PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 145–164
DOI: 10.31648/przegldpsychologiczny.9680

A Credibility Crisis in Psychology?

Jerzy Marian Brzeziński[1]
Adam Mickiewicz University, Poznań
Faculty of Psychology and Cognitive Science
https://orcid.org/0000-0003-1582-4013

Abstract

The interest in the global result obtained by B. Nosek's team increased significantly, not only among psychologists, after an article presenting the results of a large-scale international replication of psychological empirical research had been published in *Science* (cf. Open Science Collaboration, 2015). While 97% of the original research yielded statistically significant results ($p < .05$), only 36% of the results were significant in the replication. The author of the present article postulates that this result laid the ground for unjustified generalizations about the methodological weaknesses of psychology as an empirical science. Psychology is an empirical science, but it also has its peculiarities due to the specificity of the subject matter and the method (e.g. Orne, 1962, 1973; Rosenthal, 1966/2009; Rosenzweig, 1933). Equally importantly, psychology is not practiced in social or cultural isolation. Finally, psychological research is bound by rigorous ethical standards/constraints, and psychologists (as well as researchers in other fields) who publish the results of empirical research to be analyzed statistically are constrained by the editorial practices of scientific journals. Journals have an interest only in papers that present statistically significant results (where "$p < .05$"!), which leads to the so-called file-drawer effect (Rosenthal, 1979). As strongly emphasized by the author, the debate cannot be limited to the statistical significance of psychological research (in particular the power of statistical test which has emerged as a popular trend in recent years). In this article, the author discusses (and presents his point of view) the following problems: 1) the methodological specificity of psychology as an empirical science, 2) the triad of statistical significance (the problematic criterion of "$p < .05$"), effect size, and the power of a statistical test, 3) the socio-cultural context of psychological research, 4) researchers' failure to follow methodological and ethical guidelines, and 5) possible precautions and remedies.

Keywords: science, intersubjectivity, stability, rationality, credibility, replication, psychological research, statistics, statistical test, confidence interval, $p < .05$, power of a statistical test, effect size, data fishing, *p*-hacking, HARKing, interpersonal expectations, demand characteristic, file-drawer effect, pre-registration research

[1] Correspondence address: brzezuam@amu.edu.pl.

Psychology (presumably like other scientific disciplines) also succumbs to popular trends, where an increased interest in some **theoretical** issues can be observed in the longer or shorter perspective. In some cases, it can be argued with a high degree of probability that a particular theoretical solution has entered the canon of theoretical achievements in psychology. In my opinion, in the field of intelligence theory, such claims can be made about the achievements of the probably generally accepted Cattell-Horn-Carroll theory of intelligence (commonly abbreviated as "CHC"[2]). It may be worth noting that the CHC theory provided the theoretical basis for the latest editions of David Wechsler Intelligence Scales, namely the Wechsler Adult Intelligence Scale (WAIS-IV) (the Fifth Edition is pending final standardization in the US; the delay has been due to the COVID-19 pandemic) and the Wechsler Intelligence Scale for Children (WISC®-V[3]). However, research has shown that some promising "stars" shed a false light (years later), as was the case with the pseudo-scientific concept of Bert Hellinger's Family Constellations.

The history of our discipline has also witnessed the rising popularity of **workshops** (some of which were borrowed from other empirical sciences) dealing with methods of data collection or data analysis (especially statistical analyses whose dynamic growth was driven by advances in computer technology and the incredible development of highly sophisticated statistical software; programs such as SPSS or SAS are already outdated toys). Specific topics are addressed by monographs, scientific articles, technical papers[4], or conference and workshop speeches. The above also applies to self-report inventories (personality questionnaires) or pseudo-scientific projective tests (such as the Rorschach Test or Koch's Tree Test).

To fully comprehend the achievements of a scientific discipline, in particular a discipline as young as psychology (which emerged around 150 years ago; how much is that in comparison with the long history[5] of physics, mathematics or biology?), one should not only examine the factors that connect psychology to other empirical sciences, but also its peculiarities or specificities. The unifying factor in all scientific disciplines is the structure of the research process. A psychologist, a sociologist, a biologist, or a psychiatrist all reach for the same statistical tools when testing hypotheses, whereas the specific nature of the subject of research imposes both methodological (relating to the method) and ethical (relating to the psychologist's conduct towards the patients) constraints. This specificity is primarily responsible for the inadequacies of psychological research, such as poor reproducibility between replications.

---

[2] Cf. for a good review of the CHC theory, refer to Schneider and McGrew (2012).

[3] In 2020, the fifth edition of the Wechsler Intelligence Scale for Children® (2014) was adapted to Polish by Pracownia Testów Psychologicznych PTP (Joanna Stańczak, Anna Matczak, Aleksandra Jaworowska, and Iwona Bac). See: https://www.practest.com.pl/wisc%C2%AE-v-skala-inteligencji-wechslera-dla-dzieci-%E2%80%93-wydanie-piate.

[4] Cf. the development of the "R" programming language which is also used to design statistical programs (cf. e.g. Schwarzer, 2022).

[5] As understood by the historian Fernand Braudel (1902–1985).

Primary Problem: What Kind of Science Is (or Should Be) Psychology?

Several decades ago, the rapid development of **experimental tools in psychology** was influenced by the **analysis of variance** (ANOVA)[6], which was **borrowed from the natural sciences** and invented by Ronald A. Fisher, an outstanding statistician dealing with experimental agricultural research (1925/1938, 1935/1971)[7]. Along with MANOVA, ANOVA is a statistical model of modern experimentation in psychology (that has been used since the 1950s) that has enabled psychologists to move beyond comparisons of two groups only (experimental and comparator) and has created two new research opportunities, namely analyses of curvilinear relationships and interactions between two and more independent variables. These tools have contributed to significant advances in testing new research hypotheses, although the achieved progress might be long forgotten or underestimated today. Psychology has been importing statistical innovations for decades (thus moving closer to the natural sciences in this regard).

In recent years, a number of Polish psychology studies have dealt with **the power of a statistical test**, in particular the applicability of significance tests such as Student's t-test or tests used for ANOVA and MANOVA models. The problem is not new[8], but it has been brought to light by international teams dealing with the issue. The **incomplete reproducibility** of results has been a source of embarrassment for psychologists conducting empirical research.

Also, I wish to emphasize (and I will address this issue later) that **the unsatisfactory level of reproducibility in empirical research is not caused solely by insufficiently sophisticated statistical methods used by psychologists.**

---

[6] Fisher's textbook entitled "Statistical Method for Research Workers" has had as many as fourteen editions, two of which were published after his death. His other textbook, entitled "The Design of Experiments", was also popular (Fisher, 1935/1971: nine editions, the last one was published in 1978 after Fisher's death in 1960). Fisher was born in 1890 and died in 1962. In my opinion, after the Second World War, there were three main textbooks (significantly revised in subsequent editions) for psychologists that have shaped the research practice of psychologists planning experiments according to the requirements of the ANOVA statistical model. These textbooks were written by: Edwards (1950/1960/1968/1972), Winer (1962/1971; last: Winer, Brown, and Michels, 1991), and Kirk (1968/1982/1995; last: Kirk, 2012).

[7] In Polish psychology: Brzeziński and Stachowski (1981/1984), Brzeziński (2012).

[8] For example, two anthologies of older studies conducted by psychologists (which did not attract significant attention in Poland), namely Henkel and Morrison (1970), and Harlow, Mulaik, and Steiger (1997). Most of all, I wish to refer to the seminal papers by Jacob Cohen (1990, 1994) (included in various anthologies of methodological papers), which I, together with my friend Professor Jerzy Siuta (1943–2018) of the Jagiellonian University, decided to make available in Polish years ago (Brzeziński & Siuta, 1991); please note that original versions were not highly accessible at that time, especially for students. In my opinion, the views and conclusions formulated in these papers **are still relevant** today, and they should be placed on the **required** reading list of master's degree and doctoral students of psychology.

The methodological immaturity of psychology has been well addressed in the literature, although the formulated arguments are strong and, in my opinion, overly critical. The paper published by 125 authors in *Science,* one of the two most prestigious scientific journals (where *Nature* probably takes the lead) probably attracted the greatest interest (cf. Open Science Collaboration, 2015)[9].

The efforts of this research team deserve at least brief attention. The empirical research presented in 100 (out of 488) articles published in 2008 in three prestigious psychological scientific journals, namely *Psychological Science* (PSCI)*, Journal of Personality and Social Psychology* (JPSP)*,* and *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP: LMC), has been **replicated**. Thirty-two (out of fifty-five) articles were published in JPSP, twenty-eight (of thirty-seven) in JEP: LMC, and thirty-nine (out of sixty-four) in PSCI. Two articles involved two replications each. The thematic scope of the research involved 43 cognitive studies and 57 social-personality studies. The following brief conclusions can be formulated (refer to the link in the References section for detailed source data): while 97% of the original research yielded statistically significant results ($p < .05$), that result was significantly lower in the replicated research (i.e. 36%). An analysis of effect size indicators revealed that only 47% of the indicators obtained in the original research fell within the 95% confidence interval for the replicated indicators. Please note that **the analyses were limited to the studies presented in three journals only, all of which were published in 2008 and focused on the cognitive profile (43 studies) and the social-personality profile (57 studies).** However, psychology is not limited to these research domains. Nevertheless, these data laid the groundwork not only for serious theoretical and methodological debates, but also for some hate comments against **psychology as a whole**.

The debate on the demise of psychology (bolstered by this article) focused on statistical indicators, including **statistical significance**, namely $p < .05$ versus $p > .05$, and unsatisfactory **effect size** indicators obtained in replicated studies.

In my opinion, this project is a good starting point for further replication analyses which have already been undertaken. Other empirical studies that are relevant to cognitive research, have practical implications, and propose different methodological approaches should also be considered. Be that as it may, the replications (which, in addition to providing raw data, contribute test statistics such as t or F whose values do not meet the "magic" criterion $p < .05$, or meta-analyses) serve as a therapeutic measure for addressing the chaos and ethical violations resulting from the overwhelming and irrepressible desire to print anything anywhere (such as predatory journals)[10] and the destructive submission to the "publish or perish" paradigm. Scientific mediocrities must also comply with this paradigm, in particular in contemporary Poland.

[9] 7,777 quotes according to Google Scholar as of November 14, 2022.

[10] Suspicious titles can be checked on the list of predatory journals, namely *Cabell's Journalistic and Predatory Reports*; the list is commercial (paid access) and available on: https://www2.cabells.com/predatory (all the details are available there).

Nonetheless, shouldn't the problem be analyzed from a much broader perspective than that dictated by editorial practices in scientific journals (against the recommendations of APA expert panels[11])? In no way do I disregard them, but I also do not regard the statistical significance indicator, identified with the $p < .05$, as inviolable (cf. Skipper, Jr. and Guenther 1967/1970). I am also aware of these limitations, in particular when they are applied to **poorly measured data**, sometimes quite thoughtlessly. Please note that psychology can only invoke "hard" measurement results. Unfortunately, more often than not, these data are self-reported (various personality questionnaires and estimation scales). Regrettably, the application of powerful (as defined by statistical models) statistical tools to such results will only create an appearance of precision and scientific accuracy. In that regard, perhaps the results obtained by the Open Science Collaboration team should not be excessively questioned.

Similarly to other scientific disciplines, psychology is not practiced in social isolation (in an ivory tower). Research practice is not only affected by internal psychology-specific interactions that are planned by the researcher and effectively controlled by **internal forces**. It is also affected by a wide range of **external factors**. This is something we should be aware of. If these forces are not taken into account (out of ignorance or insufficient education), the reproducibility of the results of empirical research is unlikely to be satisfactory, unless they involve trivial questions with predictable answers. But then, why should one spend any time and money (often taxpayers' money) on endeavors pretending to be scientific research?

Unfortunately, when we look closer at the content of Polish psychological journals, so-called research bulletins or collective works, we will find a large number of articles that are substantially meagre (they may be "scholastically" correct, but they do not take into account all statistical procedures). Most articles present new personality questionnaires or review the results obtained through self-reporting methods. All you need to do is find a few questionnaires in the right drawer. This is not a highly challenging task, but the results are not always inspiring. Well, I need to write something for the writing's sake, so here we go. Some predatory magazines will print it for a price (in English, of course).

However, **chasing statistical analyses will not suffice**, nor will putting subjects into a CT scanner or scanning the brain (these tools have become available to researchers, including psychologists, only recently). Regrettably, psychology (the one with a capital "P") is a difficult scientific discipline to research. This difficulty does not stem from the degree of technological complexity (which cannot be overlooked), such as that encountered in cutting-edge research conducted by nuclear physicians at CERN in Geneva.

Psychological research is also difficult because one person (the experimenter) conducts research on another person (the subject). This is often the case in clinical

[11] Cf. Wilkinson and Task Force on Statistical Inference, American Psychological Association Science Directorate (1999); APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008); American Psychological Association (2020).

psychology. It can be really very difficult indeed (especially when the subjects are little children or people with various disabilities, not just intellectual ones).

Many years ago, the American psychologist Saul Rosenzweig (1907–2004) identified three peculiarities in experimental research in psychology (Rosenzweig, 1933; cf. Larsen, 2005)[12]: (1) the experimenter becomes an element of the experimental situation, (2) the subject's behavior in the experimental situation is influenced by variables related to and characterizing the subject, such as personality, motivation, etc.; (3) an interaction is established between the experimenter and the subject. In my opinion, this is a very important article that preceded the work of psychologists such as Orne (1962, 1973) or Rosenthal (1966/2009; also: Blanck, 1993; Trusz, 2013[13]).

The work of these psychologists drew attention to the fact that the subjects are able to identify the purpose of the study and modify their behavior accordingly during the experiment. Orne spoke of cues that suggest the content of the research hypothesis to the subject (he used the term *demand characteristics*). The subject is able to predict the type of behavior that is expected by the investigator, and he or she will try to adjust their behavior to comply with the identified hypothesis or act against it, depending on how they perceive the investigator (friendly or threatening). In contrast, the studies (which have been replicated multiple times) conducted by the social psychologist and methodologist Robert Rosenthal (who is also a psychotherapist, a teacher, a judge, and a sports coach) revealed that the experimenter generates interpersonal expectations towards the study subject by promoting behavioral attributes that are consistent with the hypothesis. In reference to an anthology of articles compiled by Arthur G. Miller (1972: "The Social Psychology of Psychological Research"), research on the psychological conditioning of the research process can be referred to as *social psychology of psychological research* or *psychology of psychological methodology*.

Today, only a handful of psychologists, in particular in methodologically "soft" research areas (such as clinical psychology or health psychology), recognize the interaction between the examiner and the subject as a separate and important source of variance. This is most unfortunate.

When forced to answer questions such as "What kind of science is psychology?", "Which psychological theories can be regarded as scientific?", I always say that psychology is an empirical science that does not differ in that regard from other sciences, such as biology. Each new psychological theory (or a candidate for a theory, namely a hypothesis) must be confronted with empirical facts in an empirical test. This does not need to be a laboratory experiment. It could also be a field experiment or a clinical study, as long as it complies with

[12] As Larsen pointed out: "… in his first published article, ("The Experimental Situation as a Psychological Problem" 1933), in which he explored the reciprocal interaction between the experimenter and the subject and laid the foundation for the work on experimenter expectancy effects that flourished a generation later." (p. 259)

[13] Special attention should be paid to the extensive (817 pages) and representative selection of Polish translations of papers, both classical and contemporary, dealing with the Rosenthal effect (Trusz, 2013).

the methodological standards of professional conduct that have been approved for clinical psychology, and as long as diagnostic and therapeutic practices are based on empirical evidence (cf. American Psychological Association Presidential Task Force on Evidence-Based Practice, 2006; Brzeziński, 2016).

Kazimierz Ajdukiewicz (1949–2003, 1958), a prominent representative of the Lviv-Warsaw School of philosophy and logic, argued that: (1) a scientific proposition must be consistent with the **principle of intersubjectivity: communicability and testability**, and (2) the degree of certainty with which a given proposition is formulated should be proportional to the degree of certainty regarding the rationale, which depends on the certainty of the method (ideally, the research should be experimental, rather than correlational) used by researchers to confront their statements with empirical facts; this is the **principle of rational recognition of beliefs.** Both of these principles, namely (1) the principle of intersubjectivity and (2) the principle of rational recognition of beliefs, constitute the **principle of rationality.** In my previous work (Brzeziński, 2019), I have argued that:

> Research psychologists should always adhere to the **principle of rationality**. The fact that psychology is an empirical science means that only statements based on the results of well-**controlled** empirical studies may be deemed scientific by a psychologist. (p. 17)

Obviously, this approach to psychology rules out the psychology of various diagnostic (sometimes very exotic) or nearly miraculous healing practices. Evidently, nonsensical diagnoses or conclusions are not prohibited by law. Some researchers may care about the opinions voiced by pop psychologists, while others may study exorcisms, as is the case in faculties of Catholic theology at secular universities (but let's not pretend that this is science). We live in a free country. However, it is hard to accept that such views are expressed in some universities. These institutions are supervised by the Ministry of Education and Science. Individuals who have obtained a master's degree in psychology (or only a bachelor's degree in psychology) can monetize such "scientific" views during private "psychological" practice, and such practices should be rigorously controlled. The above requires active (and not dormant) regulations on the psychological profession, which remains a pending issue. As a result, psychology is being destroyed both as a science and as a practice with the support of the Ministry of Education of Science and the Polish Accreditation Committee.

How To Avoid Trivial Results? Should $p < .05$ Determine the Scientific Value of Empirical Research?

Statistical compilations of empirical data from articles published in highly-acclaimed scientific journals and psychological science journals differ significantly from those that had been published in the same journals fifty years ago.

The above can be attributed not only to theoretical and methodological progress of our scientific discipline, increasing social methodological awareness, but also to technological advances such as more accurate measuring equipment and psychological tests, powerful personal computers, and specialized software (including statistical software!). Philosopher Hans Reichenbach (1938/1989) identified two contexts: the **context of discovery** and the **context of justification**, where the former relates to the scientific robustness of psychology and new empirical theories, whereas the latter accounts for the methodological robustness of empirical research which aims to verify such theories (which, in my opinion, should be conducted in the spirit of Popperian **falsification** (Popper, 1974) by making rigorous attempts to disprove or falsify, rather than **confirm** a theory or a hypothesis, that is by searching for empirical facts that confirm the theory or hypothesis.

More often than not, serious scientific research is conducted by psychologists in interdisciplinary teams. New non-standard scientific solutions are developed on the verge of different disciplines. The rapid development of cognitive science could serve as a positive example. I believe this is the future of empirical psychology, contrary to isolating it from other disciplines, in particular those that are more methodologically advanced (devices, measurements, advanced quantitative analyses). Therefore, psychology researchers should look for partners in the field of brain science or biology, rather than pedagogy.

The assertion that not every scientific proposition, even one that has emerged in print, adds something new to the field of psychology is not new. What about statements that purport to be serious rather than trivial, including those formulated by psychologists (perhaps too often)? Again, let me refer to the outstanding philosopher and methodologist Kazimierz Ajdukiewicz (1957/2020) who, in an essay entitled "O wolności nauki" ["On the freedom of science"], wrote that any subject undertaken by a researcher deserves to be called scientific when it meets the four following conditions. **Firstly**, the topic of scientific inquiry should address "matters that are important for science" (p. 9). In addition, the formulated proposition must "enrich science significantly." **Secondly**, the researcher's proposal is expected to be "formulated with due accuracy" (p. 9). **Thirdly**, the formulated conclusions should account for the exploratory power of the applied method, namely "…the certainty with which a proposition is made should correspond to the certainty of the underlying rationale" (p. 9). **Fourthly**, a researcher must have extensive knowledge of the studied field (p. 11).

If statistical significance were the **only** criterion for deciding whether a research hypothesis has been empirically positively tested (in my opinion, satisfactory plausibility is a better term in this context), its correctness would have to be tested by increasing the group size *ad absurdum*. This approach was openly ridiculed by Cohen (1994):

> In an unpublished study, Meehl and Lykken cross-tabulated 15 items for a sample of 57,000 Minnesota high school students, including father's occupation, father's education, mother's education, number of siblings, sex, birth order, educational plans, family attitudes toward college, whether they liked school,

> college choice, occupational plan in 10 years, religious preference, leisure time activities, and high school organizations. AH of the 105 chi-squares that these 15 items produced by the crosstabulations were statistically significant, and 96% of them at $p < .000001$. (p. 1000)

Cohen ironically concluded that "Everything is related to everything else" (p. 1000).

The magic of large samples (even those that lead to absurd conclusions, but are justified by $p < .0001$!) sometimes creates artifacts (as recognized by Cohen in the above citation). Considerable damage (not only in psychology) has been wrought by the new **methodological pseudo-standard** which assigns statistical meaning to the term "**significant**" (e.g. a significant correlation between two variables, a significant difference between two means, etc.) that relates solely to NHST[14] statistical methods and links the rejection of H_0 with a probability of $p < .05$. Why do researchers do this, and why do they shelve results that do not meet this criterion? The answer is they have been effectively trained by statistical role models in the field of statistical applications to objectively recognize achievements that are worthy of dissemination in the research community. Editorial practices in scientific journals and review practices used in promotion and competition have played a significant role in this process. These practices were correctly interpreted by fast-learning authors (instead of the Milgram electric shock experiment, prospective authors were offered the shock of having their article rejected for not exceeding the $p < .05$ criterion), and they deeply affected honest researchers. These practices have been perpetuated generation after generation. Are we dealing with a fast-acting mechanism of natural selection?

I do not wish to repeat the arguments made by the critics of this binary approach. After all, the acceptable risk when rejecting the true H_0 (the possibility of making a *type I error* with probability α) when its social cost is small (the student defends an undergraduate thesis that will sink into the abyss of the digital archive) is totally different from recognizing that result as a basis for further research on a new therapeutic method (high social cost of error). In the latter case, it is socially justifiable to move to a more stringent level of probability, for example $p < .001$, which increases **confidence in the researcher's decision**, but does not prove its theoretical validity. However, the researcher can manipulate the size of N to achieve statistical significance at any cost. In "Statistics for the Social Sciences", a popular textbook authored by the outstanding psychologist and statistician William L. Hays (1925–1995; 1973, pp. 422–424), section 10.22 ("Can a sample size be too large?") reads: "trivial associations may well show up as significant results when the sample size is very large" (p. 424).

A researcher should also control the risk a *type II error* when a false H_0 with probability β is not rejected. In psychological literature, a recent special issue of *The Review of Psychology* (better late than never) was dedicated to the **power of a statistical test**, which clearly indicates that the importance of the problem

[14] An acronym for Null Hypothesis Significance Testing.

has been recognized. It is generally known that the power of a test can be increased by increasing the sample. The researcher should (at the stage of planning the research!) be flexible. Cohen's seminal work, "Statistical Power Analysis for the Behavioral Sciences" (1988), cannot be ignored in any serious study on the power of a test.

I will not elaborate on the power of a test because there are many insightful studies on statistical significance and the power of a test (including in Polish psychological literature). The most notable recent examples include the studies conducted by Piotr Wolski (2016a, 2016b, 2016c), Tytus Sosnowski and Liliana Jarmakowska-Kostrzanowska (2020), as well as a translation of a statistics textbook for psychologists and educators by Bruce M. King, and Edward W. Minium (2009). I will also refer to the first study in the Polish psychological literature, where the power of a test was discussed in the context of planning experiments according to the ANOVA model (Brzeziński and Stachowski, 1981/1984).

The following triad of statistical indicators must always be considered when planning empirical research, and not only after the research has been completed (e.g. by formulating *ad hoc* hypotheses): **statistical significance level, the power of a statistical test, and the effect size** (cf. American Psychological Association, 2020; Grissom & Kim 2005, 2011; King & Minium, 2003/2022; Rosenthal et al., 2000; Wilkinson & Task Force on Statistical Inference, American Psychological Association, Science Directorate, 1999). This approach is now a **standard procedure** (but not yet in Poland!) that is recommended by APA reports (cf. footnote 11). In the latest textbook on psychological research methodology (Brzezinski, 2019), which has been considerably updated, I included and reviewed two figures recommending the effect size indicator(s) for each significance test (Figure 10.4, p. 221 and Figure 10.5, p. 223).

Sanford Labovitz (1970) proposed 11 criteria for selecting a significance level. However, when **exploring** a scientifically interesting problem, these criteria should be relaxed, and $p = 0.15$ or $p = 0.20$ should not be disregarded. Before making a fateful statistical decision, a researcher should first take a close look at the results and data distribution, and only then draw conclusions with the appropriate statistical tools (which is something that I have learned from John B. Tukey's "Exploratory Data Analysis" (1977), a must-read for all scientists).

To sum up, a researcher should break with the bad tradition of searching for a statistically significant result (through an ethically questionable procedure of data fishing or *p*-hacking) at any cost (most often than not through ill-reasoning and increasing sample size). This is often the case in research conducted by sociologists, social psychologists, health psychologists or educators on the Amazon Mechanical Turk (AMT) website (cf. Aguinis et al., 2021; Brzezinski, 2023; Buchanan & Scofield, 2018; Buhrmester et al., 2018; Keith & Harms, 2017; Saad, 2021; Webb & Tangney, 2022) or similar Polish websites. This type of research is not highly sophisticated in terms of methodology. It involves self-reporting methods such as personality questionnaires, attitude scaling, and surveys. To pass muster, $p = .05$ is always achieved on large samples. However, large samples are not always feasible (for example, when access to prospective subjects is

limited, including in clinical studies that involve subjects with rare disabilities or when the costs of individual assessments are very high). After all, some tests have been specifically designed to analyze the significance of differences in "difficult" studies with a small N, including Student's t-test, Fisher's exact test for 2×2 tables, chi-square test, Wald's sequential analysis or nonparametric tests for ordinal scales, such as the Mann-Whitney test by ranks, Wilcoxon signed rank test, Kruskal-Wallis test by ranks, and Friedman test by ranks. These tests are discussed in contemporary statistics textbooks, and they are included in statistical packages.

Conclusions

This article (I would like to thank the editorial team for inviting me to write this paper) was written largely in response to the Open Science Collaboration paper published in *Science* (2015) and the resulting criticism of research practices in psychology (clearly overgeneralized). To summarize this discussion, I will present the conclusions in two separate sections: (1) Why did this happen and is it still happening? (2) What can be done to minimize losses (including the loss of image) and remedy the situation? (which I believe is possible).

There is one more thing: **psychologists are not the only scientists who manipulate research findings** (by *p*-hacking, HARKing[15], falsifying or inventing results, adding their name to papers in which their participation was minimal, of little significance or null, or even plagiarizing), and low reproducibility is a problem that affects all scientific disciplines. To some degree, academic science is also being turned into scientific junk in other fields (Grabski, 2015, p. 180; "if you are a rational thinker, you cannot expect to be practicing real science this way")[16]. According to Maciej W. Grabski:

[15] An acronym for Hypothesizing After the Results are Known.

[16] To back up my claim, let me give a spectacular example of abuse in "hard" sciences. On 28 October 2022, the popular *Gazeta Wyborcza* daily published an article by Paulina Mozolewska entitled "Scandal in Groundbreaking Research" (p. 16) with the following runner: "What about the Treatment of Alzheimer's Patients?". It was preceded by information from the author, "According to investigative reporters, studies on Alzheimer's disease may have been manipulated [the criticism concerns an article by S. Lesné and K. H. Ashe published in *Nature* in 2006 – a note by J. M. B.]. Physicians and scientists wonder how this will affect the research on potential treatments for Alzheimer's disease" (p. 16). According to the article, based on the results of a six-month-long investigation, the *Science* editorial team concluded that: "… key results that serve the basis for numerous studies over Alzheimer's disease conducted for many years might have been manipulated or deliberately falsified" [manipulation of photographic documentation – a note by J. M. B.]. The article is accompanied by a long interview with Professor Tomasz Gabryelewicz of the Mossakowski Medical Research Institute and President of the Polish Alzheimer's Association. In that interview, Gabryelewicz said, "The accusations relate to photographs presenting the results of protein level analyses. In these tests, protein levels are represented by bands.

> The softer the data, the more we drift away from the main topics of scientific inquiry, the weaker the relationship with major research institutions, the higher the risk of scientific dishonesty, the lower the probability of fraud detection. In such situations, academic science is easily transformed into something that is increasingly referred to as junk science. … The amount of junk science grows exponentially with the number of scientific institutions and local journals operating outside the peer review system, and even reputable periodicals are not immune. …
>
> To make matters worse, junk science is often a useful element of manipulation, because by deliberately falsifying and misinterpreting data, and manipulating scientific analyses, junk science supports preconceived viewpoints, fosters a supportive environment for manipulators, hoaxers and fraudsters who are often titled and act with impunity, and contributes to sensationalist media coverage. (pp. 180–181)

If you look critically at the development of psychology in Poland, which is (also) influenced by the quality of second-cycle psychology programs (after all, whether we like it or not, we are all mortal, and academic staff is being gradually and naturally replaced), I am deeply concerned about the rapid spread of centers training future psychologists. Given the size of the academic population (persons who hold at least a doctoral degree), the large number of psychologists who are presently being trained in Poland cannot expect to receive a decent education, in particular in non-public universities (private schools are a lucrative business, especially if they are cheap to run, which applies particularly to schools of pedagogy, management, political science and … psychology)[17].

The defacement of psychology begins when a student writes a mediocre master's thesis (the responsibility rests mainly with the supervisor and the reviewer, but where do you get the required number of well-prepared thesis supervisors?).

Why?

Now, let us consider the possible reasons for the proliferation of mediocrity. In my opinion, several factors contribute to undesirable, embarrassing or even reprehensible behaviors in the research community.

In simple terms, band size and thickness denote the concentration of a given protein. It is unclear if these photos are a consequence of a premeditated fraud or an attempt to 'tweak' the results. … According to most opinions, Lesné might have tweaked the photos, but this is not the most important allegation. Some experts have suggested that the alleged manipulations may be digital artifacts that occurred accidentally during image processing. **Regardless of whether the photographs were intentionally manipulated or only 'tweaked' through photoshopping, the whole situation leaves one with a sense of embarrassment and distaste."** [Bolded by J. M. B.]. (pp. 16–17)

[17] Cf. https://radon.nauka.gov.pl/dane/studia-prowadzone-na-okreslonym-kierunku.

Above all, mediocrity is caused by hubris, rivalry, and the desire to stay at the top. Working conditions and financial risks (precarious employment) also play a role. If you wish to be one of the best, you must be prepared to live under constant stress. People like Diederik Stapel of Tilburg University (who belonged to the elite of social psychologists, after all, and was no stranger to financial concerns) are motivated only by the desire to stay at the top, attend prestigious conferences, and to be printed (and quoted!) in the best professional journals. As they run out of ideas, so does their resistance to temptation and they start spiraling downwards.

Secondly, there is pressure from the employer, namely the head of the department, the director of the institute, a faculty dean or a rector. In recent years, Polish academics (again, that's our Polish peculiarity!) have experienced increased pressure from the management of scientific institutions to publish articles in high-ranking journals and accumulate sufficient academic credit to initiate the post-doctoral procedure. After all, an institution which has earned enough credit will be classified in a higher category during the evaluation that is performed every four years by the Ministry of Education and Science and the Science Evaluation Board based on the results of scientific research in a given scientific discipline (psychology in this case). In extreme cases, an excessive and mechanical evaluation of publishing success prompts researchers to cut corners by adding their names to papers authored by other scientists, copying or plagiarizing papers in whole or in part, buying statistical reports in whole or in part (advanced statistical analyses conducted by specialized companies or experts, or paying cash under the table), attempting (often successfully, regrettably) to publish quasi-scientific articles in predatory journals or books, or manipulating data (cf. Brzeziński & Oleś, 2021, Chapter 10, Section 10.2: "Ethical principles of scientific research", pp. 411–475).

Thirdly, there is social acceptance and hardly any consequences for such actions. Pathological behavior is also fostered by the lack of an unequivocally firm response from the academic community to violations of academic standards, especially from university authorities (at every level! Ethical issues that undermine an institution's reputation are also swept under the carpet).

Fourthly, the number of low-quality institutions of higher education continues to increase (in particular in the non-public sector). Staffing requirements have been already reduced in new fields of study, and schools that should have been closed for the sake of decency are being maintained (or "resuscitated"). The academia attracts people who have no interest in scientific advancement, are not talented, or able to write a decent scientific article. Please note that the principle of normal distribution also works in this case. These people can either leave the university (but where would they go?) or try to cut corners.

The fifth factor is the publication practices of psychology journals. In order to publish an article, psychologists are required to demonstrate that their research proves "something". This means that only studies with a statistically significant result at the minimum required level of $p < .05$ have a chance to be published. This is why some researchers would do anything (including data manipulation!) to make it work. Statistical significance is identified with the extent to which

the independent variable influences the dependent variable. It was only recently that serious journals have begun asking the authors to describe the effect size indicators, which inform about the power of one variable (or the interaction between two or more variables) over the dependent variable, and not only the *level of significance*.

Precautions and Remedies

What precautions and remedies can be undertaken to address these problems? In my opinion, **four complementary measures** are possible.

The first measure is the **principle of transparency.** Due to the limited framework of an empirical paper (in particular a short report), the information about the studied groups and the results of statistical analyses cannot be presented in sufficient detail. However, a researcher should be ready to provide such data, for example, by making it available to the journal that will act as a depository of these data (for a certain period of time). Some editors request **raw data** (to avoid an ethics violation and "brilliant" publications, such as those submitted by Stapel; cf. Budzicz, 2015) that can be re-analyzed. I do not agree that data belong solely to the researcher and that the researcher is the only person who has the right to use them. This is intolerable, in particular when the research was financed from public funds (taxpayers' money!), as is the case in the grant system of the National Science Center.

The second measure is **reproducibility** (Neuliep, 1991; Wolski, 2016b). Only the results replicated by other researchers may be deemed to be of any scientific value.

The third measure is **the publishing policy of scientific journals.** Today, journals are reluctant to publish articles that replicate previously published research findings. The editors reserve the right to publish original results only! As a consequence, **we do not know how many unpublished articles have been shelved** solely because the authors did not obtain a value of $p < .05$ and were unwilling to violate ethics principles and "correct" the data. This is known as a negative file-drawer effect. Please note that such a biased publishing policy affects the bias of meta-analyses (unpublishable negative results are not available for the meta-analysis), which inflates their results.

To prevent that, editors should develop a new approach for approving papers for publication. The initiative promoted (although to a limited extent) by renowned journals inspires hope (but not for all researchers). This initiative involves reviewing not only the manuscript, but the entire research concept (before it was conducted). If the study receives positive feedback and is accepted by the reviewers, the editors assure the author that the results (whether $p < .05$ or not $p > .05$), obtained in accordance with the reviewed concept, will be published. This new publication format is known as **"pre-registration research.**"

The fourth measure, which in a sense is a consequence of the third measure (as mentioned above), is a **flexible approach to the *p*-value**. Why should $p < .05$ be deemed as an absolute measure of the scientific value of the result?

This is merely a matter of **convention**. This approach was adopted by Fisher (1925/1938), the author of the acceptance threshold regarding normal distribution characteristics[18]. Instead, it is rational to switch to **effect size** indicators and **confidence intervals** (cf. King & Minium, 2003; Loftus, 2008, 2012). That approach is recommended by many serious researchers, both methodologists and statisticians. In particular, **providing values for the limits of the confidence interval is of extreme scientific value.** As Cohen (1994) put it, sometimes, they are "embarrassingly large."

> "Everyone knows" that confidence intervals contain all the information to be found in significance tests and much more. They not only reveal the status of the trivial nil hypothesis but also about the status of non-nil null hypotheses and thus help remind researchers about the possible operation of the crud factor. Yet they are rarely to be found in the literature. I suspect that the main reason they are not reported is that they are so **embarrassingly large! But their sheer size should move us toward improving our measurement by seeking to reduce the unreliable and invalid part of the variance in our measures** (as Student himself recommended almost a century ago). Also, their width provides us with the analogue of power analysis in significance testing – larger sample sizes reduce the size of confidence intervals as they increase the statistical power of NHST. [Bolded by J. M. B.] (p. 1002)

I should point out that it has been almost 30 years since Cohen's article was printed, but not much has changed when it comes to publishing the results of psychological research.

* * *

I hope I have managed to convince you that psychology is a difficult research discipline if it is taken seriously. It is not easy to apply psychological theories or individual statements that have emerged victorious after a confrontation with empirical tests in the field of social (either diagnostic, judiciary, penitentiary or therapeutic) practice. This approach makes sense (and is also ethical) only when it complies with the increasingly stringent methodological requirements. Therefore, the effectiveness of the assistance procedures used by professional psychologists is correlated with the progress of scientific research (such as the development of new measuring procedures that enable the formulation of more reliable and accurate diagnoses, as well as such therapies). To that end, psychologists should not overly focus on specific technical problems relating to,

[18] According to Fisher: "… The value for which P = .05, or 1 in 20, is 1.96 or nearly 2; it is **convenient** to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviations are thus formally regarded as significant" (p. 46).

for example, the statistical test for evaluating the null hypothesis, which has already been referred to herein. In other words (and in no way do I underestimate the importance of that issue), the elaboration of correctly collected results in the language of modern statistics is only one component of the entire research procedure. Statistics is only a tool that can be used responsibly only when the researching psychologist knows it well (technical proficiency in using an SPSS statistical package will not suffice) and when he or she (also) knows the limits of its substantiated (defined by the assumptions) model, namely the significance of differences test, effect size indicators, confidence intervals, and correlation measures. Unfortunately, statistics provide the reviewers and editors of scientific journals with what seems to be a simple and quantitative criterion for assessing the validity of the obtained result, namely a statistical significance indicator of $p < .05$. I have tried to show in this paper that this is the wrong approach and that the statistical significance of $p < .05$ cannot be a binary criterion for considering a result to be scientifically interesting. The publishing decisions made mechanically as part of the procedure of qualifying an article for publication in a scientific journal, based on the criterion of "$p < .05$," inhibit scientific progress (by leading to the file-drawer effect). My point is that a flexible approach is needed to selecting the p-value. The effect size is equally important. When evaluating the significance of differences, one should also refer to confidence intervals if practicable by the measurement level of the data.

There is one more thing. Psychologists seem to forget that psychological research is psychological in nature and that the effects detected years ago by Rosenthal or Orne cannot be ignored. Researchers who ignore this will not be reporting facts, but – as Rosenthal rightly noted – artifacts.

References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Ajdukiewicz, K. (1949/2003). *Zagadnienia i kierunki filozofii. Teoria poznania. Metafizyka* [*Issues and directions of philosophy. Epistemology. Metaphysics*]. Czytelnik.

Ajdukiewicz, K. (1957/2020). O wolności nauki [On freedom of science]. *Nauka*, *2*, 7–24. https://doi.org/10.24425/nauka.2020.132629

Ajdukiewicz, K. (1958). Zagadnienie racjonalności zawodnych sposobów wnioskowania [The issue of the rationality of unreliable ways of reasoning]. *Studia Filozoficzne*, *4*, 14–29.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Author.

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Blanck, P. D. (Ed.). (1993). *Interpersonal expectations. Theory, research, and applications*. Cambridge University Press.

Brzeziński, J. (2012). *Badania eksperymentalne w psychologii i pedagogice* (wyd. popr.) [*Experimental research in psychology and education* (Rev. ed.)]. Wydawnictwo Naukowe Scholar.

Brzeziński, J. (2016). Towards a comprehensive model of scientific research and professional practice in psychology. *Current Issues in Personality Psychology*, *4*(1), 2–10. https://doi.org/10.5114/cipp.2016.58442

Brzeziński, J. M. (2019). *Metodologia badań psychologicznych. Wydanie nowe*. [*Methodology of psychological research. New edition*]. Wydawnictwo Naukowe PWN.

Brzeziński, J. M. (2023). Pytania do psychologów prowadzących badania naukowe. [Questions for psychologists conducting research] In A. Jonkisz, J. Poznański SJ, & J. Koszteyn (Eds.), *Zrozumieć nasze postrzeganie i pojmowanie człowieka i świata. Profesorowi Józefowi Bremerowi SJ z okazji 70-lecia urodzin* [*To understand our perception and comprehension of the human and the world. Papers dedicated to Professor Józef Bremer SJ on the occasion of his 70th birthday*] (pp. 289–311). Wydawnictwo Naukowe Akademii Ignatianum.

Brzeziński, J. M., & Oleś, P. K. (2021). *O psychologii i psychologach. Między uniwersytetem a praktyką społeczną* [*On psychology and psychologists. Between university and social practice*]. Wydawnictwo Naukowe PWN.

Brzeziński, J., & Siuta, J. (Eds.). (1991). *Społeczny kontekst badań psychologicznych i pedagogicznych. Wybór tekstów* [*The social context of psychological and pedagogical research. A reader*]. Wydawnictwo Naukowe UAM.

Brzeziński, J., & Siuta, J. (Eds.). (2006). *Metodologiczne i statystyczne problemy psychologii. Wybór tekstów* [*Methodological and statistical problems of psychology. A reader*]. Wydawnictwo Naukowe UAM.

Brzeziński, J., & Stachowski, R. (1981/1984). *Zastosowanie analizy wariancji w eksperymentalnych badaniach psychologicznych* (2nd ed.) [Application of analysis of variance in experimental psychological research]. Państwowe Wydawnictwo Naukowe.

Buchanan, E., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*(3), 2586–2596. https://doi.org/10.3758/s13428-018-1035-6

Budzicz, Ł. (2015). Post-Stapelian psychology. Discussions on the reliability of data and publications in psychology. *Annals of Psychology*, *18*(1), 25–40.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. Perspectives on Psychological Science, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304 .

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997 .

Edwards, A. L. (1950/1960/1968/1972). *Experimental design in psychological research*. Holt, Rinehart and Winston.

Fisher, R. A. (1925/1938). *Statistical methods for research workers* (7th ed.). Oliver & Boyd.

Fisher, R. A. (1935/1971). *The design of experiment* (8th ed.). Oliver & Boyd.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research. A broad practical approach*. The Psychology Press, Taylor and Francis Group.

Grissom, R. J., & Kim, J. J. (2011). *Effect sizes for research. Univariate and multivariate applications* (2nd ed.). Routledge, Taylor and Francis Group.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* L. Erlbaum.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). Holt, Rinehart, and Winston. [1st ed.1963: *Statistics for psychologists*; 5th ed.1994: *Statistics*].

Henkel, E., & Morrison, D. E. (Eds.). (1970). *The significance test controversy. A reader*. Butterworths.

Keith, M. G., Tay L., & Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, *8*, 1359. https://doi.org/10.3389/fpsyg.2017.01359

King, B. M., & Minium, E. W. (2003). *Statistical reasoning in psychology and education* (4th ed.). John Wiley & Sons.

Kirk, R. E. (1968/1982/1995). *Experimental design: Procedures for the behavioral sciences*. Brooks/Cole.

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage.

Labowitz, S. (1970). Criteria for selecting a significance level: A note on the sacredness of .05. In E. Henkel & D. E. Morrison (Eds.), *The significance test controversy. A reader* (pp. 166–171). Butterworths.

Larsen, R. J. (2005). Saul Rosenzweig (1907–2004). *American Psychologist*, *60*(3), 259. https://doi.org/10.1037/0003-066X.60.3.259

Loftus, G. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (pp. 339–390). John Wiley & Sons, Inc. https://doi.org/10.1002/0471214426.pas0409

Miller, A. G. (Ed.). (1972). *The social psychology of psychological research*. The Free Press.

Neuliep, J. W. (Ed.). (1991). *Replication research in the social sciences*. Sage.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://www.researchgate.net/publication/281286234_Estimating_the_reproducibility_of_psychological_science

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. https://doi.org/10.1037/h0043424

Orne, M. T. (1973). Communication by the total experimental situation: Why it is important, how it is evaluated, and its significance for the ecological validity of findings. In P. Pliner, L. Krames, & T. Alloway (Eds.), Communication and affect: Language and thought (pp. 157–191). Academic Press. https://doi.org/10.1016/B978-0-12-558250-6.50014-6

Popper, K. (1974). *The logic of scientific discovery*. Hutchinson.

Reichenbach, H. (1938/1989). Trzy zadania epistemologii [Pol. transl. W. Sady: §1: *The three tasks of epistemology*. In H. Reichenbach, *Experience and prediction* (pp. 3–16). University of Chicago Press]. *Studia Filozoficzne* [*Philosophical Studies*], *7-8*, 205–212.

Rosenthal, R. (1966/2009), Experimenter effects in behavioral research. Appleton-Century-Crofts. In *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books* (pp. 287–666). Oxford University Press.

Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 838–641.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review, 40,* 337–354.

Saad, D. (2021), Nowe narzędzia i techniki zwiększające trafność badań internetowych [Increasing validity of online research by implementing new tools and techniques], *com.press*, *4*(1), 106–121. https://doi.org/10.51480/compress.2021.4-1.248

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of Intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144). The Guilford Press.

Schwarzer, G. (2022). *General Package for Meta-Analysis. Version 6.0-0*. https://cran.rproject.org/web/packages/meta/meta.pdf

Skipper, J. K. Jr., Guenther, A. L., & Nass, G. (1967/1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. In R. E. Henkel & D. E. Morrison (Eds.), *The significance test controversy. A reader* (pp. 155–160). Butterworths.

Sosnowski, T., & Jarmakowska-Kostrzanowska, L. (2020). Do czego potrzebna jest moc statystyczna? [What is statistical power needed for?]. In M. Trojan & M. Gut (Eds.), *Nowe technologie i metody w psychologii* [*New technologies and methods in psychology*] (pp. 449–470). Liberi Libri. https://doi.org/10.47943/lib.9788363487430.rozdzial21

Trusz, S. (Ed.). (2013). *Efekty oczekiwań interpersonalnych. Wybór tekstów* [*Interpersonal expectation effect. A reader*]. Wydawnictwo Naukowe Scholar.

Tukey, J. B. (1977). *Exploratory data analysis*. Addison-Wesley.

Webb, M. A., & Tangney, J. P. (2022). Too good to be true: Bots and bad data from Mechanical Turk. *Perspectives on Psychological Science*, 1–4. https://csl.mpg.de/427800/webb_tangney__too_good_to_be_true_2022.pdf; https://doi.org/10.1177/17456916221120027

Wilkinson, L. & Task Force on Statistical Inference American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.https://doi.org/10.1037/0003-066X.54.8.594

Winer, B. J. (1962/1971). *Statistical principles in experimental design*. McGraw-Hill.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). McGraw-Hill.

Wolski, P. (2016a). Istotność statystyczna I. Nieodrobiona lekcja [Statistical significance I. A lesson not learned]. *Rocznik Kognitywistyczny* [*Yearbook of Cognitive Science*], *9*, 27–35. https://doi.org/10.4467/20843895RK.16.003.5471

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

Wolski, P. (2016c). Istotność statystyczna III. Od rytuału do myślenia statystycznego [Statistical significance III. From ritual to statistical thinking]. *Rocznik Kognitywistyczny* [*Yearbook of Cognitive Science*], *9*, 71–85. https://doi.org/10.4467/20843895 RK.16.007.6413