditions for inferring causal relationships in a clear and general way.

The characteristic feature of causal relationships is that (unlike correlations) they are relationships that are exploitable for manipulation and control: Intervening on the cause is a way of bringing about a change in the effect. The interventionist theory takes this as the starting point and defines causation (roughly) as follows: *X* is a cause of *Y* if (and only if) it is possible to intervene on *X* to change *Y* when other variables are held fixed to their values. The intervention should be an unconfounded manipulation of *X* with respect to *Y*: The manipulation of *X* should not change *Y* via any other route that does not go through *X* (for more precise definitions, see Eronen, 2020; Woodward, 2003). It is not always necessary to actually perform an intervention; sometimes it is possible to gain knowledge about the effects of interventions indirectly, for example, on the basis of observational data. The same ideas also appear in different forms in other approaches to causation that are more familiar to psychological scientists, such as Rubin’s causal model (e.g., Rubin, 2005) or Campbell’s causal model (e.g., Shadish et al., 2002).

Randomized controlled trials are usually taken to be the “gold standard” for causal inference and for satisfying the above conditions. For example, in a drug trial, participants are randomly assigned to treatment and control groups, and this randomization generates the effect of “holding fixed” other variables than the cause (the drug) and the effect (recovery). The intervention of administering the drug to participants in the treatment group should be unconfounded: For example, there should not be other ingredients in the pill that would affect recovery through a causal route that goes around the drug itself.

Many psychological experiments involve the manipulation of *nonpsychological* causes, such as drugs, educational materials, or visual and auditory stimuli (Eronen, 2020). In such cases, performing the right kinds of interventions is in principle not more difficult than in other fields. Therefore, the following arguments do not concern the venerable experimental tradition, going all the way back to Wilhelm Wundt, of manipulating external independent variables and tracking their psychological effects. However, if the aim is to develop substantive psychological theories that describe causal mechanisms of the mind, establishing causal relationships between external independent variables and psychological variables is not enough: We also need to learn causal relationships *between* psychological variables. And to do this, we need to learn about the effects of interventions on psychological variables.

The problem with interventions on psychological variables is that they are typically “fat-handed” (Eronen, 2020)⁴: They do not change just the one variable that is targeted but several other variables as well. This is because there is no direct way of manipulating psychological variables such as thoughts or affects (Chiesa, 1992; Hughes et al., 2016). Instead, they have to be manipulated indirectly via verbal instruction or other external stimuli, and such techniques are typically not precise enough to change just one variable. For example, it is (at least currently) impossible to manipulate feelings of loss of control without changing any other psychological states, such as motivation, attention, or feelings of anxiety. Moreover, psychological variables can be measured only indirectly, for example, on the basis of self-reports or behavioral proxies (De Houwer, 2011). This makes it very difficult to verify or check what variables the intervention precisely changed and therefore to what extent it was fat-handed.

This creates a problem for finding psychological causes because when interventions are fat-handed, we cannot assume that they are unconfounded manipulations that license causal inferences. More specifically, we cannot assume that they change putative effect *Y* only via a route that goes through the putative cause *X*. To illustrate this, let us again focus on ego-depletion research. In ego-depletion experiments, self-control is manipulated in very diverse ways (e.g., by letting participants engage in a complex or frustrating task or game or by letting them resist the temptation to eat delicious food; Friese et al., 2019). To warrant the conclusion that self-control is the *cause* of impaired performance in the second task, these interventions should be unconfounded manipulations of self-control with respect to the putative effect (i.e., impaired performance in the second task). In other words, they should change self-control in such a way that *other* possible causes of the effect are not affected (e.g., motivation, attention, feelings of anger). However, given the rather general nature of the interventions and our lack of knowledge of the causal structure of self-control and related constructs (motivation, attention, etc.), we cannot realistically assume that this is the case (Friese et al., 2019). For example, resisting the temptation to eat cookies might also affect motivation or induce feelings of anger and frustration. This means that ego-depletion experiments do not provide sufficient evidence that a diminished self-control resource is the cause for the impaired performance in the second task which is indeed in line with the conclusion reached in recent reviews of the state of the research (Friese et al., 2019; Inzlicht & Friese, 2019).

In sum, interventions on psychological variables are likely to be fat-handed, and such interventions do not

provide a reliable basis for causal inference. The experimental tradition of manipulating external factors and tracking their psychological effects cannot simply be extended to manipulate psychological variables, as interventions on psychological variables are entirely different in kind and far more difficult than interventions on external variables (see also Chiesa, 1992; De Houwer, 2011). Insofar as psychological theories should track causal relationships, this is an important factor in explaining why there are so few good theories in psychology and why they are so difficult to develop.

Discussion

In this article, we have discussed three fundamental difficulties in developing good psychological theories: the lack of (sufficient) robust phenomena, the lack of validity and epistemic iteration for psychological constructs, and the problem of establishing psychological causes. These issues should be addressed and discussed to make progress in resolving the theory crisis. We now outline several recommendations for psychological research on the basis of these issues.

First, our discussion supports the recent calls for more “phenomena detection” or “phenomenon-driven research” in psychology (Borsboom et al., 2021; De Houwer, 2011; Haig, 2013; see also Trafimow & Earp, 2016). By discovering new phenomena and gathering more robust evidence for those already discovered, the space of possible theories will be constrained.

Another important reason to support phenomenon-driven research is that phenomena can also be extremely important for science and society as such (Eronen, 2020): Consider, for example, the broad range of cognitive biases that psychologists have discovered, such as confirmation bias, most of which are very robust phenomena (Gilovich et al., 2002). Various theories have been proposed to explain these phenomena, such as the attribute-substitution theory, according to which people substitute difficult computations with simple heuristics, or the more general dual-system theory (Kahneman & Frederick, 2002). However, these theories are far more controversial than the phenomena themselves. Moreover, knowing that these phenomena exist is extremely important for science and society, even if we do not know the theory or mechanism behind them. The same holds for a broad range of other robust phenomena discovered in psychology, for example, the phenomenon that people tend to prefer familiar stimuli to unfamiliar ones (i.e., the mere-exposure effect; Bornstein, 1989). Simply knowing that these phenomena exist and describing them is useful, even in the absence of an accepted theory that would explain them.

In addition to being discovered and being described, phenomena can also be further analyzed by looking for shared abstract structures in different phenomena (Hughes et al., 2016). For example, at an abstract level, phenomena as different as constantly checking your phone and rewarding the good behavior of children with candy can both be seen as instances of (positive) reinforcement (Hughes et al., 2016). For all of these reasons, phenomena detection should be seen as an important goal in itself and as a central part of psychological research (see also Fiedler, 2017; Haig, 2013; Rozin, 2001).

However, we by no means intend to suggest that theorizing in psychology is hopeless or a waste of resources or that we should return to a kind of behaviorism in which theories about mental processes are rejected as unscientific. The issues we have raised should not be seen as insurmountable obstacles but rather as challenges that need to be met before good psychological theories can be developed in a given domain.

This brings us to our next point: It is doubtful whether making psychological theories more mathematical or formal, which is a common theme in the recent literature (e.g., Borsboom et al., 2021; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; van Rooij & Baggio, 2021), will lead to significant advances in psychology as a science.⁵ None of the problems we have discussed is solved by formalizing psychological theories: There will still be no large body of robust phenomena to constrain the theories, the constructs used do not become more valid, and a formal treatment alone does not solve the problem of causality and fat-handed interventions. Moreover, many successful and extremely important theories in the life sciences are not formalized or mathematical theories (e.g., the fermentation theory or the theory of synaptic transmission; Bechtel & Richardson, 1993; Machamer, Darden & Craver, 2000). As pointed out by Rozin (2001; see also Morey et al., 2018), using complex statistical and computational models does not make psychology more scientific and can be even counterproductive if the conceptual and empirical basis (e.g., robust phenomena) is not yet solid.

Finally, it is hard to overemphasize the importance of having clearly and transparently defined concepts as the basis for theories. Note that this is not the same as formalization of theories: Concepts can be well defined in qualitatively formulated theories as well (e.g., Darwin's theory of evolution), and formal theories can have poorly defined concepts as their elements (e.g., models in memetics that have a clear mathematical structure but for which the central concept "meme" is not well defined; Kronfeldner, 2011). Conceptual clarification and construct validation should be seen as an important and valuable parts of research, and validation should

be taken to be an iterative and ongoing process instead of just a hurdle that needs be crossed. In our view, strengthening the conceptual basis of psychological theories is at least as important as improving statistical techniques and practices in psychological research.

In the long run, this will also help with the problem of causal inference, as having clearly defined and clearly measurable constructs makes it easier to perform targeted interventions and to track their effects. With sufficiently well-defined constructs and valid measurements, it may also be possible to eventually infer causal relationships from purely observational data (for more, see, e.g., Eronen, 2020; Rohrer, 2018). Another possible reaction to the problem of finding psychological causes is to develop noncausal theories, for example, in the form of abstract functional principles extracted from phenomena (De Houwer, 2011; Hughes et al., 2016), although whether noncausal theories can be truly explanatory is a matter of ongoing debate (see, e.g., Reutlinger & Saatsi, 2018).

Fortunately, there are ongoing research programs in psychology that exemplify the good practices we have describe above. For example, after the recent disappointments in ego-depletion research, there are now increasing efforts to better define the key constructs, such as self-control and related concepts, and to validate different ways of measuring them (Frieze et al., 2019; Inzlicht & Frieze, 2019; Lurquin & Miyake, 2017). A broader example is the functional-cognitive paradigm (De Houwer, 2011; Hughes et al., 2016) that aims at first establishing environment-behavior relations (robust phenomena) and then formulating explanations for them in terms of clearly defined mental constructs that act as mediators. Finally, as a more concrete example, Robinaugh et al. (2020) propose a theory for panic disorder that is tailored to this specific disorder and thereby constrained by phenomena (there is robust evidence for many central phenomena related to panic attacks), and the authors also explicitly focus on defining the key concepts.

To conclude, we believe that the most fundamental factor underlying the theory crisis is that the subject matter itself, psychology, makes it very hard to develop good theories (Meehl, 1978). Drawing on contemporary philosophy of science, we have discussed three central challenges to developing psychological theories: There are often not enough robust phenomena to constrain theories, not enough attention is paid to defining and validating constructs, and establishing psychological causes is very hard. We hope that this article brings more attention to these crucial issues and thereby helps to provide more solid building blocks for the theoretical foundations of psychology.

Transparency

Action Editors: Travis Proulx and Richard Morey

Advisory Editor: Richard Lucas


Editor: Laura A. King

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iDs

Markus I. Eronen  <https://orcid.org/0000-0003-2028-3338>

Laura F. Bringmann  <https://orcid.org/0000-0002-8091-9935>

Acknowledgments

We thank Freek Oude Maatman and Sven Ulpts for helpful comments on an earlier draft. We are also very grateful to Travis Proulx and the two reviewers for their extensive and constructive feedback.

Notes

1. This example is borrowed from Lakens (2019).
2. Because there is no consensus on the definition of “theory,” we use the term very broadly in this article to include also models, nonquantitative theories, and descriptions of mechanisms.
3. Meehl (1990) made an analogous point regarding the testability of psychological theories:

There exists an implicit misconception, ubiquitous among students and professors studying soft areas . . . This misconception is that, if a theoretical conjecture is “scientifically meaningful” (not theological or metaphysical or so vague as to cover anything), then it must be possible to test it at the present time. Even a slight familiarity with the history of astronomy, physics, chemistry, medicine, and genetics shows that such a metatheoretical notion is plainly false. . . . The most dramatic example from biological science in recent times, and one of the two or three greatest scientific discoveries ever made, is Crick and Watson’s theory of the DNA. No amount of theoretical ingenuity would have enabled them to do this, let alone test it, until chemical methods were sufficiently precise to be able to show that in any organism the adenine and thymine are always precisely equal in the number of molecules present, as are the guanine and cytosine. (p. 239)

4. The notion of fat-handed interventions was introduced to the philosophy of psychology by Baumgartner and Gebharter (2016) and Romero (2015) as an alternative to Craver’s mutual manipulability criterion for constitutive relevance (Craver, 2007). The kind of fat-handedness that we are discussing in this article is independent from the fat-handedness due to constitution discussed by these authors.
5. Of course, if “formal theories” is understood in a very general sense as theories that are clearly and explicitly formulated and not necessarily quantitative or mathematical in structure, we agree that formal theories are preferable to nonformal ones.

References

- 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, 1252–1265.

- Baumeister, R. F., Muraven, M., & Tice, D. M. (2000). Ego depletion: A resource model of volition, self-regulation, and controlled processing. *Social Cognition*, 18(2), 130–150.
- Baumgartner, M., & Gebharter, A. (2016). Constitutive relevance, mutual manipulability, and fat-handedness. *The British Journal for the Philosophy of Science*, 67, 731–756.
- Bechtel, W. C. (2008). *Mental mechanisms*. Routledge.
- Bechtel, W. C., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton University Press.
- Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*. Advance online publication. <https://doi.org/10.1093/bjps/axy051>
- Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, 97(3), 303–352.
- Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, 106(2), 265–289.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071.
- Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2021). Theory construction methodology: A practical framework for theory formation in psychology. *Perspectives on Psychological Science*, 16(4), 756–766. <https://doi.org/10.1177/1745691620969647>
- Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*, 26(1), 27–43.
- Browne, J. (2006). *Darwin’s origin of species: A biography*. Allen & Unwin.
- Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
- Chang, H. (2016). The rising of chemical natural kinds through epistemic iteration. In C. Kendig (Ed.), *Natural kinds and classification in scientific practice* (pp. 53–66). Routledge.
- Chiesa, M. (1992). Radical behaviorism and scientific frameworks: From mechanistic to relational accounts. *American Psychologist*, 47(11), 1287–1299.
- Craver, C. F. (2007). *Explaining the brain*. Oxford University Press.
- Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Darwin, C. (1859). *On the origin of species by means of natural selection*. John Murray.
- De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, 6(2), 202–209. <https://doi.org/10.1177/1745691611400238>
- De Jonge, P., Wardenaar, K. J., & Wichers, M. (2015). What kind of thing is depression? *Epidemiology and Psychiatric Sciences*, 24(4), 312–314.
- Eronen, M. I. (2015). Robustness and reality. *Synthese*, 192, 3961–3977.

- Eronen, M. I. (2019). Robust realism for the life sciences. *Synthese*, 196, 2341–2354.
- Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59, Article 100785. <https://doi.org/10.1016/j.newideapsych.2020.100785>
- Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 12(1), 46–61. <https://doi.org/10.1177/1745691616654458>
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378.
- Fried, E. I. (2017). Moving forward: How depression heterogeneity hinders progress in treatment and research. *Expert Review of Neurotherapeutics*, 17(5), 423–425. <https://doi.org/10.1080/14737175.2017.1307737>
- Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, 31(3), 29–30. <https://www.psychologicalscience.org/observer/measurement-matters>
- Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*, 23(2), 107–131.
- Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology*, 20(6), 733–743.
- Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.
- Hagger, M. S. (2014). Avoiding the “déjà-variable” phenomenon: Social psychology needs more guides to constructs. *Frontiers in Psychology*, 5, Article 52. <https://doi.org/10.3389/fpsyg.2014.00052>
- 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., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., . . . Zwienerberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573. <https://doi.org/10.1177/1745691616652873>
- Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *The American Journal of Psychology*, 126(2), 135–153.
- Hoskin, M. (1997). Astronomy in antiquity. In M. Hoskin (Ed.), *The Cambridge illustrated history of astronomy* (pp. 22–47). Cambridge University Press.
- Hughes, S., De Houwer, J., & Perugini, M. (2016). The functional-cognitive framework for psychological research: Controversies and resolutions. *International Journal of Psychology*, 51(1), 4–14.
- Inzlicht, M., & Friese, M. (2019). The past, present, and future of ego depletion. *Social Psychology*, 50(5-6), 370–378. <https://doi.org/10.1027/1864-9335/a000398>
- Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, 7(5), 450–463.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (pp. 49–81). Cambridge University Press.
- Kendler, K. S. (2012). Epistemic iteration as a historical model for psychiatric nosology: Promises and limitations. In K. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 305–322). Oxford University Press.
- Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, 24(3), 326–338.
- Kronfeldner, M. (2011). *Darwinian creativity and memetics*. Routledge.
- Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science*, 83, 227–247.
- Lakens, D. (@lakens) (2019, September 20). *The Scheel Theorem: Things get more personal in psych because people have their own theory. Consequence: Books like the ABC . . .* [Tweet]. <https://twitter.com/lakens/status/1174963097158578176>
- Lurquin, J. H., & Miyake, A. (2017). Challenges to ego-depletion research go beyond the replication crisis: A need for tackling the conceptual crisis. *Frontiers in Psychology*, 8, Article 568. <https://doi.org/10.3389/fpsyg.2017.00568>
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, 67(1), 1–25.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 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(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195–244.
- Michaelian, K., & Sutton, J. (2017). Memory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2017 ed.). <https://plato.stanford.edu/archives/sum2017/entries/memory>
- Michie, S. F., West, R., Campbell, R., Brown, J., & Gainforth, H. (2014). *ABC of behaviour change theories*. Silverback Publishing.
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3), 363–386.
- Morey, R., Homer, S., & Proulx, T. (2018). Beyond statistics: Accepting the null hypothesis in mature sciences. *Advances in Methods and Practices in Psychological Science*, 1(2), 245–258.
- Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature*, 553, 399–401.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3, 221–229.
- Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, 18(3), 301–319.

- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596–1618.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146.
- Reber, R. (2016, April 30). The theory crisis in psychology. *Psychology Today*. <https://www.psychologytoday.com/intl/blog/critical-feeling/201604/the-theory-crisis-in-psychology>
- Reutlinger, A., & Saatsi, J. (Eds.). (2018). *Explanation beyond causation*. Oxford University Press.
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Robinaugh, D., Haslbeck, J. M. B., Waldorp, L., Kossakowski, J. J., Fried, E. I., Millner, A., McNally, R. J., van Nes, E. H., Scheffer, M., Kendler, K. S., & Borsboom, D. (2020). *Advancing the network theory of mental disorders: A computational model of panic disorder*. PsyArXiv. <https://doi.org/10.31234/osf.io/km37w>
- Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese*, 192(11), 3731–3755.
- Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, 5(1), 2–14.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322–331.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin.
- Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, 71(3), 549–570.
- Stanford, K. (2017). Underdetermination of scientific theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017 ed.). <https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination>
- Thomas, J. G., & Sharp, P. B. (2019). Mechanistic science: A new approach to comprehensive psychopathology research that relates psychological and biological phenomena. *Clinical Psychological Science*, 7(2), 196–215.
- Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology*, 26(4), 540–548.
- Tulving, E. (2007). Are there 256 different kinds of memory? In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 39–52). Psychology Press.
- van Rooij, I. (2019, January 18). *Psychological science needs theory development before preregistration*. Psychonomic Society. <https://featuredcontent.psychonomic.org/psychological-science-needs-theory-development-before-preregistration/>
- 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. <https://doi.org/10.1177/1745691620970604>
- Wimmer, M. C., Dome, L., Hancock, P. J., & Wennekers, T. (2019). Is the letter cancellation task a suitable index of ego depletion? *Social Psychology*, 50(5-6), 345–354.
- Wimsatt, W. C. (1986). Developmental constraints, generative entrenchment, and the innate-acquired distinction. In W. Bechtel (Ed.), *Integrating scientific disciplines. Science and philosophy* (pp. 185–208). Springer.
- Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.
- Woodward, J. (1989). Data and phenomena. *Synthese*, 79(3), 393–472.
- Woodward, J. (2003). *Making things happen. A theory of causal explanation*. Oxford University Press.
- Woodward, J. (2015). Methodology, ontology, and interventionism. *Synthese*, 192, 3577–3599.
- Zumbo, B. D., & Chan, E. K. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences* (Vol. 54). Springer.