PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 127–143
DOI: 10.31648/przegldpsychologiczny.9679

Two Voices on the Credibility Crisis in Psychology

Arkadiusz Białek[1]
Jagiellonian University, Institute of Psychology
https://orcid.org/0000-0002-9002-4764

Piotr Wolski[2,3]
Jagiellonian University, Institute of Psychology
https://orcid.org/0000-0002-7028-6142

Abstract

While various shortcomings and flaws in the conduct of research and analysis of results in psychology and other social sciences have been recognized for a long time, recent years have witnessed greater prevalence and wider scope of this criticism. There are also more proposals for improvement. In this article, we focus on selected, key areas of the credibility crisis in psychology. Piotr Wolski discusses those related to the improper understanding and application of significance tests, while Arkadiusz Białek characterizes some of the research practices that undermine the credibility of psychological studies and demonstrates how to counteract them. Although the use of good research practices can improve the reproducibility and replicability of research results, the proposed reform should also encompass the way theories are developed. The discussed proposal for theory development in psychology leads to a series of practical steps. Unlike the hypothetico-deductive model, it starts with the identification and description of the phenomenon. The explanation of the phenomenon formulated through abduction is then formalized in mathematical equations or computer simulations and verified. Adhering to good research practices and proper theory development has the potential to provide psychology with more solid foundations and make it a cumulatively evolving science.

Keywords: credibility crisis, statistical inference, *p*-value, significance tests, questionable research practices, theory development

[1] Correspondence address: a.bialek@uj.edu.pl.
[2] Correspondence address: piotr.wolski@uj.edu.pl.
[3] Both authors contributed equally to this manuscript.

The recent years mark a period of turbulent discussions and changes in psychology. Challenges in replicating results (Ioannidis, 2005; Open Science Collaboration, 2015), limitations in their generalizability (Yorkoni, 2022), and deficiencies in psychological theories (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019) have led to questioning the credibility of research and declaring a state of crisis. While many of the identified contemporary problems are not entirely new – similarities between the current crisis and debates in social psychology in the 1960s and 1970s are pointed out by Lakens (2023), and criticism of testing null hypotheses dates back to the 1930s, before it became established in psychology (Cohen, 1994) – the present crisis is distinguished by the widespread awareness of the issues and the presence of corrective procedures and practices. This distinctiveness is partially linked to technological development, such as the emergence of platforms enabling preregistration of studies, data, and analysis codes (e.g., OSF, Zenodo, GitHub) or document formats facilitating result reproducibility (e.g., R Markdown, Quarto).

The limitations of space prevent us from providing a comprehensive discussion here on the manifestations, causes, and proposed solutions to the credibility crisis in psychology (a comprehensive and up-to-date overview can be found in Nosek et al., 2022). Below, we focus only on selected, key areas, in our opinion. Piotr Wolski writes about statistical inference and how its improper understanding and application contribute to a decrease in research credibility. Arkadiusz Białek characterizes questionable practices that lower the credibility of studies and discusses an interesting proposal for the principles of theory development in psychology.

Statistical Inference

Methodology is rarely a favorite subject for psychology students. In their survey, Haller and Kraus (2002) observed problems with the correct interpretation of the typical t-test result among 100% of surveyed students, around 90% of researchers, and approximately 80% of methodology instructors. Although the specific nature of tasks and sample limitations may warrant treating the obtained values somewhat anecdotally, their results align with the commonly observed sentiment among methodology teachers that, when it comes to statistical inference, most students, as well as a significant number of researchers, feel less confident and prefer to rely on ready-made schemes.

The most important interpretative scheme, replicated in numerous textbooks, lectures, countless teaching materials, and online guides, and passed directly among researchers, concerns the interpretation of the test probability value p. Its origin lies in the early works of Ronald Fisher, later developed and modified by Jerzy Neyman in collaboration with Egon Pearson. However, today, it is mainly disseminated through a "game of telephone" of many mutually inspiring and repeating sources. In essence, it can be summarized as follows: if the calculated probability value p in a statistical test is less than the threshold

of 0.05, then the tested effect is considered statistically significant. The assertion of statistical significance implies a kind of validation of the tested effect, confirming that the data justifies the statement – with a sufficiently small risk of error – that the observed effect in the sample is not a random fluctuation but reflects a genuine regularity. The *p*-value represents the conditional probability of observing an effect in the sample as large as or larger than the one observed if this effect were not present in the population, i.e., if the null hypothesis were true. If the null hypothesis can be rejected, the effect can be considered statistically significant. According to Fisher (1971), "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (p. 16). In practice, this interpretative scheme usually reduces to a simple rule: if the effect is statistically significant (i.e., $p \leq 0.05$), the results are reliable, and they can be generalized to the population – with a 5% or smaller, depending on the *p*-value, risk of error. Unfortunately, as shown, among others, by the results of the mentioned survey by Haller and Kraus, this interpretative scheme, especially in its simplified version, often leads users to erroneous conclusions, including overestimating the meaning of statistical significance, making overly broad conclusions about the population, misunderstanding the nature of error, and inadequate assessment of its risk.

A significant problem is the incorrect understanding of the *p*-value. We have become accustomed to treating it as if it were some objective, independent criterion of the reliability (significance) of the result. However, the *p*-value is not a population parameter but a sample statistic – laden with random error, just like the tested result. Simple simulations show that – especially with small samples – repeatedly conducting the same experiment leads to fundamentally different estimates of the effect size and, consequently, significantly different assessments of the *p*-value and decisions about statistical significance (Cumming, 2008; Halsey et al., 2015).

Statistical significance testing is a useful tool for distinguishing real effects from random fluctuations, but it works correctly only when both the probability of Type I error (false positive – declaring random fluctuation as an effect present in the population) and Type II error (missing – recognizing an existing effect in the population as a random fluctuation present only in the sample) are appropriately small. The popular convention ensures that the risk of a false positive is no greater than 5%. However, there is no equally widespread custom protecting us from the error of omission while simultaneously guaranteeing a sample size that limits random error to an acceptable level. Statistical power, or the probability of not committing a Type II error, is notoriously low for typical studies in psychology – less than 50% (Bakker et al., 2012; Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). In everyday life, it would be impossible to function relying on such unreliable ways of assessing the state of affairs. Imagine if, every other time, you looked at the bakery shelves full of bread and left disappointed, thinking it had just run out; or if, every other time you tried to wash your hands, you unnecessarily unpacked new soap, wrongly assuming that the previous one was used up; or if you opened the door every other time someone rang the bell, and even then, half the time, you closed it in the faces of guests right after opening

it, stating that there was no one outside and the bell just played a trick on you… Fortunately, evolution has ensured that in our daily functioning, we "instinctively" rely on sufficiently robust data and create sufficiently reliable models of reality. However, paradoxically, when it comes to statistical inference, our decisions are sometimes less sensible than those in everyday life. Statistical intuitions (including those of professionals, Bakker et al., 2016) are fallible, just as golden rules such as the $N = 30$ boundary supposedly distinguishing large from small samples, or the rule of a minimum of 30 observations per cell in factorial experiments. Therefore, just as we do not estimate p-values by eye, we should not establish the required sample size ad hoc but calculate it – according to the expected effect size and assumed statistical power. Basic calculations are not demanding, and suitable software tools, including free ones, are readily available (e.g., Faul et al., 2007). Research is costly and time-consuming, so the sample size usually results from a compromise between the desire to maximize statistical power and practical considerations. We mentioned that the p-value is sometimes absolutized – often users of statistical tests are not aware of its random error, and even results of studies with low power are considered reliable if they meet the "magical" criterion of statistical significance $p \leq 0.05$. Such thinking may encourage a kind of research "poker" bet: although we have no time or money for large studies, the theory justifying the expectation of the existence of the effect seems convincing, so perhaps luck will be on our side, and we will achieve the necessary p-value ≤ 0.05 already in a small sample. However, the laws of large numbers cannot be deceived – every conclusion based on a study with too small a sample is burdened with a high risk of error, both positive and negative. We can believe – with limited confidence – only in those results that come from studies with a sufficiently large statistical power. If credibility is understood in line with Fisher's original idea (1971), as a justified expectation that subsequent studies of the same effect will repeatedly yield statistically significant results (p. 14), then in psychology, only less than every other published statistically significant result is credible (Boyce et al., 2023; Open Science Collaboration, 2015). We can improve this not very glorious statistic by adopting a more rigorous approach to research planning. This important element should receive greater attention in teaching, reviewers should more commonly expect authors to justify their sample size decisions, and, above all, we should plan our projects to provide reasonably high certainty of detecting the effects we are looking for if they exist.

Another important issue in the interpretation of statistical tests is the automatic acceptance of statistically significant results as practically significant. This overinterpretation is favored (and also reflected) by a significant change that has occurred in methodological language: although Fisher (1971) originally referred the adjective "significant" to the inconsistency of results with the null hypothesis, today we rather refer it to the results themselves. We suggest misleadingly that it is not their inconsistency with the null hypothesis that is significant, but the results themselves. Fisher also warned against confusing statistical significance with practical significance. In most typical cases, a statistically significant effect is one for which we only have a reasonable certainty that it is not equal to 0. However, this does not mean that it has practical significance. There are many

effects that, although different from 0, are so small that they are not worth bothering with. Therefore, methodologists recommend calculating confidence intervals and complementing significance tests with effect size estimates (Wilkinson et al., 1999).

From a purely linguistic point of view, it might seem that when we talk about a statistically significant effect, we mean one that is not only significant but also finds confirmation of its importance in statistical procedures. However, in reality, we are talking about effects that are only inconsistent with the null hypothesis – typically, different from 0, but it is not known whether meaningfully so. Paradoxically, therefore, the term "statistically significant" means less, not more than the adjective "significant" in its basic sense. This leads to confusion because, whether we want it or not, we process the meanings of words automatically, and it is challenging to ignore the associations that such a phrase invokes. This is probably why statistical significance is so often overinterpreted and overvalued. It would probably be different if, instead of the term "significant," we used the term "nonzero." The connotations of these terms are fundamentally different– although a nonzero effect exists, its size is unknown, it may be important, or it may not be, and it is challenging to say anything about it without further, more detailed research or analysis, which we are encouraged to conduct. On the other hand, a significant effect is rather an important, large effect, deserving attention, found with sufficient certainty, leaning more towards considering the issue resolved and closing the study. It would be challenging to change the linguistic tradition established over the decades, but it is worth taking educational actions to popularize the correct understanding of statistical significance.

Novice researchers sometimes tend to think that the confirmation of statistical significance validates the observed effect and justifies the belief that its size in the population is close to the value in the sample. They can be blamed for not paying attention in statistics class, but it must also be admitted that such a wishful interpretation of the result corresponds to a psychologically valid expectation: we would like to have a statistical validation tool that determines the level of risk that the population effect deviates from the observed value by more than a reasonable margin of error. Perhaps that is why some users of statistical tests forget that the probability concerns the sample they are studying and not the population, which they – apparently – have no reason to ask about since they have just examined it. However, within classical frequentist statistics, it is impossible to determine the probability of the truth of hypotheses about the population. Therefore, the statistical significance must be decided based on the probability of observing certain data under the null hypothesis being true. This counterintuitive solution encourages misunderstandings and a wishful interpretation of significance. The latter leads to unjustified conclusions and reduces the effectiveness of appeals for broader use of effect size indicators. Someone who falsely understands statistical significance as confirmation that the population effect does not significantly depart from the observed one does not need to additionally check the size of the population effect because they wrongly think they already know it.

Some of the problems with the interpretation of significance tests can be attributed to users who misunderstand them. However, there are also issues

arising from the limitations of the method itself, or more precisely, its controversial logic (Westover et al., 2011). The probability that interests the user can be denoted as P(H | D). This is the conditional probability of the truth of hypothesis H about the population, in light of the observed data D in the sample. Since it concerns subjective certainty about a certain state of affairs, rather than the expected frequency of this state occurring in a long series of repetitions, it makes no sense in the traditional frequentist statistical categories. It can be meaningfully defined only within Bayesian statistics, popular today but considered erroneous by Fisher. Therefore, in significance testing, instead of P(H | D), we calculate P(D | H), the probability of observing data D in the sample, assuming the truth of hypothesis H. The logic of significance testing rests on the assumption that if the probability P(D | H), i.e., *p*, is small, then the null hypothesis is also unlikely, meaning P(H | D) is small. This assumption generally works well, but it must be remembered that the probabilities P(H | D) and P(D | H), although related, are not identical. In an extreme case of a large difference between P(H | D) and P(D | H) the significance test can be unacceptably liberal. Extra caution should be exercised with the results of significance tests when the null hypothesis is highly probable *a priori* (Wolski, 2016).

Some critics propose replacing significance tests with confidence intervals (Cumming, 2014), while others advocate for Bayesian methods (Wagenmakers et al., 2018; Westover et al., 2011). Some radicals even go so far as to ban statistical inference altogether (Woolston, 2015). However, the moderate option, which might not resonate in media so much but is more convincing for many, calls for more balanced corrective actions – better understanding of traditional methods, complementing significance tests with effect size estimation, more careful experimental design, and greater emphasis on replicability (Wasserstein, 2015; Wilkinson et al., 1999). Regardless of the standpoint, everyone agrees that the scientists should make their research and interpretative decisions independently, rather than delegate them to some ritualized scheme.

Questionable Research Practices

Above, we discussed the issues associated with statistical inference and how to correctly apply and interpret the results of hypothesis testing. Simultaneously, many problems in contemporary psychology arise from the application of hypothesis testing methods outside of their intended context, meaning beyond confirmatory research. The same researchers who convincingly demonstrated the inability to replicate a large portion of psychological research results (Open Science Collaboration, 2015) also emphasize the necessity of clearly separating hypothesis generation from testing, the context of discovery from the context of justification, exploratory research from confirmatory research, and prediction from postdiction (Nosek et al., 2018). Blending these contexts and presenting exploratory research as confirmatory or postdiction as prediction may indeed enhance the attractiveness of results and consequently contribute to their

publication. However, it simultaneously raises the risk of committing a Type I error and is one of the causes of the non-replicability of research results in psychology. Searching for any statistically significant relationship in the collected dataset is a form of result manipulation known as *p-hacking*. Moreover, presenting *post hoc* hypotheses in a research report as if they were *a priori* hypotheses is referred to as HARKing (Hypothesizing After the Results are Known; Kerr, 1998). Because *p-hacking*, formulating hypotheses after learning the results are considered the most harmful among questionable research practices (QRPs[4]), it is worthwhile to discuss them more extensively.

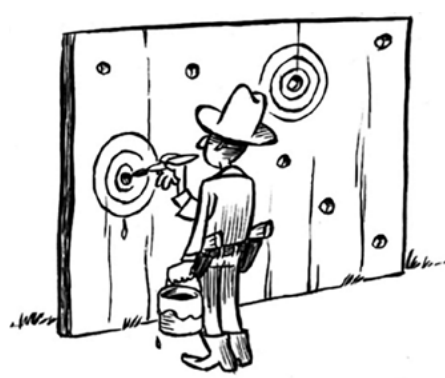
De Groot (1956) was the first to draw attention to the issue of mixing hypothesis generation with hypothesis testing. However, due to being published in Dutch, his work remained unnoticed for a long time (his publication was translated by Wagenmakers et al. in 2014). De Groot argued that while there is, of course, a need for exploratory research in psychology and broader scientific contexts, using the same dataset for both hypothesis generation and testing is an improper research practice. It leads to an increased likelihood of committing a Type I error. He strongly emphasized that the significance level alpha, used in hypothesis testing procedures, refers to an individual hypothesis. By adopting this conventional threshold – in psychology commonly set at 0.05 – researchers allow for the possibility of making a mistake 1 in 20 times. That is, when testing a null hypothesis (e.g., about no association between social anxiety and academic achievement), they may incorrectly reject it and conclude that such an association exists, even though it does not. In psychological research, such a risk of making a Type I error is accepted. However, in medical research, it might be considered too high. Therefore, the significance threshold there is set at 0.01 or 0.005 (allowing for the possibility of making a mistake 1 in 100 or 1 in 200 times). The situation changes, however, when, after collecting data, we start searching for statistically significant results, following the principle of "data torture": continuously examining the data until we find any statistically significant relationship, sometimes even any at all. In such a situation, the probability of making a Type I error and erroneously rejecting the true null hypothesis significantly increases. Let's discuss this process using an example from Dorothy Bishop (2019; 2021). Let's assume after her that we have a large dataset, and we are searching for a relationship between handedness and ADHD. The analysis shows that the relationship between these variables is not statistically significant. Undeterred, we decide to split the collected data by the age of the participants and look for this relationship in younger and older children. Even after such division, none of the relationships prove to be statistically significant. In the next step, because we measured both skills and preferences in the study, we search for a relationship between handedness and ADHD in these subdivided subgroups and for different measurement methods. Still, nothing! Let's pause in our journey through the "garden of forking paths" (Gelman & Loken, 2014) and consider whether the probability of making

[4] A discussion of other questionable research practices can be found in the publication of Andrade (2021).

a Type I error is still 0.05. Unfortunately, the answer is "no." At this stage of "data torture," when we are looking for relationships in four subdivided subgroups and would be satisfied with finding any statistically significant relationship in any of them, the probability is 0.19, following the formula $(1-(1-0.05)^4)$. However, we continue our search and decide to split the existing sets by gender and focus on the groups of girls and boys. Still, no statistically significant result. However, we realized that our participants differed in terms of their place of residence, so we divide them into a group from urban areas and a group from rural areas. And there it is! We found a statistically significant relationship in the group of younger girls living in cities, considering the measurement of skills (not preferences), between handedness and ADHD. As Bishop notes, now all that remains is to come up with a justification for this relationship and describe it convincingly in the article. If, additionally, we present it as a hypothesis and state that in the planning stage of the study, we expected this relationship to occur only in the group of younger girls from large cities, etc., it would be HARKing. Of course, presenting results in this way is not merely telling an untrue story. The more significant issue is that during our "data torture" and search for the "significant *p*" (*p-hacking*) in sixteen emerging groups, the alpha level is no longer the initial 0.05 but 0.56 $((1-(1-0.05)^{16}))$. Thus, we allow for the possibility of making a mistake not once in every 20 studies but in every 2. Therefore, the probability of "discovering" a nonexistent relationship is very high, similar to flipping a coin to determine that a given result is statistically significant, even though it does not exist. This exaggerated example illustrates the real problem of generating hypotheses and testing them on the same dataset in an improper manner.

Figure 1

Comic book illustration of HARKing (Dirk-Jan Hoek, CC-BY)[*]



[*] A researcher engaging in HARKing is like a gunslinger drawing the target after taking the shot, not before.

How can one address such improper practices? The first step is to clearly distinguish exploratory research from confirmatory research and to test hypotheses

only in the latter (Wagenmakers et al., 2012). However, researchers, like other individuals, are susceptible to cognitive biases, including the "I have always known this", i.e. hindsight bias. Consequently, researchers may be convinced that they "basically expected this relationship" because "it makes sense." To avoid, and effectively protect against, such biases, it is necessary – and this is the second step in countering the described situation – to implement and adhere to appropriate procedures involving preregistration of studies. In this case, before examining the first participant, we register our hypotheses, method, and data analysis plan on a platform like the Open Science Framework (https://osf.io/). Such private or public registration is timestamped and can serve to demonstrate that our research is confirmatory, and that the choice of analyses was not made after inspecting the collected data. We can go a step further and submit our research plan to a journal that accepts the "registered report"[5] publication type. A detailed description of the research question, hypotheses, methods, and analysis plan undergoes anonymous peer review. If the submission is accepted based on the review (or after making any necessary revisions), the research report will be published regardless of whether the hypotheses are confirmed and whether we obtain "statistically significant results." Of course, the research must be conducted following the accepted protocol, and the article undergoes a second round of review, mainly focusing on the results and discussion, without challenging the decisions accepted in the first stage. Such a report may additionally include a clearly delineated *post hoc* analysis, meaning it can encompass both a confirmatory and an exploratory part. Therefore, a registered report is a form of empirical publication in which the validity of research proposals is assessed, rather than the statistical significance of results, countering publication bias. Simultaneously, registered reports counteract *p-hacking* and HARKing and enable a clear separation of planned confirmatory research from exploratory data analysis[6].

The Importance of Theory in Psychology

While the proposed procedures and standards mentioned above have the potential to discourage many questionable research practices (low test power, *p-hacking*, HARKing) and institutional issues (e.g., publishing only statistically significant results) and, consequently, make the results of psychological research reproducible (i.e., obtaining the same results in a reanalysis of specific data) and replicable (i.e., obtaining the same results by conducting a study with a different group of people), they are currently considered insufficient. This is because they

[5] As of early 2024, the *The Review of Psychology* will introduce the option to submit registered reports.

[6] A more comprehensive and practical discussion of good research practices can be found in the materials of the course "Best Practices in Statistical Design and Reporting" (Heyard, 2022).

do not address the more fundamental shortcoming of psychology – the "weakness" and non-cumulative nature of its theories. Adhering to the proposed procedures and standards can correct the "machinery of the hypothetico-deductive method" (Scheel et al., 2021), but it does not address the correctness of the process of developing psychological theories and whether, based on their content, reality can be adequately explained and predicted. It is possible that researchers will successfully replicate the results of studies based on an incorrect theory (Szollosi et al., 2019) or those collected using invalid measurement tools. Realizing these eventualities resulted in the demands for reforming the psychologies are not limited only to the improve of research practices and correctness statistical analyzes and conclusions drawn on their basis, but also encompass the reform of how theories are created in psychology. The efforts made in recent years aim in various directions and involve the need to clarify concepts (Bringmann et al., 2022), methods of measuring psychological variables (Flake & Fried, 2020), or the conceptualization of mental disorders, for example (Fried et al., 2022). Due to the limited scope of this article, it is not possible to discuss all these efforts. Therefore, in the following we will limit ourselves to two issues that we consider particularly important – briefly discussing the significance of descriptive research and delving into a broader reconstruction of proposals regarding how to create theory in psychology.

Although Paul Rozin's (2001) critique is primarily directed at social psychology, it seems that the shortcomings he identified also apply to other areas of "soft" psychology. In publication, initially created in collaboration with Solomon Asch (Asch's illness prevented his full engagement in the text), they pointed out that social psychologists often strive to act as "mature researchers," which, according to them, involves formulating hypotheses and conducting experiments. However, this perception is erroneous because in "mature sciences," such as biology, greater emphasis is placed on identifying phenomena and describing them, with less focus on experiments. A phenomenon is a stable, recurring, and general property of the world (Haig, 2005). According to Rozin, many studies in natural sciences are driven by "informed curiosity," starting with the identification of a phenomenon, its description, and determining the scope of its occurrence. Often, they are not based on theory but arise from the need to capture a phenomenon in the world, precisely describe its regularities, and only then develop a theory to explain it. Many breakthroughs in science, such as Darwin's theory of evolution or Watson and Crick's discovery of DNA, emerged in the manner described above – research was not guided by a hypothesis or model but by curiosity and had a descriptive nature. Rozin argues that, in the justified need to become a more advanced science, social psychology skips a crucial stage of describing the studied phenomenon. He advocates taking a "step back" and returning to the observation and description of social behaviors. Methodologists aiming to reform the ways theories are created in psychology share a similar starting point, and their proposal is discussed below.

Almost half a century ago, Paul E. Meehl (1978) in his insightful and constructive critique of the scientific nature of psychology acknowledged that "Theories in 'soft' areas of psychology lack the cumulative character of scientific

knowledge. They tend neither to be refuted nor corroborated, but instead merely fade away as people lose interest" (p. 806). This suboptimal situation becomes even worse when we realize that in many areas of psychology, theories are either not formulated at all or expressed only verbally, often imprecisely (Robinaugh et al., 2021). Psychological theories are sometimes so vague that they cannot be considered incorrect (Scheel, 2022). Recognizing that the main reason for this state of affairs is the "methodological repertoire" possessed by most psychologists includes only designing studies to empirically test hypotheses (most often within the framework of frequency statistics and null hypothesis significance testing), but does not include the correct theory creation and striving to change this state of affairs, the procedure for building theories in psychology developed by Borsboom and colleagues (Borsboom et al., 2021; Haslbeck et al., 2022; Van Dongen et al., 2022) is reconstructed below.

Borsboom and collaborators, in their efforts to improve the theory-building process in psychology, propose a sequence of practical steps helpful in constructing theories properly. An explanation of these steps should be preceded by clarification of the terms they use. Theories formulated in psychology serve to explain phenomena, which should not be equated with data. Data provide evidence of the existence of a phenomenon but are not identical to it, as they are always particular, i.e., collected in a specific place and time; they are ephemeral and idiosyncratic (Haig, 2005). On the other hand, relations or statistical patterns identified in collected data go beyond the particularity of a specific dataset and should emerge in other datasets as well. These statistical relations – identified in different datasets – represent phenomena. For example, the positive correlation between scores on depression and anxiety scales is an identified statistical relation in various datasets, and it represents a stable and general property of the world, i.e., a phenomenon. Therefore, in the metatheory of Borsboom and colleagues, a phenomenon is identified with an empirical generalization[7].

Due to the inherent imprecision of verbal theories, researchers should strive to create formal theories. Borsboom and colleagues (Borsboom et al., 2021; Haslbeck et al., 2022; Van Dongen et al., 2022) propose dividing the process of constructing theories into steps[8]. In the first step, it is necessary to identify the phenomenon. This can be either an empirical generalization, e.g., some people experience panic attacks and worry they will experience them in the future, or a capacity (van Rooij & Baggio, 2021), e.g., the ability to use pointing gestures, or social behavior or social interaction, e.g., teasing. In the next step, a proto-theory must be formulated.

[7] It is not the only possible approach, as van Rooij and Baggio (2021) consider abilities, such as language acquisition, as the phenomena that psychologists should explain. However, due to the perceived accessibility and potential usefulness for a broader audience, we have limited our discussion to the resolutions adopted within the framework proposed by Borsboom and colleagues.

[8] These steps are not limited to those outlined in the main publication by Borsboom and colleagues (2021) but have been expanded and modified to include resolutions found in other publications by members of their team (Haslbeck et al., 2022; Van Dongen et al., 2022).

It has a verbal character and is formulated through abduction, explaining a given phenomenon. Abduction[9] is one of the modes – alongside induction, Bayesianism, and the hypothetico-deductive method – of formulating explanations in science (Fidler et al., 2018). The use of abduction in science dates back to the works of Charles S. Peirce, who stated that "abduction consists in studying the facts and devising a theory to explain them" (Haig, 2005). Therefore, if a researcher formulates a hypothesis or proto-theory to explain a phenomenon and considers it worth further exploration because it provides a better explanation than alternative hypotheses, they employ abduction. Hence, abduction is often, though not always, identified with inference to the best explanation (Haig, 2005). In the next step of theory creation, the explanation of the phenomenon formulated through abduction (verbal proto-theory) should be formalized and expressed in the form of a formal model. This model can be expressed in at least two ways: in the form of mathematical equations or agent-based simulations (Borsboom et al., 2021). In the first case, using differential equations, we attempt to capture the most important components of the phenomenon and the relationships between them. In the second approach, we specify the properties of agents and the rules governing interactions between them and the environment, and then simulate the process of their development using Agent-based Modeling (ABM, Smaldino et al., 2015). Such a model, which formalize the proto-theory's explanation of the phenomenon, is a "toy model," representing only the most important and selected properties of the phenomenon and parts of the real world that give rise to it (Beer, 2020; Haslbeck et al., 2022). It is a "thinking tool" – allowing us to explore the theoretical consequences of explanations (Borsboom et al., 2021). Such a model is not a data model or a statistical model but is a theoretical model. Unlike verbally formulated explanations, explanations expressed in mathematical equations or programming codes are precise, allowing for a strict deduction of the system's development (Haslbeck et al., 2022) and are transparent, which significantly facilitates communication between scientists and, consequently, cumulative growth of knowledge.

The basic verification of the explanatory value of a proto-theory formalized in model comes from deducing or emerging the studied phenomenon from equations or simulations (Borsboom et al., 2021). However, a more detailed verification of the correctness of the formal model involves deducing or simulating data (theory-implied dataset). Then, these data undergo the same statistical analysis used in the analysis of empirical data. Finally, the results of these two analyses are compared (Haslbeck et al., 2022; Van Dongen et al., 2022). Therefore, in this next step of theory creation, we evaluate to what extent the formal model provides data whose analysis yields results similar (a similar pattern of statistical relationships) to those obtained in the analysis of empirical data. If such similarity is

[9] While this term may sound unfamiliar to some readers, the mode of inference described by it is common. If a physician, based on the clinical presentation of a patient, concludes that they have a throat infection and prescribes appropriate medication, or if someone, seeing a person running, looking around and holding a leash, thinks that their dog has escaped and asks if they need help, both of these inferences have the character of abduction.

achieved, we can consider that the formulated theory explains the phenomenon (Van Dongen et al., 2022). In case of discrepancies between the results of these two analyses, we try to explain – again through abduction – their causes and modify the formal model accordingly. Thus, the process of constructing the theory described here is an iterative process of refining it (Beer, 2020; Haslbeck et al., 2022). If the expected similarity is eventually achieved[10], in the next step, we evaluate the value of the theory based on, for example, properties of a good scientific theory formulated by Kuhn, i.e., its accuracy, consistency, scope, simplicity, and fruitfulness (Borsboom et al., 2021), or based on the explanatory goodness of the theory, i.e., its precision, robustness, and empirical relevance (Van Dongen et al., 2022). Finally, in the final step of theory construction, an assessment of its predictive value is recommended. Using the hypothetico-deductive method, bold predictions are derived from it, exposing the constructed theory to the possibility of rejection (Borsboom et al., 2021; Haslbeck et al., 2022). This is a strictly confirmatory procedure that should include preregistrations and simulations of data and analyses related to specific predictions (Haslbeck et al., 2022; Wagenmakers et al., 2012). If the theory passes this test and can correctly predict empirical data, it can be considered confirmed and helpful not only in explaining but also in predicting and controlling psychological phenomena (Haslbeck et al., 2022).

The theory creation process described above poses significant challenges for researchers. Conducting research in accordance with these guidelines requires more time and effort than in the traditional hypothetico-deductive model. However, it seems that such an approach has the potential to provide psychology with more robust foundations, make its results replicable, and, as Meehl (1978) advocated, turn it into a field of knowledge with a cumulative character.

As stated at the beginning of this article, it is worth emphasizing that many of the shortcomings discussed in psychological research have long been recognized[11]. Nowadays, there is a growing acceptance of the need for more thoughtful planning and conducting of studies, as well as the careful analysis and interpretation of gathered data and results. We hope that the above considerations will also contribute to the intensification of these processes.

References

Andrade, C. (2021). HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices. *The Journal of Clinical Psychiatry*, *82*(1), 20f13804. https://doi.org/10.4088/JCP.20f13804

[10] A shortcoming of the discussed proposal is the fact that its authors do not specify what degree of similarity or divergence is expected and acceptable.

[11] It is worth mentioning the article by J. Cohen (1990). While discussing the properties of hypothesis testing, the issue of multiple comparisons, test power, and effect size, the author also adds that he has been writing about some of these matters since the 1960s.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Beer, R. D. (2020). Lost in words. *Adaptive Behavior*, *28*(1), 19–21. https://doi.org/10.1177/1059712319867907

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435. https://doi.org/10.1038/d41586-019-01307-2

Bishop, D. (2021). UBL & Elsevier seminars on Reproducible Research. YouTube https://www.youtube.com/watch?v=C-rk22as870&t=214s

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

Boyce, V., Mathur, M. B., & Frank, M. C. (2023, July 31). Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. https://doi.org/10.31234/osf.io/dpyn6

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to Basics: The Importance of Conceptual Clarification in Psychological Science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal Social Psychology*, *65*, 145–153.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2008). Replication and *p* Intervals: *p* Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, *25*(1), 7–29.

de Groot, A. D. (1956/2014). The meaning of "significance" for different types of research (translation and annotated by E.-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, & H. L. J. van der Maas), *Acta Psychologica*, *148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical

sciences. *Behavior Research Methods*, *39*(2), 175–191. https://link.springer.com/article/10.3758/BF03193146

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. https://doi.org/10.1177/2515245918770407

Fisher, R. A. (1971). *The design of experiments* (9th edition). Hafner Press.

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

Fried, E., Flake, J., & Robinaugh, D. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*, 358–368. https://doi.org/10.1038/s44159-022-00050-2

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465. https://doi.org/10.1511/2014.111.460

Haig, B. D. (2005). An Abductive Theory of Scientific Method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Haller, H., & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online*, *7*(1).

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*(3), 179–185. https://doi.org/10.1038/nmeth.3288

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

Heyard, R. (2022). *Best practices in statistical design and reporting*. University of Zurich. https://osf.io/t9rqm/

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Lakens, D. (2023, July 24). Concerns about Replicability, Theorizing, Applicability, Generalizability, and Methodology across Two Crises in Social Psychology. https://doi.org/10.31234/osf.io/dtvs7

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

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS* (*Proceedings of the National Academy of Sciences of the United States of America*), *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1–8.

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

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years. *Journal of Consulting and Clinical Psychology*, *58*(5), 646.

Rozin, P. (2001). Social Psychology and Science: Some Lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

Scheel, A. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295. https://doi.org/10.1002/icd.2295

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309.

Smaldino, P. E., Calanchini, J., & Pickett, C. L. (2015). Theory development with agent-based models. *Organizational Psychology Review*, *5*(4), 300–317. https://doi.org/10.1177/2041386614546944

Szollosi, A., & Donkin, C. (2021). Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796

van Dongen, N. N. N., van Bork, R., Finnemann, A., van der Maas, H., Robinaugh, D., Haslbeck, J. M. B., … Borsboom, D. (2022, April 13). Productive Explanation: A Framework for Evaluating Explanations in Psychological Science. https://doi.org/10.31234/osf.io/qd69g

van Rooij, I., Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, *31*(4), 321–325. https://doi.org/10.1080/1047840X.2020.1853477

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J. Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N.,

& Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wasserstein, R. (2015). ASA comment on a journal's ban on null hypothesis statistical testing. Retrieved 05 Aug 2015, Sente.

Westover, M. B., Westover, K. D., & Bianchi, M. T. (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine*, *9*, 20. https://doi.org/10.1186/1741-7015-9-20

Wilkinson, L., APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Wolski, P. (2016). Istotność statystyczna II. Pułapki interpretacyjne [Statistical significance II. Interpretive pitfalls]. *Rocznik Kognitywistyczny* [*Yearbook of Cognitive Science*], *9*, 59–70 (in Polish). https://doi.org/10.4467/20843895RK.16.006.6412

Woolston, C. (2015). Psychology journal bans P values. *Nature*, *519*(7541), 9.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, E1. https://doi.org/10.1017/S0140525X20001685

Acknowledgments

Arkadiusz Białek extends gratitude to Róża Krycińska and Monika Szczygieł for their comments on the earlier version of the article.