ROCZNIKI PSYCHOLOGICZNE / ANNALS OF PSYCHOLOGY
2015, XVIII, 1, 25-40

ŁUKASZ BUDZICZ
Adam Mickiewicz University in Poznań
Department of Psychology

POST-STAPELIAN PSYCHOLOGY. DISCUSSIONS ON THE RELIABILITY OF DATA AND PUBLICATIONS IN PSYCHOLOGY

In 2011, Diederik Stapel's fraud was discovered. It turned out that not only did Stapel forge data but also journals failed to notice many obvious errors and encouraged distortions (e.g., not reporting studies with non-significant results). Simultaneously, Simmons et al. (2011) published an article dedicated to questionable research practices that could significantly increase the number of false-positive results through arbitrary decisions pertaining to data analysis and presentation. Shortly after, there appeared results of studies suggesting that a large number of researchers confess to such practices and that they are, in fact, commonly accepted. These events sparked off a wide debate about the reliability of data in psychology. The author of the present paper discusses the most important points of this debate, showing how the low level of theoretical maturity, the lack of consensus on the rules of applying research techniques and interpreting results, and the unrealistic demands of editors of empirical journals may have contributed to this crisis.

Keywords: fraud in psychology, Diederik Stapel, "false-positive psychology," publication bias.

STAPEL AND THE CRISIS IN PSYCHOLOGY IN 2014

Although announcing a "crisis" in psychology is a regular phenomenon, which has been with us for a long time (Asch, 1952/1987; Ring, 1967; Elms,

Corresponding author: ŁUKASZ BUDZICZ – Department of Psychology, Adam Mickiewicz University in Poznań, ul. Szamarzewskiego 89, 60-568 Poznań; e-mail: lukasz.budzicz@gmail.com

Acknowledgments. The author wishes to thank Jerzy M. Brzezinski and anonymous reviewers for their valuable remarks on the earlier versions of the paper.

1975; Bevan, 1991; Staats, 1999; Rozin, 2001; Rand & Ilardi, 2005), it seems that the last few years have been special in this respect. In a relatively short period of time, there appeared a number of papers questioning the reliability of typical data in psychology (the entire issue of *Perspectives on Psychological Science 7*(6), 2012; Asendorf et al., 2013; Ferguson, 2013; John, Lowenstein & Prelec, 2012; Kepes & McDaniel, 2013; LeBel & Peters, 2011; Masicambo & Lalande, 2012; Mitchell, 2012; Murayama, Pekrun, & Fiedler, 2014; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, 2013; Simonsohn, Nelson, & Simmons, 2014; Tressoldi, 2012; Wicherts, Bakker, & Molenaar, 2011). Meaningfully, the special issue of *Perspectives on Psychological Science* was entitled *crisis of confidence* (Pahsler & Wagenmakers, 2012), which can be understood as either "crisis of reliability" or "crisis of trust."

The most important cause of this "escalation" was the discovery of Stapel's fraud, which had lasted for many years (a majority of the articles mentioned above refer to it). The story was discussed extensively by specialists (but not only), and there is no need here to recall its details, which have been widely presented elsewhere (Bhattacharjee, 2013; Levelt Committee, Noort Committee, & Drenth Committee, 2012; further: Levelt, 2012). It seems sufficient to mention that after an investigation conducted by a special committee, almost 60 papers published within the previous 15 years in the most renowned psychological journals were withdrawn.

The significance of Stapel's story does not consist in the fact that a fraud was detected in the field of psychology. If frauds are found in other disciplines, it is hard to expect that psychology will be any different in this respect. The problem lies in the fact that dozens of texts reporting more than two hundred of studies were challenged neither at the stage of peer-review nor – most importantly – by the mechanism of intersubjective verifiability (i.e., independent replications). This significantly distinguishes the Stapel affair from other scientific frauds, especially in the field of natural sciences (see Stroebe, Postmes, & Spears, 2012), where detecting a fraud was usually possible because other researchers were unable to replicate the experiments. Physicists witnessed an affair of a comparable caliber. Jan Hendrik Schön, at some point a young star of material physics, would publish one article every two weeks in journals such as *Science* or *Nature*. No laboratory was able to obtain materials of a quality similar to that allegedly produced by him, which is why experts started to take a closer look at his studies. From the moment when Schön's first publications appeared, experts needed approximately two years to detect systematic anomalies in the data and to question the reliability of his work (Reich, 2009). By contrast, before Stapel's fraud

came to light, no constructive criticism of his findings had been formulated and there had not appeared, to my best knowledge, any unsuccessful replication of his studies. Given the scale of his fraud, Stapel himself got caught because of very prosaic reasons (e.g., he bragged about the results although he was unable to provide raw data; Bhattacharjee, 2013).

The fraud itself was far from perfect. Analyses of statistical data conducted post factum revealed a number of irregularities. Experts would point to such unlikely situations as identical data in independent experiments (e.g., means and standard deviations), absence of missing data, very low *F* statistics wherever no significant effects were expected, and too strong effects for scales with low reliability or for scales with one item. When raw data were analyzed, investigators would find mechanically inserted columns of variables. As a matter of fact, every questioned article contained some kind of irregularities (Levelt, 2012, pp. 69-100). Fabricating data is not, as one might expect, an easy procedure. The fact that the best psychological journals would systematically approve such data means that reviewers either lacked adequate competence (above all, statistical competence) or prepared sloppy reviews. In a standard announcement about job opportunities for reviewers, APA informs that "reviewing a manuscript takes time (1-4 hours per manuscript reviewed)" (these announcements can be found in the PsycARTICLES database, after typing in "reviewers wanted"). Assuming even the best case scenario of four hours, it is hard not to ask the following rhetorical question: is four hours really enough to read carefully an article that includes three to six studies on average, analyze its theoretical basis, check the cited sources, reflect on the rationality of the applied procedure and tools, not to mention controlling raw data or verifying the correctness of calculations? I would postulate that this amount of time can be sufficient, at best, for checking the basic editorial standards, pointing out obvious errors, and presenting one's own perspective in the review.

The committee set up to investigate the affair did not blame Stapel exclusively. Editors and reviewers of journals not only accepted Stapel's successful studies eagerly and (almost) without any reservations, but also encouraged him and his collaborators to persist in misconduct. The report reads: "a co-author [of Stapel's] stated that editors and reviewers would sometimes request certain variables to be omitted, because doing so would be more consistent with the reasoning and flow of the narrative, thereby also omitting unwelcome results. Reviewers have also requested that not all executed analyses be reported, for example by simply leaving unmentioned any conditions for which no effects had been found, although effects were originally expected. Sometimes reviewers insisted on retros-

pective pilot studies, which were then reported as having been performed in advance. In this way the experiments and choices of items are justified with the benefit of hindsight" (Levelt, 2012, p. 53).

Stapel showed that an intelligent person is able to publish and make a career using "smooth" results supported by an attractive narrative much easier than using reliably collected data, which would probably be much less unequivocal and impressive. Nevertheless, Stapel's fraud may point to a far more fundamental problem, namely, to the fact that a great majority of findings in psychology are only loosely connected to one another in respect of their theoretical dimension and that what takes place is the "gathering" of facts rather than building a well-integrated knowledge system. It is hard to place new findings within a broader theoretical perspective because either there is no such perspective or (probably more often) it is too general to clearly state what the results of particular studies should look like. In one of his fraudulent studies, Stapel "showed" that in more chaotic environments (e.g., ones contaminated with rubbish) people discriminate against minorities to a greater extent (e.g., sit further away from such people on a public bench; Stapel & Lindenberg, 2011). This and other hypotheses advanced by Stapel do not appear to be exceptional in any way. They are similar to a great number of studies present in the literature. They do not describe any kind of groundbreaking or sensational findings, especially such that would radically contradict the existing knowledge (such as those presented by Bem, 2011, in his article on precognition). It is also hard to point to any psychological (meta) theories to which they would stand in contrast. Yet, it is possible (and Stapel did that) to equip them with theoretical explanations that sound reasonable and to quote many sources that describe similar effects. The effects, although convincing, are not particularly strong (Stapel himself took care not to show too strong effects because he knew that they would have been unconvincing to the editors; see Bhattacharjee, 2013).

A fraud on such a scale is something extremely rare, even in natural sciences (see Stroebe et al., 2012). It is even the more unusual in psychology, since frauds in this field are hardly ever discovered. The fundamental question is whether Stapel was only a "black sheep" (as has been suggested by the European Association of Social Psychology, 2012) or a telling example of a more common phenomenon in psychology. There is no conclusive answer to this question; nevertheless, Stapel's case shows that we should be concerned about the ability of the scientific environment to detect frauds. Furthermore, frauds consisting in data fabrication are not the only problem that may significantly distort the picture of scientific reality. In discussions on the condition of psychology, there appear

voices that the percentage of researchers who falsify reality in a more subtle way is, in fact, considerable.

MORE SUBTLE FALSIFICATIONS: FALSE-POSITIVE PSYCHOLOGY

Roughly at the time when Stapel's affair broke out, there appeared an article by Simmons, Nelson, and Simmonsohn (2011) about "false-positive psychology" (a funny wordplay resulting from the combination of "false-positive error" and "positive psychology"). The text became very popular (about 800 citations in Google Scholar; data from April 2015), and I would speculate that the Stapel affair contributed to its popularity significantly. Basically, the article does not discuss new ideas (e.g., Maxwell, 2004; Ioannidis, 2005); nevertheless, it has been standardly quoted in the course of new discussions pertaining to the "crisis."

The article consists of two parts: "empirical" and mathematical. In the "empirical" part, the authors presented a report from a real study. Reporting data selectively, the authors "showed" the existence of absurd effects, for example, that listening to the song *When I'm 64* by The Beatles *decreased* the age of the investigated individuals. In the mathematical part, the authors generated randomly selected distributions of data that were supposed to simulate real studies. Four practices of "cranking up" data were applied in order to check how they would increase the probability of a significant result. The four practices were:

1. Using many dependent variables and reporting the one that came out significant.
2. Adding additional groups of investigated individuals until statistical significance has been reached (and, of course, abandoning further investigation when significance has been reached).
3. Including a divalent variable as an additional covariable (e.g., sex).
4. Conducting experiments on more than two experimental groups and selectively reporting only those in which significant differences have been observed.

Using one of the above practices practically doubles the probability of making the false-positive error from the "traditional" 5%; however, using all of them increases the probability of finding a significant effect in totally random data to 61%.

The fact that it is theoretically possible to use such tricks does not tell us much about the frequency of this phenomenon. Therefore, John, Loewenstein,

and Prelec (2012) made an attempt at providing an empirical assessment of the frequency of the questionable research practices mentioned above. An invitation to take part in the study was sent to six thousand psychologists, research workers of American universities. Anonymity was guaranteed to the participants, and they were motivated to tell the truth (they could designate a non-profit organization that was to be supported financially by the authors). Eventually, a little more than 2,100 individuals took part in the study. The researchers asked the participants whether they had ever engaged in any of the questionable practices listed in Table 1. Additionally, the respondents were asked to estimate how common these practices were among other researchers and how many of them would actually confess to using them; juxtaposing these data, the authors assessed the prevalence of the practices. The respondents were also asked to evaluate the acceptability of particular practices on a 3-point scale. The results are presented in Table 1.

In the case of particular categories, over 50% of the respondents confess to practices that lead to a very high number of false-positive results. It has been estimated that the percentage of researchers who engage in the practices analyzed by Simmons et al. (2011) – for example, not reporting dependent variables, adding participants, or selecting only "working" comparisons between research groups – is, respectively, 78%, 72%, and 42%. One of the most pessimistic conclusions is that none of the questionable research practices (except for falsifying data) is perceived as wrong: they are deemed "acceptable" (much less frequently "rather acceptable"). In future studies, it would be interesting to verify whether this stems from a low level of methodological awareness or rather from cynical pragmatism. It can be also legitimately assumed that a certain percentage of fraudulent researchers will not admit to resorting to such practices and/or will not take part in a similar study (in the reported study, over 60% of the invited researchers decided not to participate).

The results presented by John and colleagues indicate that "false-positive psychology," against which Simmons et al. (2011) warned, may actually be a fact rather than merely a mathematical curiosity. Going one step further, Bakker, van Dijk, and Wicherts (2012) simulated studies on the effects of complex strength and grouped them into "meta-analyses." It turned out that the best strategy to obtain publishable (statistically significant) results was to conduct many small studies with the use of the questionable research practices. This strategy was also found to distort reality to the greatest extent. Given that the standard sample size in a psychological study is approximately 40 persons (Tressoldi, 2012; Marszalek et al., 2011), it seems that such a strategy is applied quite frequently (see also Francis, 2014).

Table 1

Questionable research practice (QRP)	The percentage of researches who confessed to particular QRPs (self-admission rate)	Prevalence of particular QRPs as assessed by subjects	Prevalence estimate derived from admission estimate (see John et al., 2012, for details)	The defensibility rating of particular QRPs (*SD* in parentheses)
1. Failing to report all of a study's dependent variables in a paper	66.5%	60%	78%	1.84 (0.39)
2. Deciding whether to collect more data after looking to see whether the results were significant	58.0%	62%	72%	1.79 (0.44)
3. Failing to report all of a study's experimental conditions in a paper	27.4%	38%	42%	1.77 (0.49)
4. Stopping the collection of data earlier than planned because one has found the result that one has been looking for	22.5%	41%	36%	1.76 (0.48)
5. "Rounding off" the *p*-value in a paper (e.g., reporting that a *p*-value of .054 is less than .05)	23.3%	41%	39%	1.68 (0.57)
6. Selectively reporting only those studies that "worked"	50.0%	61%	67%	1.66 (0.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results	43.4%	45%	62%	1.61 (0.59)
8. Reporting an unexpected finding as having been predicted from the start	27.0%	50%	54%	1.50 (0.60)
9. Claiming that the results are unaffected by demographic variables (e.g., gender) when one is actually unsure if they are (or knows that they are not)	4.5%	2%	13%	1.32 (0.60)
10. Falsifying (fabricating) data	1.7%	10%	9%	0.16 (0.38)

Source: data provided by John et al. (2012). *Notes.* In column 1, the results of the group that was additionally motivated to tell the truth are presented. The results of the control group were, on average, a little lower (i.e., 0-7% lower). The values in column 2 were estimated on the basis of the height of the chart bar (no precise raw data are given in the article). The percentages in column 3 represent the prevalence of a given research practice assessed by the authors of the study on the basis of the percentage of respondents who confessed to using questionable practices and the percentage of people who the participants believed would be willing to confess to using such practices. The scale applied in column 4: 0 – *completely unacceptable*; 1 – *to some degree acceptable*; 2 – *acceptable*.

In psychology, the complexity of research topics, the low level of consensus on the way of measuring psychological variables, and the fluidity of theoretical assumptions make it easier for researchers to make arbitrary decisions on data

analysis and presentation. In order to better illustrate the problem, an example of a hypothetical study, taken from Gelman and Loken (2013), will be applied and creatively developed here. The study pertains to differences in solving mathematical problems between the supporters of Democrats and Republicans in the USA. The hypothetical researcher assumes that differences occur depending on the context in which these problems are embedded: Democrats will be better able to cope with a problem pertaining to healthcare, and Republicans with a mathematically analogous problem embedded in the military context. The researcher also collects a number of additional demographic data. Let us assume that this effect occurs only in men. This can be easily explained because men have stronger ideological beliefs (here the hypothetical researcher can give many sources). And what if differences turn out to pertain only to women? The researcher will generate a copious explanation pointing out that women are more susceptible to the context (and he will support this claim with a number of sources). Moving further, we know that the researcher asked the respondents about their political preferences using a 7-point scale. There arises another question: whom is he to compare? The persons who marked 1-3 with the persons who marked 5-7? Maybe it would be better to exclude 3 and 5 as cases close to neutrality? Or maybe it would be better to ideologically compare "ones" and "sevens"? What to do with those neutral? Exclude those instances? And what if crucial analyses reveal significant differences between those politically engaged and those neutral? Maybe none of the analyses will yield a significant result, but if age is added to the model, then perhaps a significant interaction will occur, an interaction for which the researcher can develop a further story (e.g., younger individuals are more radical in their political views vs. young individuals have less crystalized and more fluid political views). In the meantime, our hypothetical researcher can make an infinite number of other arbitrary decisions if he thinks that they might bring significant effects.

On any large collection of raw data one can conduct a great number of analyses, some of which will yield significant results just by accident. Additionally, it is possible to come up with a better or worse justification (or, bluntly speaking, a "story") for every result and, from countless publications, select a source that supports it. Stapel's articles contain dozens of citations supporting his research hypotheses, although we now know that his studies were fraudulent. After obtaining results of a given kind, the researcher can present them as predicted from the very beginning by his own "theories," and thus give them alleged but in fact illegitimate effectiveness in generating empirically verified predictions. Unfortunately, this in not an uncommon practice (Kerr, 1998).

Gelman and Loken (2013) analyze several instances of real studies, pointing to an arbitrariness of some decisions that is difficult to explain. I will discuss here a different study that appears to me to be questionable for the same reason. Gervais and Norenzayan (2012) presented studies that indicate that analytical thinking is a predictor of lack of religious faith. Manipulations in four experiments conducted by them consisted, for example, in having the participants watch a sculpture of a thinking philosopher (or, in the control group – a sculpture of an athlete); priming them with words either neutral or associated with analytical thinking; forcing them to read words printed in distinct vs. blurred type, etc. In each of the studies, the authors obtained results suggesting a relationship between "analytical thinking" (more precisely: manipulation allegedly inducing such thinking) and the lack of faith. What is questionable is that in particular experiments the authors used different dependent variables. In the first experiment, the participants assessed their level of faith in God on a 0-100 scale; in the second experiment, they were asked to answer questions about God, angels, and devil on a 1-7 scale; in another experiment, they were asked to respond (on a 1-7 scale) to 10 assertions pertaining to their religiousness (e.g., "My religious beliefs are the very thing that underlies my life philosophy"). Nevertheless, even in this inconsistency the researchers were *inconsistent* because in the last study they returned to "faith in God assessed on a 0-100 scale" as the dependent variable. The authors did not provide any explanation of why they had used different dependent variables in each of the studies. While different experimental procedures can be somehow understood, different dependent variables make it difficult to offer a reliable interpretation of the obtained results, because one cannot be sure whether particular dependent variables measure exactly the same. I suppose, intuitively, that the applied measures probably measure something very similar, but their arbitrary application may indicate that each time several dependent variables were used, and that the researchers reported only those in the case of which significant relationships were observed (or that there were different combinations of the manipulation condition and the dependent variable, and only those that worked were reported). What points to a selective reporting of research results is the very high negative correlation between effect size and sample size. In the four experimental studies discussed, this correlation was $r = -.97$. Looking at this from the mathematical perspective, the size of the sample is unrelated to mean effect size. There is, however, a relationship between the size of the sample and the variance in effects (see Lippa, 2009), that is, in the case of smaller samples effects will be more diversified. Quite naturally, with smaller samples, only stronger effects will turn out to be significant. Such a high correlation between

effect size and sample size may be accidental, but it may also indicate that there were additional studies which the authors did not mention (see especially Francis, 2012). In a similar vein, Fergusson (2013) showed how scientists who investigate the influence of computer games on aggression, using a procedure of shocking others with white noise, drew out from it a number of different indicators, probably depending on which of them exhibited significant patterns (e.g., the number of sessions in which the participants applied the loudest noises, square root of noise length multiplied by intensity, sums of intensity, or means of intensity and length separately, etc.).

Both fabrication and selective presentation of research results ("false-positive psychology") are types of unreliability, in my opinion, not very different from each other in respect of their practical consequences. In both cases, the scientific community is presented with a distorted picture of reality, and receives information only about successful studies, which falsely suggests ease in obtaining the results; scholars are unable to calculate the real effect size and nobody can be really sure whether the unpublished data did not contain information on some important moderators diminishing the effect. Let me formulate the following hypothesis here: while both researchers and editors of scientific journals are aware of the wickedness of falsification, selective publishing is a strategy that stems partially from the lack of knowledge of statistics (see acceptance in studies by John et al., 2012) and from an unwritten agreement between editors and researchers that the attractiveness of the manuscript is a condition practically as important as its reliability. Therefore, only data that support hypotheses will be published, which is, in fact, something we can observe in empirical journals on a regular basis. According to the bibliometric analysis conducted by Fanelli (2010), psychology and psychiatry are branches of science with the greatest number of articles in which initial hypotheses have been positively verified. I would not expect that psychologists have at their disposal theories so highly refined and investigation tools so precise as to predict reality better than representatives of natural sciences.

Lately, there have also appeared a number of other signals indicating that the reliability of data in psychology is far from perfect. Particularly alarming are the analyses pointing to: the low power of psychological studies (Bakker et al., 2012; Francis, 2012, 2014; Tressoldi, 2012), which, combined with the small number of publications that report negative effects (Fanelli, 2010), suggests that there is a strong publication bias; the distribution of p values in the literature, especially the improbable increase in the frequency of values just below .05 (Massicampo & Lalande, 2012; Leggett, Thomas, Loetscher, & Nicholls, 2013; Simonsohn,

et. al., 2014); errors in reporting the *p* value, especially by classifying non-significant values as significant (Bakker & Wicherts, 2011; Wicherts et al., 2011); the small percentage of replications in the total of studies (Makel, Plucker, & Hegarty 2012); and not sharing raw data upon request (Mitchell, 2012; Wicherts, Borsboom, Kats, & Molenaar, 2006).

FINAL REMARKS

The "crisis" has resulted in a wide discussion on the necessity of introducing changes into research and publishing practices in psychology. Some suggestions are of a more technical character, others demand a fundamental change of approach to practicing science. The former include a proposal of obligatory publication of raw data (Wicherts & Bakker, 2012; Simmonsohn, 2013), which could facilitate detecting unusual patterns of data (as I have already mentioned, in Stapel's texts a great number of errors were detected). Raw data analysis enabled Uri Simmonsohn (2013) to detect two other cases of data fabrication, conducted by two social psychologists (specifically, by Lawrence Sanna and Dirk Smeesters). In yet another recent case, statisticians pointed to the extremely low probability of data obtained by Jens Förster (Kolfshooten, 2014). Easy access to raw data can be conducive to detecting frauds, although it will not prevent them completely. Another proposal pertains to compulsory preregistration of studies in order to limit the possibility of "creative" data handling (Aveyard et al., 2013). This should not be difficult to put into practice because studies usually have to be reported to ethical committees beforehand; therefore, conducting additional registration on websites dedicated to this purpose does not seem to require great effort. Another idea to limit the phenomenon of "false-positive psychology" is to make it obligatory for authors to declare that they have reported all variables, conditions, manners of establishing the sample size, and deleted cases (Simmons et. al., 2012).

Finally, there are proposals of more fundamental reforms. Journals are mainly open to positive results, whereas even the best studies with negative results do not stand a chance of being published. If the study does not yield the expected results (which is very common in science), researchers can either choose to waste several months of hard work (and, consequently, have smaller chances getting promoted, etc.) or decide to distort the data to a greater or smaller extent (or, possibly, if they have conducted several experiments of which only some yielded the expected results, they can omit the unsuccessful ones to increase the

attractiveness of the narration). A change of the incentive system therefore appears to be crucial for promoting good studies, irrespective of their results (Nosek & Bar-Anan, 2012). An interesting solution has been suggested by the *Cortex* journal, which declared opening a section for preregistered studies (Chambers, 2013). Authors are supposed to send articles with only the initial section and a description of the research method in them. The reviewers are to evaluate whether the formulated research question is important and the study reasonably designed. When the decision is positive, the author starts to conduct the study, and the journal takes upon itself the obligation to publish it, irrespective of its results. In this way, the authors can potentially diminish the risk of devoting time and resources to studies that do not stand a chance of being published. If the data has already been collected, it has been suggested that the reviewers might evaluate the texts without knowing their results (the so-called *result blind review*; Greve, Bröder, & Erdfelder, 2013). Possible discrepancies could then be a hint for the reviewers that a good study loses because of its negative results (or that a questionable study gains because of its positive results). The final criterion of evaluating data in science is, naturally, the possibility of their independent confirmation, which is why a considerable proportion of voices in the dispute on the reliability of data in psychology pertains to conducting precise replications (Asendorpf, et al., 2013; Nosek, Spies, & Motyl, 2012; Francis, 2012).

The realization of the proposals described above does not require extensive financial resources. Nevertheless, it does require breaking certain customs and habits present in the discipline for a very long time. A majority of the most important players have not made the decision to radically change the rules of the game yet. Journals such as *Journal of Personality and Social Psychology* or *Journal of Experimental Psychology*, which suffered the most because of Stapel's fraud, should be the party most interested in introducing the changes. So many irregularities have been detected recently that it seems that, sooner or later, research and publishing practices will have to change.

In the end, I will venture to formulate the following thesis: the current crisis stems form the ill-conceived pursuit of "innovativeness." It is this pursuit that causes jumping from one hypothesis to another, without thoroughly understanding a particular effect or obtaining maximally reliable knowledge of it (to the extent to which this is possible in such a field of science as psychology). The best journals require "groundbreaking" studies, and similar requirements are formulated by institutions that award grants (here is an example from the Polish playground – on the website of the National Science Center, the following in-

formation about the Maestro program can be found: "it is a contest for financing research projects that aim at realizing pioneering scientific studies, . . . which go beyond the current state of knowledge"). Therefore, even a large and the most carefully prepared study that aims at controlling the correctness of another study does not stand a chance of receiving financial support. The signal is clear: a prominent researcher does not engage in conducting replications of other studies; he or she can, at best, repeat the study modifying it creatively (so-called conceptual replication). Still, a conceptual replication does not tell us much about the original effect; in particular, we are unable to tell whether the inability to replicate the study stems from the fact that the reported effect simply does not exist, or whether it is a result of the modifications introduced. In the end, findings cumulate, researchers do not have infallible criteria of differentiating between real and falsified results – and, if only due to logistic matters, replicating all studies will never be possible (Makel et al., 2012). A certain part of misrepresentations that result from researchers' arbitrary decisions, falsifications, or unconscious errors will never be corrected. Being neither chemists nor physicists, we are unable to obtain results that are both easy to replicate and relatively straightforward in their theoretical dimension (not to mention that representatives of natural sciences are sometimes also unable to obtain such results). The assumption that in a science such as psychology several consecutive studies should always point to the same effect is unrealistic (particularly if the effect is weak, which also contradicts the elementary probability calculation; Francis, 2012, 2014). Comparably unrealistic is the assumption that scientific journals will be filled with "discoveries" from the first to the last page. Personally, I value articles that describe new ideas, increase our understanding of reality, and break old schemata. Nevertheless, a science such as psychology must be based to an equal extent on arduous though not very creative work, consisting in verifying the truth of knowledge developed by other people. I do not think that psychology would suffer if, alongside innovative, exploratory, and relatively small studies, journals also promoted texts reporting reproductive studies – studies that would be very strong and maximally reliable.

REFERENCES

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence, 149*(1), 91-130.

Asch, S. E. (1952/1987). *Social psychology*. New York: Oxford University Press.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology: Recommendations for increasing replicability. *European Journal of Personality, 27*(2), 108-119.

Aveyard, P., Bellgrove, M., Bertamini, M., Bestmann, S., Bishop, D., Brembs, B., . . . Wolfe, J. (2013). *Trust in science would be improved by study pre-registration.* Retrieved from http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666-678.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

Bevan, W. (1991). Contemporary psychology: A tour inside the onion. *American Psychologist, 46*, 475-483.

Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times*. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-raud.html?_r=0

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

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist, 30*(10), 967-976.

European Association of Social Psychology (2012). *Statement EASP on Levelt December 2012.* Retrieved from: http://www.easp.eu/news/Statement%20EASP%20on%20Levelt_December_%202012.pdf

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*, e10068.

Ferguson, C. J. (2013). Violent video games and the Supreme Court: Lessons for the scientific community in the wake of Brown v. Entertainment Merchants Association. *American Psychologist, 68*(2), 57-74.

Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163-171.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7*(6), 585-594.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21*, 1180-87.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Retrieved from: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336*(6080), 493-496.

Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist, 18*(4), 286-294.

Ioanis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology, 6*(3), 252-268.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kolfshooten, F. (2014). Fresh misconduct charges hit Dutch social psychology. *Science, 344*, 566-567.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*(4), 371-379.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. R. (2013). The life of *p*: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology, 66*(12), 2303-2309.

Levelt Committee, Noort Committee, & Drenth Committee (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel.* Retrieved from https://www.commissielevelt.nl/wp-content/uploads_per_blog/commissielevelt/2013/01/finalreportLevelt1.pdf

Lippa, R. A. (2009). Sex differences in sex drive, sociosexuality, and height across 53 nations: Testing evolutionary and social structural theories. *Archives of Sexual Behavior, 38*(5), 631-651.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills, 112*, 331-348.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163.

Mitchell, G. (2012). What is wrong with social psychology? *Dialogue. The Official Newsletter of the Society for Personality and Social Psychology, 26*(2), 12-17.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*, 107-118.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*, 217-243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631.

Pashlev, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Ring, K. (1967). Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology, 3*, 113-123.

Rand, K., & Ilardi, S. (2005). Toward a consilient science of psychology. *Journal of Clinical Psychology, 61*, 7-20.

Reich, E. S. (2009). *Plastic fantastic: How the biggest fraud in physics shook the scientific world*. Palgrave Macmillan.

Rozin, P. (2001). Social psychology and science: Some lessons from Salomon Asch. *Personality and Social Psychology Review, 5*, 2-14.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21-word solution. *Dialogue. The Official Newsletter of the Society for Personality and Social Psychology, 26*, 4-7.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875-1888.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Staats, A. (1999). Unifying psychology requires new infrastructure, theory, method, and a research agenda. *Review of General Psychology, 3*, 3-13.

Stapel, D. A., & Lindenberg, S. (2011). Coping with chaos: How disordered contexts promote stereotyping and discrimination. *Science, 332*, 251-252.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*(6), 670-688.

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology, 3*(218). doi: 10.3389/fpsyg.2012.00218.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence, 40*(2), 73-76.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE, 6*(11), e26828.

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