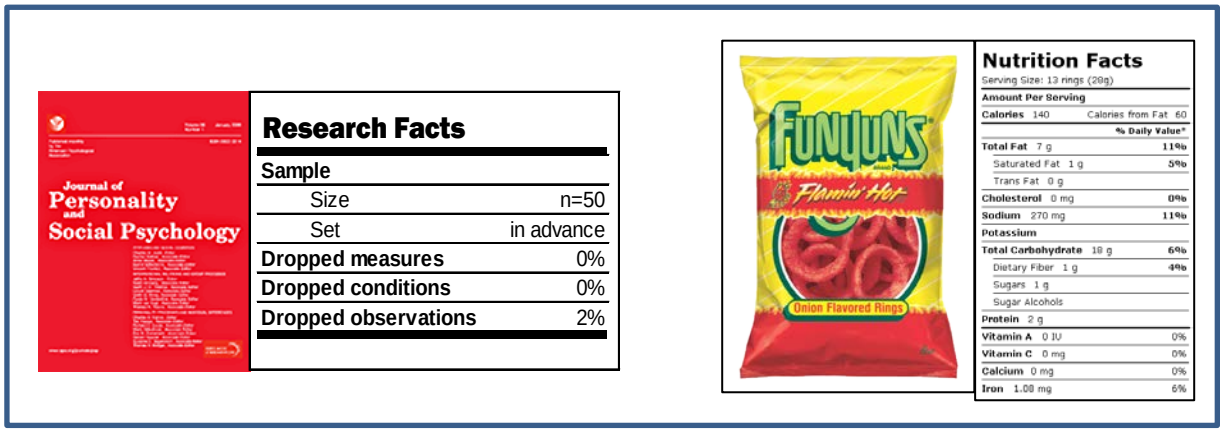
Not unlike objections that it is easier to comply with disclosure for researchers in some fields than in others, sardines importers complained that labeling the exact type of oil in their cans was harder on them than on domestic producers, etc.[2]

Fortunately for all of us today, the Federal Government knew better than to let the perfect be enemy of the good. If only perfect solutions were implemented, we would still be drafting the Pure Food Act of 1906.

We hope that editors will emulate the pragmatic politicians of the 1900s, deciding to implement disclosure requirements in our journals before a perfect solution with no detractors is arrived at. In the meantime, those of us who realize transparency is a necessary condition for evidence to be scientific can start adding 21 words to our papers.

Figure 3. One of these labels is not mandatory



What Can We Do to Reduce Scientific Misconduct?

Wolfgang Stroebe (Utrecht University and University of Groningen), Tom Postmes (University of Groningen) and Russell Spears (University of Groningen)

The news that the highly respected social psychologist Diederik Stapel had committed large-scale scientific fraud came as a wake-up call to our scientific community. Even though there had been major fraud cases before in physics (e.g., Jan Hendrik Schön, 2002) and in medicine (e.g., Darsee, 1981; Wakefield, 2004; Woo Suk Hwang, 2006), most of us had never considered that such major fraud would happen in our midst. After all, the fraud of Karen Ruggiero (2001) was comparatively minor, resulting in only two retractions of scientific articles, and Marc Hauser (2002) was a biologist by training and thus not really a proper psychologist, even if he did hold a position in psychology (for information, see Stroebe, Postmes & Spears, in press). Finally, the case of Sir Cyril Burt and his invented twin data happened long before most of us were born.

The international press had a field day and lambasted psychology, suggesting that the Stapel case "exposes deep flaws in the way science is done in [..] psychology" (Carey, 2011). Journalists also wondered why we did not discover the fraud earlier (e.g., Campbell, 2011) and frankly, we asked ourselves the same question (Stroebe et al., in press). We had always assumed that science is self-correcting in that findings that are based on fraudulent research will be discovered either in the peer review process or through (failed) replications (Broad & Wade, 1982; Goodstein, 2012). As Crocker and Cooper (2011) recently asked: "Scientists generally trust that fabrication will be

[2] "Grocers complain of Wiley," *New York Times*, November 20th, 1904, pp.7

uncovered when other scientists cannot replicate (and therefore fail to validate) findings. In this particular case, however, reliance on replication did not work. Why?"

In order to see how other frauds were identified, we began to study reports of fraud cases. There is abundant material (e.g., Case summaries from the Annual Reports of the Office of Research Integrity http://ori.hhs.gov/case_summary), but official reports typically do not disclose how the fraud had been discovered. For this information we had to rely on newspaper reports. This limited our sample to 40 of the more spectacular cases. As we reported in an article to appear in "Perspectives on Psychological Science", we found to our surprise that hardly any of these cases had been discovered during the review process or as a result of failed replications (Stroebe et al., in press).

In retrospect, one can think of good reasons for this. Reviewers evaluate a manuscript in terms of whether hypotheses are clearly derived from theory, whether the research is sound, whether alternative explanations are ruled out, etc. Because they have to rely on information provided by the authors, fraud may be difficult to detect. However, our research also shows that in several of the cases we studied, the fraudsters were sloppy and left clear signs of wrongdoing that could (perhaps should) have raised reviewers' suspicion (Stroebe et al., in press). This suggests that in general, reviewers are not always sufficiently alert to the possibility of fraud.

With regard to replications, a frequently deplored problem is that journals typically do not accept replications for publication. As a result, they are rarely done and if they are done, they do not become known to the scientific community. However, as we will discuss below, there are also other reasons why replications are not always very effective means for the identification of fraudulent research.

Incentives for Fraud

We all know that the way to achieve success in our discipline is to publish in high impact journals. And to get accepted in these journals, one not only has to develop hypotheses that are novel and interesting, but ideally the predictions should be supported unambiguously by the data. Obviously, researchers who fabricate or falsify their data have an advantage here. It is therefore not surprising that journal impact measures correlate positively with number of retracted articles (Fang & Casadevall, 2011). However, this system not only rewards fraudsters, it also rewards deserving researchers. Furthermore, the system is shaped by market forces. As long as there are more good researchers than there are jobs, departments will be selective and as long as there are more manuscripts than there is publication space, editors will be too. Thus, we reasoned that there was little likelihood for this system to change.

Strategies For Fraud Reduction

Instead of trying to change the system, we decided to focus on making improvements to the process of conducting research. In this section, we discuss five potential strategies of fraud reduction. Some of these will be quite familiar (i.e., courses on research ethics, increasing the accessibility of data sets, facilitating replications) but others are less so (i.e., strengthening the position of whistleblowers, instituting research audits).

Research ethics. Many have pointed out that the field needs clear standards and clear procedures to deal with suspicions about research or researchers. Indeed, our research suggests that universities have often been reluctant to investigate fraud cases or, if they investigated them, to make their findings public. This hinders us in combatting this problem. Institutions may also need to devote more attention to research ethics. A course on research ethics should therefore be part of every graduate program (and although Stapel himself also taught such a course, this is no reason not to). In such courses one should discuss the obvious rules and good practices (e.g., if data of participants are eliminated, accepted rules have to be followed and this has to be reported in the article), as well as the grayer areas of research practice, such as failure to report null findings.

But we believe that the development of a macro-level infrastructure for dealing with fraud (rules and procedures) should be complemented by a consistent micro-level commitment to maintaining the highest standards of research integrity in our everyday research practices. Research on fraud points to the strong influence that the immediate social environment's norms and practices have on one's ethical conduct. Accordingly, the micro-level maintenance of ethical standards by the local research group should be the most impactful way of guaranteeing that standards are upheld. Our research confirms that fraud is most often flagged up by insiders, aware that

something is amiss. Extending this, the local level is also the best place to ensure that ethical research practices are promoted. In some sense, this is also a heartening conclusion: we can take matters into our own hands.

***Strengthening the position of whistleblowers*.** As in the Stapel fraud case, reports by whistleblowers are by far the most frequent way in which fraud is discovered. These whistleblowers are mostly research collaborators, who have inside knowledge of the research practice in their laboratory. They are often PhD students or postdoctoral researchers. It is therefore important that graduate students are not only taught proper research practices but are also informed that fraud does happen and what should be done when it is suspected. There should also be assigned people of trust at each department to whom people can turn in cases of suspicion. And there should be clear protocols stating how such discussions should be handled. Furthermore, the anonymity of whistleblowers must be safeguarded whenever possible. Being a whistleblower often has negative effects on people's careers–we need to ensure that the reputation and careers of whistleblowers do not suffer, both for individual whistleblowers and for the institutions who decide to self-investigate.

***Increasing the accessibility of data sets*.** Even though the APA rules clearly specify that authors should share research data with others on request (e.g. American Psychological Association, 2010), authors are often reluctant to do so. For example, Wicherts, Bakker and Molenaar (2011), who contacted the corresponding authors of 141 articles published in psychology journals, reported that most authors failed to send their data. One of the most widely accepted strategies of fraud detection is the creation of a publicly accessible repository of the data of published studies, which would at least discourage the most egregious example of fraud based on obviously dubious or even nonexistent data. However, this still leaves the possibility that researchers could "massage" their data (e.g., by omitting participants, who did not respond in line with hypotheses). At least for studies conducted by computer, this could be prevented, if all research institutions stored data of studies in read only files and keep these data for a decade or more. With studies using written questionnaires, these could be scanned and also stored electronically.

The public availability of data sets would also facilitate the application of statistical methods designed to expose scientific fraud. Such methods have been used by Simonsohn (2012) in identifying problems in the articles of social psychologists Smeesters and Sanna. Both resigned as a result of these accusations. In the case of Smeesters, a university investigation committee concluded that the findings reported in three of his articles were "probably the result of data selection by Smeesters" (Erasmus University Rotterdam, 2012). The case of Sanna was investigated by a committee at the University of North Carolina, where he had worked when he published the suspected research (before he moved to the University of Michigan). The findings of the committee were not made public. However, Sanna resigned his position at the University of Michigan and withdrew three of his published articles (Yong, 2012). Other methods of statistical fraud detection have been suggested by Diekman (e.g., 2007). While the development of such methods is certainly an extremely promising way to identify fraudulent research, it still needs to be clarified how well such methods discriminate between fraudulent and non-fraudulent research. Also, public availability of data would have little effect, unless there was some probability of the data being scrutinized and reanalyzed. As we will discuss later, one way of assuring this would be through the institution of random audits being conducted by research institutions.

***Facilitating replications*.** Another widely accepted strategy of fraud detection is to encourage replications of studies. For example, Crocker and Cooper (2011) argued: "Despite the need for reproducible results to drive progress, studies that replicate or fail to replicate) others' findings are almost impossible to publish in top scientific journals. This disincentive means fraud can go undetected, which was the case with Stapel". And similarly, Chambers and Sumner (2012) write: "Replication is our best friend because it keeps us honest. In science, false results have a short (albeit potentially damaging) lifespan because regardless of how they come about, other scientists won't be able to reproduce them. On the other hand, true results will be replicated time and time again by different scientists." Mummendey (2012, p. 7) goes even further and suggests: "Scientific journals could expand their already high standards of the peer review system by adding the requirement for a thorough external replication. Authors submit their manuscript together with their data. Once the publication has been approved by a preliminary group of reviewers, the editors invite suitable experts to attempt a replication of the results. After this has been accomplished, both the original manuscript and the replication study are published together."

Our perspective, however, is that this trust in the power of replications is somewhat idealistic and even misguided. First, purely in practical terms there is the problem of the doubling of resources needed to conduct publishable research (and such resources may not be easily or equally available to all in these difficult economic times). Second, we have numerous examples in the psychological literature, where "true" results repeatedly failed to be replicated. [The more ancient among us will still remember the controversy surrounding the Festinger and Carlsmith (1959) results, which could only occasionally be replicated, until it was discovered that freedom of choice and negativity of consequences were essential for the effect to emerge.] Since there are always numerous reasons for a given finding not to be replicated, failure to replicate cannot be seen as a reliable indicator of fraud. Furthermore, since due to their high productivity fraudsters are often highly respected in their field, even blatant failures to replicate their findings might not arouse suspicion. Finally, even successful replication cannot be seen as indication that the original result was *not* fraudulent. Since fraudsters are typically careful in suggesting plausible hypotheses, it is quite possible that these hypotheses might have been supported by an empirical study had the fraudster cared to conduct it. In the case of Stapel, one of the committees examining his publications has suggested that some of his PhD research was fraudulent (Keulemans 2012), but these findings have been replicated on occasion, at least conceptually.

Clearly, information about multiple failures to replicate a study is important because it suggests that a given finding is not very reliable or stable. Furthermore, the indication that different findings of a particular researcher or research group cannot be replicated might signal that there could be a problem. For example, the failures to replicate research by the physicist Jan Hendrik Schön motivated his colleagues to have a close look at his publications (Reich, 2009). This led to the discovery that he had published similar performance curves for different devices and ultimately to the discovery of his fraud (Reich, 2009). Therefore the recent creation of a website, where researchers can upload and view results of replication attempts in experimental psychology is a useful initiative (PsychFileDrawer.org, 2012). However, it is difficult to decide on the basis of a brief summary how, or how well a given study was done. Furthermore, the researchers who report their replications to this web site should be required (rather than merely advised) to download their data.

We are somewhat less convinced of the usefulness of the Open Science Collaboration and their plan to replicate all studies published in three journals during a given year (http://www.openscienceframework.org/). Although the initiators of this collaboration emphasize that they do not target fraud per se, but hope to check the extent to which psychological research can be replicated, one can doubt whether the information that a certain percentage of studies did not replicate will justify the enormous investment in research time and resources that this task requires. Furthermore, since social psychological research is more sensitive to social and other context factors than is research in psychophysics, it is very likely that the project will find that social psychological research is less replicable.

***Instituting research audits*.** Another frequent method through which fraud was identified in our sample of cases was through research audits. During such an audit, researchers have to disclose all the material used in a given study and typically their data are reanalyzed. Such research audits are frequently conducted in medicine, but mostly when there is already suspicion of research fraud. However, to be successful in fraud prevention, audits should be conducted on a random basis and not only once there has been reason for serious suspicion. Although this seems impracticable, we know of at least one research institution, where such random audits are being practiced http://www.emgo.nl/kc/Audit/1%20Internal%20Project%20Audit%20Procedure.html

Such audits would not only discover outright fraud, they would also discourage behaviors in the grey areas between good practice and scientific misconduct. As part of such an audit all members of a research group would be interviewed and unusual research practices could be identified. For example, Stapel claimed to have done field research (e.g., Stapel & Lindenberg, 2011). Since it is implausible that senior researchers collect such data themselves, a research assistant would have been involved in real data collection (assuming this occurred), who could have been interviewed in an audit. (The cases of Sir Cyril Burt and also Karen Ruggiero, where such assistants appeared not even to exist, might have brought these frauds to light earlier had they been audited). Since only a small proportion of research projects could be audited in this way, the probability that one of Stapel's projects would have been audited does not seem all that great. However, given that he was an active inventor/researcher at three research institutions, the chance is not negligible. Furthermore, since his fraud would most likely have

been discovered in such an audit, the knowledge that such audits are being conducted might have discouraged his behavior.

Conclusions

Even though the prevalence of research fraud is likely to be low [most estimates put it around 1% to 2% (Stroebe et al., in press)], scientists have a particular responsibility to society and it is understandable that reports of research fraud are greeted with a public outcry. Although the Stapel case hardly exposed deep flaws in the way we conduct our science, it clearly demonstrated that any trust-based system, as science is, is open to exploitation. We therefore need to look at our procedures and check whether they can be tightened. Any system can be improved, and lessons can be learned from the Stapel case as well as from the many other cases of fraud. And the major lesson to be learnt is that the assumption that science is self-correcting and that findings based on falsification will eventually be discovered and rejected is an illusion.

We have been criticized for drawing this conclusion by colleagues who have argued that such claims will cause people to lose trust in science. In our opinion, the trust in science is undermined by cases of research fraud and not by analyses of underlying causes of fraud (although we would add that exposing cases of fraud should help us to rebuild trust in the long run). The Stapel fraud was a wake-up call that motivated social psychology to scrutinize their research practices. And although there is much good in Social Psychology and although the discipline has been very successful in recent years, the case threw light on some procedural weakness (which our research suggests can be found in our own discipline as well as many others) that need to be addressed and fixed. If we use the case as a learning experience rather than deciding to return to "business as usual", something good will have come out of this painful episode.

References

American Psychological Association (2010). Publication Manual of the American Psychological Association. Sixth Edition. Washington, DC: American Psychological Association.

Broad, W. J., & Wade, N. (1982). *Betrayers of the truth*. New York, NY: Simon & Schuster.

Carey, B. (2011). Fraud case seen as a red flag for psychology research. *The New York Times.* Retrieved from http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html

Chambers, C. & Sumner, P. (2012). Replication is the only solution to scientific fraud. The Guardian, September 14. Retrieved from http://www.guardian.co.uk/commentisfree/2012/sep/14/solution-scientific-fraud-replication

Crocker, J., & Cooper L. (2011). Editorial: Addressing scientific fraud. *Science, 334*, 1182. doi:10.1126/science.1216775. Retrieved from

http://www.sciencemag.org.proxy.library.uu.nl/content/334/6060/1182.full

Diekmann, A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics, 34*, 321–329.

Erasmus University of Rotterdam. (2012, June 1). *Report by the committee for inquiry into scientific integrity.* Retrieved from http://www.eur.nl/fileadmin/ASSETS/press/2012/Juli/report_Committee_for_inquiry_prof._Smeesters.publicversion.28_6_2012.pdf

Fang, F. C., & Casadevall, A. (2011). Retracted science and the retraction index. *Infection and Immunity, 79*, 3855–3859.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 59*, 203–210.

Goodstein, D. (2010). *On fact and fraud*. Princeton, NJ: Princeton University Press.

Keulemans, M. (2012b, June 1). Diederik Stapel pleegde al fraude tijdens promotie in Amsterdam [Diederik Stapel already committed fraud during his dissertation in Amsterdam]. *De Volkskrant*. Retrieved from http://www.volkskrant.nl/vk/nl/2844/Archief/archief/article/detail/3264084/2012/06/01/Diederik-Stapel-pleegde-al-fraude-tijdens-promotie-in-Amsterdam.dhtml

Mummendey, A. (2012). Scientific misconduct in social psychology: Towards a currency reform in science. *European Bulletin of Social Psychology, 24*, 4–7.

Reich, E. S. (2009). *Plastic fantastic*. New York, NY: Palgrave Macmillan.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *Observer.* Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html

Simonsohn, U. (2012). *Just post it: The lesson from two cases of fabricated data detected by statistics alone* (Unpublished report). The Wharton School, University of Pennsylvania, PA.

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

Stapel, D. (2011). Stapel betuigt openlijk 'diepe spijt' [Stapel declares publicly "deep regret"]. *Brabants Dagblad.* Retrieved from http://bd.nl/nieuws/tilburg-stad/stapel-betuigt-openlijk-diepe-spijt-1.121338

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

Yong, E. (2012). Uncertainty shrouds psychologist´s resignation. *Nature News*, http://www.nature.com/news/uncertainty-shrouds-psychologist-s-resignation-1.10968

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, e26828. doi: 10.1371/journal.pone.0026828

What Is Wrong With Social Psychology?

Gregory Mitchell (School of Law, University of Virginia)

Those weary of discussions of Hauser, Stapel, Sanna, and *p*-hacking may be relieved to discover that my title does not refer to the recent revelations of manufactured data and other questionable research practices by social psychologists to generate statistically significant results (e.g., John, Loewenstein & Prelec, 2012; Simonsohn, 2012). My title is based on two other troubling facts that have received less attention on the blogs, in the popular press, and in our journals.

First is my recent finding that social psychology fared much worse than other psychological subfields in a comparison of results in the laboratory and the field (Mitchell, 2012). This replication and extension of Anderson, Lindsay and Bushman (1999) collected 82 meta-analyses in which effects in the laboratory were compared to effects in the field (e.g., were the effects of alcohol on behavior the same in a "bar lab" as in a real bar?) and examined the correlation, relative magnitude, and constancy of effect direction for 217 pairs of effects obtained from these meta-analyses for a wide range of phenomena from many psychological subfields. I found that industrial-organizational psychology performed remarkably well in the field (r = .89 for paired lab and field effects), and the magnitude of effects were similar in the lab and field; laboratory studies from personality psychology also held up well in the field (r = .83), but there were considerably fewer paired effects for this subfield than for I-O and social psychology. Social psychology performed much worse: over 20% of effects from social psychology laboratories changed signs in the field, the correlation of lab and field results was much lower (r = .53 if we exclude an outlier pair of effects), and the relative magnitude of effects differed greatly between the lab and field. Social