PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 165–184
DOI: 10.31648/przegldpsychologiczny.9681

A Few Remarks on the State of Research in Social Sciences.

A Conversation with Professor Jarosław Górniak[1]

Jarosław Górniak[2]
Jagiellonian University, Institute of Sociology
https://orcid.org/0000-0001-9210-5712

Arkadiusz Białek[3]
Jagiellonian University, Institute of Psychology
https://orcid.org/0000-0002-9002-4764

Piotr Wolski[4]
Jagiellonian University, Institute of Psychology
https://orcid.org/0000-0002-7028-6142

Piotr Wolski: The high-profile Open Science Collaboration project (2015) highlighted alarmingly low replicability of results published in three prestigious psychological journals. Psychology is currently engaged in a broad debate on

[1] A sociologist and economist, a professor of social sciences specializing in social research methods, statistical data analysis, economic sociology, organization sociology, and public policy analysis, with a particular focus on science and higher education policy, competence development, and the labor market. Vice-Rector for Development at Jagiellonian University, former dean (2012–2020) and vice-dean (2005–2012) of the Faculty of Philosophy at Jagiellonian University, head of the Department of Economic Sociology, Education, and Social Research Methods at the Institute of Sociology at Jagiellonian University. The creator and first director of the Center for Evaluation and Analysis of Public Policy at Jagiellonian University. From 2016 to 2018, he served as the chairman of the National Congress of Science Council. He is the author and co-author of 7 books, 80 articles and chapters, and the scientific editor of 16 collective works. According to Google Scholar, his works have been cited 1675 times, with an h-index of 21. He has promoted 14 PhDs.

[2] Correspondence address: jaroslaw.gorniak@uj.edu.pl.

[3] Correspondence address: a.bialek@uj.edu.pl.

[4] Correspondence address: piotr.wolski@uj.edu.pl.

the reasons behind this troubling situation, actively seeking ways to address and reform standards. The discussion appears to extend beyond psychology, resonating, for example, in the field of medicine. Do you observe similar reformatory movements in sociology or other social sciences that you are familiar with?

Jarosław Górniak: I think that among the social sciences, psychology is most deeply rooted in experimental research. It establishes its empirical results based on studies that can achieve the status of conclusiveness in verifying causal hypotheses. In other social sciences, including sociology, which I represent, the situation is different. That means… of course, experimental research is not entirely absent in sociology, but it must be said that it is rare. It more often occurs in the realm of microsociology, bordering on psychology, which includes the theory of group dynamics in its competence. Microsociology is not a wide stream in sociology, more like a streamlet, but it is present, although one would probably wish for it to be practiced on a larger scale. Unfortunately, at our Institute of Sociology at the Jagiellonian University, it ended with the death of Jacek Szmatka and then the dispersal of his team. If not for that, we could better compare, even on our own turf, how the approaches of our disciplines differ. Nevertheless, it was clearly evident that microsociology, under Szmatka and his team's leadership, was moving towards experimental research due to the subject of research and the scientific style, which was, however, based on the pursuit of verifying causal hypotheses. These hypotheses often resulted from formalized theoretical constructs.

If, however, we were to seek an answer to what is specific to sociology in everyday perception, our discipline is rather associated with observational, descriptive, and demoscopic research. Its characteristic is obtaining answers to questions about the prevalence of certain phenomena, more than establishing strict causal relationships. Although the causal trend is present in both psychology and sociology, pay attention to the characteristic difference between our disciplines. In sociology, at least in this dynamically developing paradigm that seeks to build clear causal models, verification of these models is rather based on results from observational studies. The tool for such verification of causal relationships is structural equation modeling with latent variables. This is what connects psychology with sociology – the presence of hidden constructs in our research. I don't want to delve into deeper debates about the ontological status of these characteristics. Much depends on whether we are realists or antirealists… Realists would speak of hidden characteristics that truly exist and are measured inferentially using a certain set of observed indicators. In contrast, antirealists would argue that these are only theoretical constructs, useful in observation and making judgments about reality… which, in technical terms, ultimately comes down to the same thing. The debate here is more about philosophical foundations.

PW: From a more pragmatic point of view, what's important is that constructs, as you mentioned, are undoubtedly useful. If a construct, such as the general intelligence factor (g), explains a higher percentage of variance in intelligence research than any of the raw measures of cognitive abilities that make up the intelligence quotient, it vividly demonstrates this utility, doesn't it?

JG: Yes, because it gains a kind of predictive validity and is able to predict a broader spectrum of derivatives of what we call intelligence than any single indicator. Even if this characteristic were observable, as long as we're talking about a universe of typical moderately correlated indicators, the main component, being simply a linear combination of these indicators, would explain each of them better than any other single indicator.

The predictive value makes us turn to these types of constructs. As I said, we can set aside the debate about their ontological status. Returning to the specificity of the discipline, sociology, even if methodologically disciplined and striving to verify a relatively strong hierarchy of evidence, theories, or claims about causal relationships – which, to me, is theory in the social sciences: a claim about causal relationships – does this more based on results from observational data. It employs structural models or other types of modeling, allowing valid statements about causal relationships. I mean modeling that tests the causality of interaction based on a counterfactual model. It can be regression with instrumental variables, the difference-in-differences method, or, for example, a type of object matching analysis, such as propensity score matching, or other similar types of matching… Of course, each of these approaches has its merits, each also has some recognized weaknesses. But generally, each of them is better than using naive methods or drawing causal conclusions ad hoc from pure descriptive analyses…

Overall, however, it must be said that in the social sciences, such a more disciplined approach is unfortunately a minority trend. Maybe in economics – of course, the one that seeks econometric verification of its theories – we encounter various cases of modeling, also those that meet the necessary conditions for causal interpretation. Of course, just making a regression model does not provide a basis to claim that even features in strong correlations are causally related. An element of counterfactual analysis is needed for this, such as an experiment allows. Because the experiment is the gold standard here. It is based, one might say, on the principle of seeking counterfactuality through the collision of the test sample with a control sample representing a counterfactual situation, i.e., a situation in which the stimulus does not affect the people under study compared to a situation in which it does.

Seeking the difference between the impact and non-impact of the stimulus on individuals undergoing intervention, or the treatment effect, effectively allows determining the strength of the causal influence. An experiment provides this opportunity. Although it must be said here that it is not surprising that the number of successful replications is limited. There are several different issues contributing to this. It's not just a problem related to improper use of statistical measures, although, of course, that can be the case too. It must be said that experiments are often not prepared with enough care from the perspective of their statistical framework. For example, when planning the conditions of the experiment, the appropriate statistical power is not ensured.

PW: Fully agree.

Should We Use Significance Tests, and How? Replicability of Parameters or Mechanisms?

JG: People often forget that something like statistical power is also an important component in evaluating an experimental situation. We usually rely on significance. We simply want to determine whether an interaction is occurring.

PW: In psychology, perhaps more than in sociology?

JG: In psychology, in psychology. I'll talk about sociology in a moment, but right now, I'm thinking about psychology… Replication issues are one thing there, but another, as discussed, is *p-hacking*, the misuse of significance tests to claim an impact.

PW: And associated with that is publication bias – fixation of authors, editors, and reviewers on the magical $p = 0.05$.

JG: Yes, the thing is that these elements are not planned in conjunction. We know, after all, that adopting a specific critical level of significance does affect the power of the test. Along with the sample size, it influences the overall statistical conditions of the experiment. This should be collectively determined at the planning stage. But what particularly puzzles me is the issue of this vehement criticism of statistical tests and the practice of reporting the test probability, *p*. The exact *p*-value is provided – computers easily calculate it – and it is checked whether it falls below the threshold. This threshold is usually arbitrarily set at the five-hundredths level, and it is expected that the tested relationship should replicate appropriately often, since the probability of committing a Type I error is less than five in a hundred repetitions.

JG: People forget that something like statistical power is also an important component in evaluating an experimental situation. We usually rely on significance. We simply want to determine whether an interaction is occurring.

PW: In psychology, perhaps more than in sociology?

JG: In psychology, in psychology. I'll talk about sociology in a moment, but right now, I'm thinking about psychology… Replication issues are one thing there, but another, as discussed, is *p-hacking*, the misuse of significance tests to claim an interaction.

PW: And associated with that is publication bias – fixation of authors, editors, and reviewers on the magical $p = 0.05$.

JG: Yes, the thing is that these elements are not planned in conjunction. We know, after all, that adopting a specific critical level of significance does affect the power of the test. Along with the sample size, it influences the overall

statistical conditions of the experiment. This should be collectively determined at the planning stage. But what particularly puzzles me is the issue of this vehement criticism of statistical tests and the practice of reporting the test probability, *p*. The exact *p*-value is provided – computers easily calculate it – and it is checked whether it falls below the threshold. This threshold is usually arbitrarily set at the five-hundredths level, and it is expected that the tested relationship should replicate appropriately often, since the probability of committing a Type I error is less than five in a hundred repetitions[5].

What has been proposed as an alternative to significance tests? Editors of *Basic and Applied Social Psychology*, in response to the criticism of statistical test abuse, announced that their journal would not publish articles containing the *p*-value (Trafimow et al., 2015, cited in: Woolston, 2015), which is an extreme and unnecessary reaction. Personally, I lean towards the stance reflected in the title of another publication involved in this debate: “The Practical Alternative to the *p* Value Is the Correctly Used *p* Value” (Lakens, 2021).

Expecting exact replication of parameters in psychological or other social sciences research, which is not strictly neuronal or biochemical (where there is variability but less), is unrealistic. In research areas like social psychology, and even partially cognitive psychology, where we lack robust data akin to hard biochemistry, exact parameter replication is challenging. What we can hope for, and it’s good if we achieve it, is the replication of mechanisms – the structure of the model, pattern, or dependencies. Friedrich Hayek in economics and Ludwig Mises earlier emphasized this. Mises, in his famous work “Human Action”, stated that expecting the same price elasticity of demand for potatoes in the same place after a year is the expectation of a fool. However, this doesn’t undermine the general principle that, under normal conditions, when the price rises, the demand for a good should decrease. Unless we are dealing with a specific situation, as described by Robert Giffen, where due to competition between goods – also considering substitution possibilities and other factors – a peculiar situation occurred in Ireland. There, an increase in the price of potatoes paradoxically led to an increase in demand because people could no longer afford anything else, so they chose the cheapest way of sustenance in the face of famine, namely potatoes. Therefore, even though their price was rising, demand also increased. There are certain deviations that are explainable, but they don’t invalidate the general principle. Nobody, however, expects identical parameters…

PW: Psychology differs from economics in that we are usually not interested in predictions concerning parameters that are crucial in economics. In economics,

[5] Note that we are talking about the conditional probability of making an error of the first kind, i.e. rejecting the null hypothesis when it is true. When performing a statistical significance test, we do not know whether the null hypothesis is true or false, so it would be a mistake to consider that the p-value represents the probability of a wrong decision or a failed replication in this particular case. We mention this because this and similar misinterpretations are very common in the practice of using significance tests (editorial note).

understanding the mechanism is necessary for precise predictions. In psychology, we need replications of studies that serve to identify mechanisms, e.g., studies on dependencies.

JG: I disagree here. In the end, it's about prediction. If the causal theory is correct, in science, it always comes down to prediction. If we adopt a hypothetical-deductive system, confirmed, meaning empirically verified and the structure of the theory is based on causal relationships, such a system inherently allows predictions. Another matter is whether these predictions should be in terms of specific values…

PW: Exactly. If we know, for example, the mechanism of child development, we want to be able to predict whether and when specific difficulties will appear in this development. This level of prediction is present in psychology, but nothing more ambitious… We don't need, for example, to estimate a child's IQ at the age of seventeen based on its IQ at the age of twelve…

JG: I'm thinking more about predictions in terms of patterns. Although economics aspires to predict at a more precise level, and sometimes it succeeds, it's not always the case… Once, my collaborators and I wrote a text analyzing forecasts regarding the labor market made by all Polish institutions dealing with it. The text was written around 2008, analyzing the 90s, almost until the first financial crisis. And in conclusion, we had to state that everyone was wrong in their predictions. Not only in terms of the level but also the direction. And systematically… So, I wouldn't exaggerate with this ability of economics to predict. Various models are created, for example, inflation models, which can occasionally provide accurate forecasts – under conditions of a certain stability in the institutional system, unless additional events modify this mechanism, which is not isolated. However, it's about predicting patterns. The point is that we should know that if we act in way A, in the face of alternative B, we have a chance to achieve a better outcome in a given situation. For example, that the development of a child will be more favorable in scenario A than in scenario B.

PW: Fully agreed.

JG: I wouldn't expect much in the case of sociology or psychology. Exceptions are only phenomena related to processes of a fundamental, biochemical nature, as I mentioned briefly earlier. Although, of course, even in their case, we deal with a kind of random variability. It is always present; it's just less in the case of these processes and phenomena, allowing for more satisfying, more precise results.

Let's return to the reason we're talking about this – as a solution to the problem of limited replicability of studies, it has been proposed, in an extreme variant, to report effect sizes. As if the effect size would save us from something when we don't know if the effect is statistically significant because someone prohibited publishing the result of a statistical test…

The problem is not in that. The issue is that statistical test results are published, applied to results from poorly conducted studies… The test itself is not bad. It can serve very well if the conditions for its legitimate use are met – if the experiment was properly planned and executed. Even then, of course, these experiments need to be replicated. However, journals were rejecting replications, not dealing with them, and yet replications are necessary! Because even if we make a statistical decision error only five times out of a hundred, it is not said when this error will occur…

The probability of a plane crash is very low, and yet many people take a sedative just in case or – using the onboard service – anesthetize themselves differently.

PW: Even though the probability of a plane crash is much lower than the probability of making a mistake in research…

JG: Replications are necessary because you never know when that probability will materialize.

PW: Fisher is blamed for inventing the 0.05 criterion, but he always emphasized the inevitability of the possibility of error and the necessity of replication. He accepted the fact that researchers adopt such a criterion, considering that the value of 0.05 is not a bad idea because it corresponds to a range of roughly two standard deviations, which seems like a reasonable criterion. A super-small criterion would require unrealistically sophisticated studies…

JG: Large samples, simply.

PW: So, on a daily basis, this is a useful criterion, but we must remember that it is relative, and the researcher should decide what tool to choose for which problem, what level of significance to adopt, and then apply it. It's a qualitative decision. In another matter, Fisher is often misunderstood – we often think of statistical significance as the significance of observed results, while he did not talk about significant results but about their significant inconsistency with the null hypothesis. Observing a very low *p*-value strongly contradicts the null hypothesis and allows us to strongly reject it. However, strong rejection of the null hypothesis and acknowledging that the effect is practically significant, or "meaningful," are different things.

JG: It's a problem of incorrect interpretation, a misunderstanding of the term… One can invoke the Sapir-Whorf semantic theory here, which says that language decides how we see the world and generates our actions. Like in that famous example with empty gasoline barrels, where differences in the meaning of the word "empty" in two different languages led users of one language to more often ignore safety regulations and smoke cigarettes near supposedly "empty" barrels containing explosive vapors. Similarly here, we shouldn't use the term statistically significant.

PW: Exactly.

JG: When I discuss this issue with students, I say that the probability of erroneously rejecting the true null hypothesis is less than 0.05 or 0.01, or simply is, for example, 0.003… It is a small probability; therefore, we reject the null hypothesis. Making such a decision, we rarely make a mistake.

We must also remember the statistical preparation of studies, which I mentioned earlier. The level of significance and the power of the test are related. In the planning stage, we can determine what minimum effect size will be significant for us in terms of the weight of the mechanism, its substantive significance. This will allow us to establish appropriate thresholds for the sample size, test power, and level of significance. All these elements are interconnected. There are computer applications designed for this purpose. They should just be used. They are commercial, free, such as R. At our university, we have several different packages available, for example, Stata has it done very well, allowing you to make very nice charts to help make the right decisions regarding the planned studies[6].

Returning to the previous thought, for me, replacing the reporting of *p*-values with reporting effect sizes in publications is not a solution. Because the bare value itself doesn't say much. I also wouldn't expect an exact replication of the observed effect size in the next study!

Arkadiusz Bialek: In developmental psychology, the predictive value is additionally diminished by the cohort effect. We can't expect to observe exactly the same thing in a few years. Similarly, as I understand, it is in economics…

JG: The universality of theories can be verified precisely through the replication of experiments in different population segments. If the hypothesis predicts a cohort effect, then the experiment needs to be replicated in different cohorts.

AB: Exactly. There are research plans that allow separating the cohort effect. However, longitudinal studies, which are most significant in developmental psychology because they allow identifying developmental changes, are not only difficult in themselves but also very difficult to replicate.

JG: These are panel studies, repeated with the same individuals. They allow measuring gross change, i.e., the change that occurs at the individual level between individual measurement points over time. For example, related to the phases of the life cycle. However, the cohort effect is something else. A cohort carries certain socialization processes. They can be studied in independent samples. And independent replications can be done here. Firstly, it's necessary to replicate the study with the same cohort after some time; this can be an independent sample, as long as it's taken from the same cohort. Secondly, we should,

[6] Analysis of the test power and determination of the sample size is made possible by the free software G*Power (https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower).

of course, conduct studies with different cohorts each time, in every wave of measurements. Then, having all these results, it's necessary to separate the cohort effect from the effect of the life cycle phase.

AB: Exactly.

JG: There is still a third effect – the historical period effect, the action of a certain combination of factors, you could say historical, that can affect all cohorts.

AB: Like the pandemic, right?

JG: Exactly. It can affect all cohorts, and the worst part is that when we try to make one model later, these parameters are interrelated, and unfortunately, we don't have enough degrees of freedom to estimate such a model. But this can be bypassed. There are ways that help deal with this problem. This shows how science creates tools to deal with diverse research questions. Research design, i.e., the research plan, is always a derivative of the research question we pose, the problem we want to solve.

I have the impression that both in education and self-education, and then in the work of a researcher, too little importance is attached to proper research planning, research design.

AB: I completely agree.

JG: The problem lies here. Banning the *p*-value won't solve the problem of improper use of tests or the fact that our studies are not being replicated because someone assumed incorrect assumptions or too hastily stated that they are dealing with a statistically significant effect… just because they were happy…

I had a period in my life when I was intensively involved in imparting knowledge about statistical analyses. I conducted various workshops and seminars. Once, after one of these meetings where we talked about exact tests available in the SPSS package, a participant approached me. He worked for a medical university and said he didn't like SPSS; he preferred Statistica because there he gets the results of many different tests in one table and immediately sees which tests to choose to confirm the hypotheses the researcher cared about. Clients are satisfied, and such a summary saves him a lot of work… Of course, if someone conducts research in this way, it creates many problems…

PW: I would like to address the demand you're talking about, replacing the *p*-value with the effect size. Indeed, many methodologists remind us of the need to pay more attention to the effect size. Still, when it comes to replacing *p* with another measure, perhaps the most influential is Geoff Cumming's voice, the author of the "New Statistics" textbook, who advocates replacing p-values with confidence intervals. Mathematically, both are closely related, so the change is – you could say – cosmetic but significant. If a researcher is reasonably familiar with the basics of statistical inference, they believe that if *p* is less than 0.05, then they can assume that the regularity they observed in the sample also occurs in the population.

JG: Well, if he has a confidence interval, he'll reach the same conclusion, seeing that the confidence interval doesn't include 0.

PW: Not really. Cumming bases his postulate and teaching program on the results of studies showing that if researchers are presented with two effects, and for one of them, the *p*-value exceeds the magical significance criterion of 0.05, and for the other it doesn't, they are more likely to consider these effects as qualitatively different than when they are shown two confidence intervals…

JG: For me, that's a false trail. To be clear, I have nothing against confidence intervals. Since it doesn't cost much, I would advise researchers – show both. The approach where something needs to be discarded is erroneous. But the argument of naivety is not a good one. After all, we're not talking here about popular texts where measures need to be carefully chosen so that the reader's intuitive reception aligns with the popularizer's intention, i.e., for the popularizer to achieve the effect of correct reception of scientific content. However, when it comes to communication within science, we demand that the communicating parties possess the appropriate workshop competence. Researchers must be well-prepared.

We, of course, have a problem that not all researchers are well-prepared.

PW: The results of studies on understanding the basics of statistical inference suggest that most of them are not.

JG: Well, that's what reviewers are for. Good journals take care – reviewers should be very well-prepared from a statistical perspective. If they have a problem with that, they can appoint separate methodological-statistical reviewers who will check if everything in the submitted papers is fine in this regard. This needs to be controlled because scientific knowledge is a specific segment of human, social knowledge that is characterized by the fact that it arises in a particular way – based on the scientific method. I won't dwell on the concept of the scientific method because we probably understand each other well here, but for this method, research methodology is crucial. It tells, in line with the best current knowledge, how to conduct studies so that their results are scientifically credible. This knowledge undergoes changes, so sometimes old results may be refuted when new, better methods of studying a phenomenon emerge.

In the natural sciences, we've also had cases where researchers were given better instruments and discovered anomalies inexplicable within older theories, which became unsustainable and had to be discarded.

Therefore, in the field of methodology, each of us must educate ourselves throughout our lives. If someone wants to pursue science, they must continuously develop their toolbox.

PW: That's a great conclusion to this part of our conversation…

The Significance of Exploratory and Confirmatory Research in Theory Building

PW: I would like us to change the topic slightly now and talk about theory building. The pretext for this could be the ongoing discussion about the practice known as HARKing, which stands for Hypothesizing After the Results are Known (Kerr, 1998) – formulating research hypotheses after obtaining the results. This is considered a mistake. So, how should research be conducted to promote the development of good theories?

JG: We need to ask whether it is indeed a mistake. Here, I probably won't follow the mainstream again because I find the stance closer to Charles Peirce more appealing. Peirce drew attention to what happens before the moment when – as Popper wanted – we formulate and test a hypothesis, subjecting it to an attempt at refutation. Popper said: I'm not very interested in where the hypothesis comes from. It's a somewhat nonchalant statement… But that's precisely what interested Peirce. He spoke of the process of abduction, which results from systematic observations, as a natural stage in the process of scientific discovery. He talked about scientific insight, serendipity. A recognizable example of such insight is Archimedes jumping out of the bathtub and shouting: Eureka!

An example closer to life can be what the fictional Dr. House did with his colleagues. They collected certain medical parameters of the patient, and then formulated hypotheses. In their case, these were theories about the nature of the disease because diseases have specific symptoms. There are causal relationships between the observed parameters and the occurrence of a particular condition. They tried to guess what was associated with a specific data configuration, hypothesizing after collecting these parameters. Then, of course, the second stage followed, in the spirit of Popper, where, having a hypothesis about causal relationships, we derive a prediction from it, concerning what should occur if our theory is true. So, we formulate a hypothesis about a causal relationship: if the patient suffers from a particular ailment, we should observe a specific element. Let's conduct an additional study and check. This is something that occurs in that disease and not in others, so it will allow us to determine what we are dealing with. Therefore, we plan a study that either refutes our hypothesis or allows us to maintain it.

We conduct a test of the causal consequences arising from the theory. However, the theory arises from the fact that we constantly operate within a certain realm of knowledge. We are not a scientific *tabula rasa*; we do not clear our minds before formulating hypotheses that we subject to falsification. We are immersed in a certain paradigm, in a set of theories that sometimes compete with each other. We gather observations and say: among the theories known to me, this one fits these data. If it is true, what should follow from it? Or we say: this theory doesn't fit here, so I'll risk creating my own theory. This is, in a sense, *post hoc* hypothesizing, but it is the creation of a theory inspired by the data. Importantly, I don't stop there. My scientific endeavor does not end with deriving concepts from the obtained data but with posing and verifying a hypothesis that leads me to a conclusion. I would call this first phase provisionally exploration and the second phase provisionally confirmation.

I am a strong advocate of exploratory research. I believe that many interesting things have emerged thanks to it. Exploration is quite natural for science. Let's look at how many interesting things have emerged through it, even if they often did not have a conclusion later.

For example, the distinctions of Pierre Bourdieu, a very popular sociologist, an icon of late 20th-century sociology. How did his concept of the class structure of society in connection with lifestyles and tastes, which he called 'tastes,' come about? Bourdieu examined what people do professionally, treating it as a manifestation of their class situation, their position in the social structure. He noted what they dedicate their free time to, analyzing these choices from the perspective of tastes or flavor. Using correspondence analysis, he created a perceptual map, saw how everything fit together, and began to draw conclusions. He built a theoretical concept, but now another stage appears. Predictions must be derived from this concept: if it is true, the appearance of element A should lead to A', and the appearance of B should result in the observation of B'. Again, we are talking about replicable patterns of relationships, not replicable parameter values.

AB: I'm very glad that in our conversation, we are arriving at conclusions that resonate so much with ideas emerging and proposed in the context of reforming psychology. Dutch psychologist and methodologist, Denny Borsboom, believes that the process of scientific knowledge has an iterative character. As you say, we start with descriptive research and the identification of certain patterns of relationships…

JG: Yes, structures. Structures that govern the correlations observed between phenomena.

AB: According to Borsboom (Borsboom et al., 2021), this model should be at least partially formalized and subjected to simulation. So we could talk about adding another stage. What is particularly interesting, however, is that they also refer to Charles Sanders Pierce and explicitly mention abduction as a preliminary stage…

JG: I'm glad I'm not alone. (laughter)

AB: In my opinion, in the case of HARKing, the problem is that the hypothetico-deductive pattern of practicing science is such a widespread norm – sometimes accepted even unconsciously, enforced by the structure of scientific articles – that many researchers feel a very strong need to enter into that confirmatory pattern and formulate hypotheses after obtaining results. Instead of transparently admitting that the hypothesis emerged as a result of observations made, they present it as something they approached the research with…

JG: This is raising the evidential status of their achievement. If you follow the literature on structural modeling, you can see a similar problem there.

I won't go into the details of the debates between supporters of different approaches to model verification, but I'll mention one discussion revolving around the question of the validity of using so-called modification indices to improve structural models. These indices tell us how freeing a particular parameter will affect the overall fit of the model. This allows us to make *post hoc* adjustments that improve fit. In my opinion, such a procedure in itself is not bad. It's a kind of exploration. However, after such corrective action, the modified model should be tested on new data from an independent sample. That's what advocates of this practice recommend. They say, don't be afraid to use modification indices because they help you. Like inspecting residuals between the reconstructed matrix and the observed one…

PW: A bit like in regression…

JG: Exactly – we improve the model.

PW: However, risking overfitting – too good a fit of the model to those specific data.

JG: But all of this fits into the exploratory phase. That's why verification – checking if the model can be sustained – comes later.

There is also a discussion about what criterion should be used to accept or reject the model. Orthodox thinkers believe that the only valid criterion is the chi-square test. Others believe that it is better to use descriptive measures, such as one of the most popular, RMSEA (root mean square error of approximation), or others. Various, more or less arbitrary thresholds are accepted here. Supporters of these methods point out that if chi-square is taken as the criterion, with large samples – common in social research – practically no model would be tenable.

I don't want to settle this now, but importantly, no one – neither camp – questions that after modifications, adjustments to the model, independent confirmation is needed.

AB: But very few actually do it…

JG: Yes, very few actually do it.

Moreover, in experimental research, we have similar problems. I once presented at a psychological conference – based on Kenneth Bollen and Judea Pearl (2013) – the errors we can make in psychological experiments because both independent and dependent variables are latent traits. However, not only the stimulus we directly manipulate but also the measured quantity are just indicators of these latent traits. We assume that when we perform a specific procedure, we cause a real change in the level of the stimulus at the input.

For example, wanting to examine the influence of motivation on the level of attention concentration, we praise the participants and assume that this increases their motivation. Then we measure the rate of errors in a attention-demanding task, treating it as a (negative) indicator of attention concentration.

But how can we be sure that the interaction doesn't actually occur directly between the indicators? That the applied praise doesn't directly reduce the likelihood of errors in the used task, for example, by providing useful feedback to the participant? We assume that by manipulating our observable indicator, we actually make a change in the level of the stimulus at the input, that we increase motivation (latent variable), which improves the effectiveness of attention, resulting in a reduction in the number of errors. But we don't really know.

The proposed solution to this problem in structural modeling was to combine the experiment with modeling, using at least two or more indicators.

PW: In psychology, the problem of internal validity is so widespread that we have even become somewhat accustomed to always risking it. However, we need to do something to minimize it.

JG: In general, you could say that research methodology is the science of avoiding cognitive errors. Starting from the simplest errors of the natural approach, such as unjustified generalization, and ending with more sophisticated issues, such as the discussed problem of not recognizing the significant structure of causal relationships concerning a particular research situation. These things require more attention from researchers.

But to briefly conclude the problems of theory-building…

In my opinion, scientists naturally work exploratively. One of my researcher friends once told me, 'you know, for me, the greatest pleasure is when I approach a problem, search, conduct various exploratory analyses to create a model. But when I have that model and need to deal with its verification, it becomes terribly boring…' There, you have to do concrete, tedious work…

AB: Yes, and here is creation.

JG: Yes, creation, the aforementioned serendipity, enlightenment. And that pleasure you feel when something comes to mind that can bring you closer to solving a problem…

There is another important issue. We can contrast two approaches: scientific and practical. Science, by nature, seeks to formulate causal explanations, to understand the mechanism. Practice, on the other hand, is mainly interested in prediction.

And now, at this moment, we can say that a powerful development of computational and analytical techniques is taking place in the field of prediction. Now, all these big data analyses, what is called data science, is a powerful turn towards induction, consciously foregoing giving this induction a cognitive status, seeking some causal mechanism. It is enough to make an induction of rules that appear with sufficient probability to achieve a specific return on investment. Ultimately, that is essential. If something allows us to predict more effectively by a few percent, with multiple repetitions of a specific effect, we collect the cream from the milk. There would be nothing strange about it if it weren't for the fact that such a procedure brings tremendous success in very different fields, because artificial intelligence, in fact, also relies on the induction of rules.

AB: So, the application of machine learning.

JG: Yes, neural networks, machine learning. We need to learn how to use what the findings obtained with their help bring us. However, if we open ourselves to this fuller cycle, in which we are dealing, de facto, with exploration, then we need to learn to draw conclusions about potential causal models.

PW: Exactly, because those rules based on correlations can only predict, but they cannot provide a causal mechanism.

JG: It's not a matter of whether they can or cannot because they don't test it at all.

AB: They are created for a different purpose.

PW: Well, but this cannot even be done. A neural network, taught a certain response, does not reveal that algorithm, and the programmer cannot extract it.

JG: Yes, but there is also an attempt to automate the generation of causal models, which is related to Bayesian networks and more broadly to trying to identify in a more automated way how to ensure the internal validity of causal analysis by properly closing the 'back doors' through which information flows that could disturb the causal relationship. For example, there is a free tool called DAGitty, thanks to which we can extract from the network of potential dependencies those points that we need to control and plan the study in such a way that we can determine whether factor A really affects factor B in the presence of other factors. Additionally, a characteristic feature of these types of models, graph-based models, is that they are non-parametric models. That is, we are not interested in the values of parameters, but we are interested in whether there is a causal mechanism at all.

AB: However, in a broader perspective, one could say that nowadays, in a sense, we are witnessing the realization of what Karl Pearson wanted, namely the gathering of almost complete data. Fisher introduced statistical inference because he believed that we could not gather complete data, so we must take samples and infer about the population based on them. So, if nowadays, in the era of big data, we can collect complete data, then – perhaps – statistical inference is not needed at all?

JG: No, these are population analyses, but this only removes the issue of inference, which means statistical inference from samples to populations, an important branch of statistics loses some foundation here. However, techniques developed in the field of statistical inference are used in testing the stability of models. It can always be said that even if we study the population, the moment we take a measurement is a random moment from the universe of states in which all objects of a given universe can be found. So, there may be some random variability here. We may also want to draw conclusions about the precision of our expectations about the future.

Let's look at the fashionable big data analysis on data, which often have the nature of non-reactive data, i.e., data generated without the awareness of the subjects that they are under observation. Such data are not subject to disturbance referred to as ecological, which occurs in the case of reactive data. When conducting interviews, we treat subjects in a certain way, so there is always a research effect. If it is a questionnaire survey conducted by interview technique, the consequence is the interviewer effect. In the case of an experiment, it may be the effect of the experimental situation, and so on. We always have to take these disturbances into account. The difference between reactive and non-reactive research can be illustrated with the following example: if a dragonfly does not know that it is being observed, we have a chance to see that it sometimes sits on a stick, not just constantly flying. Because if the dragonfly saw every time that we were observing it, for its own safety, it would take off into flight, and as a result, based on such observations, we might form the false conclusion that it always remains in flight, and we would draw a false conclusion about the non-sitting insect. Non-reactive research is important, but even if we cover the entire population with it, it does not mean that the parameters generated from it, linking certain features, will be repeated in subsequent studies in a perfect way.

Prediction with a satisfying level of precision is very important operationally, but it is absolutely insufficient strategically. Strategy deals with what will happen to the studied community in the future, at a different time, possibly under slightly changed circumstances.

To achieve the ability to predict and determine what will happen when we act with other factors that did not occur when the data were collected, we need to understand how mechanisms work. And this is the place for science, and no one will take this place away from science. There will always be a demand for the practical use of science to make strategically oriented decisions. If such knowledge were not provided, decisions of a strategic nature would probably be made based on intuition. Humans have the ability to create theories that are either correct or not, based on observations, induced rules. People see that certain things are repeated, but only clever people will think about why, and they will notice that sometimes there are minor or major disturbances in them and create concepts of why this happens. People naturally create theoretical concepts, create models. They can make decisions based on a small amount of data. They can create models of dependencies, without empirical data or having very sparse, individual observations. For example, stereotypes that are models of expected dependencies that we build based on individual, sparse, and unscientifically systematic observations, or they are part of cultural transmission and may concern a completely different situation, completely different configurations, and are not suitable for transfer to the present day. Thanks to the fact that people create models, they can make decisions based on very little data.

This economy in terms of the amount of data needed to operate effectively is very attractive in terms of the vision of building artificial intelligence systems, so work on this will not stop.

AB: I would like us to go back a bit to the beginning of our conversation. For the thematic issue, we have planned an article on causal inference and the work of Judea Pearl. You mentioned at the beginning that such thinking is already present in sociology, while in psychology, the belief dominates that only experimental research is the basis for making causal inferences. The use of observational data, even with the application of structural equation modeling, is insufficient.

JG: I wish you good luck in applying experiments in astronomy.

AB: Exactly, at the same time, we will not ask people to start smoking, i.e., we will not conduct experimental manipulations.

PW: Or we will not remove their schizophrenia, or we will not give it in an experiment.

AB: Yes, we will not experimentally give schizophrenia. So how to convince psychologists that, under certain conditions, it is possible to make inferences about causal relationships based on observational data.

JG: Convince them to learn methodology. (laughter)

AB: So, we're going back to education.

JG: I think, observing the activity of the Banach Circle[7], that its members don't need much convincing. I think they are already convinced that not only experiments matter. But you have to be careful because you can easily move towards randomness. There is a lot of poorly conducted research in social sciences. Of course, research is done for various needs, such as describing reality. And if there is such a demand, such a result is provided. However, in that case, it must be as precise as possible and free of various errors. What I have a problem with in such research questions in social sciences, in sociology and related areas, is not adhering to basic principles, such as the representativeness of the sample if you conduct research on samples, and, of course, the principles of measuring with a social survey. Neglecting various factors that disturb the quality of the result, starting from negligence in data collection, that is the measurement situation itself… But above all, randomness concerns the ways of selecting and implementing the sample or, when estimating results, dealing with the situation that we have cases drawn but not realized. Or that we have missing data within realized observations. All these factors affect the results, distort them.

PW: In psychology, there are more characteristics that are more biologically determined, and a non-randomly selected sample also has them randomly distributed…

[7] A discussion circle at the Institute of Psychology of the Jagiellonian University focused on methodological issues and research integrity.

JG: I wouldn't console myself. I understand, this is the species unity of humans, right? (laughter) If we study one person, it's as if we studied everyone. Or, we can study a pigeon, (laughter) it's also a vertebrate, it's also warm-blooded, right?

AB: But a psychology student will definitely be enough. (laughter)

JG: Yes, a psychology student will definitely be enough, but pigeons are also very popular, it even later penetrated into sociology, for example, in the order of pecking.

AB: Going back to the limitations of contemporary research that we were talking about recently with Piotr and Professor Brzeziński, who is also the author of an article in this issue, it is the issue of online research using various platforms designed for this purpose. These limitations consist in the fact that research participants often – colloquially speaking – click through questionnaires, i.e., they quickly provide answers without reading their content or not reading the content at all. Although it can be controlled to some extent…

JG: But in this respect, this tool is no different from filling out a paper survey, it even provides more control. I believe that the threats associated with individual data acquisition techniques are recognized. In the case of online research, the dramatic problem is controlling the sample and the issue of the representativeness of the results. I have the impression that people have stopped caring about it altogether and just gather data. It's like in the early days of survey research when questionnaires were sent out, as in the famous example described in statistical textbooks: *Literary Digest* sent out 4.5 million surveys to its readers who declared whom they would vote for, and based on this, an election forecast was created. Meanwhile, the Gallup Institute conducted a survey on a representative sample of about 1,300 people. And it turned out that this forecast accurately predicted that Roosevelt would win, and *Literary Digest*, based on 4.5 million collected surveys, exaggerated by over 10% in favor of his opponent. Increasing the sample size itself does not reduce systematic error; it only reduces random error in a non-linear way.

PW: In psychology, most studies use small samples, so when there's a study with a large sample, we immediately think it's fantastic, we can believe the results because, after all, it's a large sample. But the larger it is, the more biased it may be, unfortunately. Isn't that right?

JG: It depends, but in short, simply increasing the sample size has no effect on systematic error because contextual factors that cause this error decide about the systematic error. If they are repeated, increasing the sample size doesn't change anything here. With an increase in the sample size, the proportion of systematic error increases, and the proportion of random error decreases in the total study error.

PW: In psychology, the key thing is this immortal *p*-value, and in large samples, we have many effects in which *p* exceeds the accepted threshold.

JG: And that's precisely why your studies don't replicate. Sometimes you would have to assume several replications right away in different contexts. It may turn out that the parameters change in a specific way, which can be the source of a very interesting hypothesis about the significance of a certain contextual factor for the overall study result, for the parameter value. If it turns out that the pattern of the relationship is maintained in replications, and the parameters change and remain in some systematic dependence on a certain feature that differentiates people in a broader community, then this can lead to the construction of a theory of the causal mechanism and its interaction with a specific contextual factor. I'm not saying that respondents have to be randomly selected for psychological studies. They can be purposefully selected. I would even say that they can be selected homogeneously, but then there should be several experimental situations with homogeneous but different groups. Then it becomes possible to examine whether a certain characteristic, due to which we make this purposeful selection, will not be a feature that causes certain variability. The aforementioned Professor Jerzy Brzeziński is a tireless advocate of the external validity of research…

AB: So, we're going back to the problem of generalizability, which is implicitly assumed in psychology, and although we are studying psychology students, we generalize to people in general.

JG: Or even generalize studies on pigeons… Homans in exchange theory took these pigeons from psychology.

PW: Probably from Skinner.

JG: We must finish because we other duties are waiting …

References

Bollen, K., & Pearl, J. (2013). Eight Myths About Causality and Structural Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301–328). Springer.

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

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. (2021). The Practical Alternative to the *p* Value Is the Correctly Used *p* Value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Woolston, C. (2015). Psychology journal bans P values. *Nature*, *519*, 9. https://doi.org/10.1038/519009f