On the quality of quality assessments

Lars Engwall[1]
Department of Business Studies, Uppsala University, Sweden

Introduction

A main thread throughout the present volume is the tendency in the scientific world to increasingly rely on bibliometrics in the evaluation of academic institutions as well as individual scholars. Impact factors, based on the citations of articles in specific journals during a 2-year period, have become important for publishers for the prestige of their journals and at the end of the day for subscription figures and subscription rates. These impact factors in turn have strong effects on the publishing behaviour of researchers, who for natural reasons, try to be published in journals with as high impact factors as possible. As a result, in many scientific fields today, academic institutions have outsourced the task of making the quality assessments to journal editors and their reviewers. Although earlier evaluations of candidates on the academic job market in many instances were based on the reading of their publications, current evaluations are increasingly based on the impact factors of the journals where the research has been published. Such an approach has the advantage of efficiency of course: comparing numbers of publications weighted by their impact factors is clearly much less time-consuming than providing a personal opinion after careful reading. However, the use of bibliometrics for evaluations rests on one very vital assumption, i.e. that the quality assessments made by journal editors and their reviewers can be trusted. The present chapter will point to the fact that this is not always true and will provide an analysis of the quality of the quality assessments provided by journal editors and their reviewers. In so doing, it will point to two types of errors: first, the reject error, i.e. the rejection of papers that eventually appear to be very important, and secondly, the accept error, i.e. the acceptance of papers that eventually turn out to be fraudulent. For both types of errors, examples will be presented. In addition, the present chapter will point to the risk that these errors may increase in number with the passage of time. It is argued that the peer-review system is gradually deteriorating for two reasons: first, the radically increasing flow of manuscripts, and secondly, the intensified pressure on individuals to publish. The former circumstance will imply that the demand for reviewers is increasing, whereas the latter has the consequence that the supply of reviewers is decreasing. Also, in this case, empirical evidence will be provided. The final section discusses the conclusions.

[1]Email: lars.engwall@fek.uu.se

Errors in the review system

In analysing the review system, it is appropriate to recall the risks involved in statistical testing (Figure 1, left-hand panel) [1]. It refers to two types of errors: type I error, implying that a correct hypothesis is rejected, and type II error implying that a false hypothesis is not rejected. As shown in the right-hand panel of Figure 1, the same type of problems appears in the review system. The type I error implies that an important significant contribution is rejected (reject error), whereas the type II error means that bad papers are accepted (accept error). In the following two subsections, I will discuss these two errors, respectively.

The reject error

Since reviewers are likely to be conservative in their quality assessments, reject errors can be expected to occur widely, although they are more difficult to identify than accept errors. However, an illustration of the occurrence of reject errors is the experiment performed by the two psychologists Douglas P. Peters and Stephen J. Ceci some 30 years ago [2]. They selected 12 articles, which had already been published in 12 highly regarded psychology journals, and sent them back for review 18–32 months after their publication. They then found that out of 38 editors and reviewers, only three detected the resubmissions. Only one of the nine remaining articles was accepted. Of the 18 referees, 16 recommended reject, in many cases referring to "serious methodological flaws".

There are also a number of rather spectacular non-experimental examples of reject errors. The 1977 Nobel Laureate for Physiology or Medicine, Rosalyn S. Yalow, pointed out in her Nobel Lecture that she, and her long-time collaborator, Solomon Berson (who could not share the prize since he died in 1972), had difficulty in publishing their results [3]:

> "*we were able to demonstrate the ubiquitous presence of insulin-binding antibodies in insulin-treated subjects* […] *This concept was not acceptable to the immunologists of the mid-1950s. The original paper describing these findings was rejected by Science and initially rejected by the* Journal of Clinical Investigation […] *A compromise with the editors eventually resulted in acceptance of the paper.*"

Figure 1

Statistical Testing		
Action	H_0 *is True*	H_0 *is False*
Reject	Type I Error	Correct
Accept	Correct	Type II Error

Reviewing		
Action	*Good Paper*	*Bad Paper*
Reject	Reject Error	Correct
Accept	Correct	Accept Error

Errors in statistical testing and in review systems

Berson and Yalow shared early rejections with another Nobel Laureate for Physiology or Medicine, Hans Krebs, who received the prize in 1953. His paper on his path-breaking findings was thus returned from *Nature* "in case [you prefer] to submit it for early publication to another periodical" [4].

And manifold other examples of the reject errors could be quoted. Although, in hindsight, they cast an aura of stupidity around rejecting editors, such errors nevertheless do not lead to scandals. This is instead the case for the accept errors.

The accept error

In the same way that an experiment could be quoted in relation to reject errors, there is also an example for accept errors. This is the submission in the 1990s of a paper by the physicist Alan Sokal to the journal *Social Text* [5], where he provided a parody of modern cultural studies. After the article had been published, Sokal in a subsequent paper [6] revealed his experiment and phrased his hypothesis in the following way:

> "*Would a leading North American journal of cultural studies* […] *publish an article liberally salted with nonsense if (a) it sounded good and (b) it flattered the editor's ideological preconceptions?*"

Needless to say, the Sokal experiment caused embarrassment among post-modern scholars and led to a lot of discussions, which Sokal has summarized in the book *Beyond the Hoax* [7]. However, there are also a number of examples of real scientific frauds. Some 30 years ago, the two science journalists William Broad and Nicholas Wade even started out their book *Betrayers of the Truth* [8] by pointing to fraudulent behaviour among historically recognized scientists such as Ptolemaios, Galilei, Newton, Dalton and Mendel. They also provided information about a number of modern scientific frauds. One of them is Elias A.K. Alsabati, who systematically published already published articles in lesser-known journals [9]. Others include various manipulations of data by the researchers themselves, but also cases where doctoral students manipulated data in order to please their professors (see Chapter 8 in [8]).

As the book by Broad and Wade [8] was published some 30 years ago, it might be argued that the system has become more professionalized and that such fraudulent behaviour is no longer possible. This is not the case, however. Several more recent examples can be mentioned. One of them is the Norwegian oral cancer researcher Jon Sudbø, who managed to publish articles in three prestigious journals, with current impact factors around 50, 18 and 30 respectively, based on manipulated data: *The New England Journal of Medicine* in 2004 [10], *The Journal of Clinical Oncology* in 2005 [11] and *The Lancet* in 2005 [12]. He was exposed and the above-mentioned and several other publications by him were retracted [13].

Another more recent scientific swindler, who has attracted considerable attention, is Woo-Suk Hwang, a professor at Seoul University involved in stem cell research [14]. He and his group even published two articles on cloning in *Science* in 2004 and 2005 [15,16]. However, as early as the fall of 2005 fraudulent behaviour was detected, and the two articles were withdrawn in January 2006 [17].

Nevertheless, in April 2013 the two papers scored 723 and 437 citations respectively on Google Scholar [18].

A third modern example, mentioned in Chapter 4 by Jane Grimson, is Andrew Wakefield, who claimed there was a relationship between vaccination for measles, mumps and rubella, and the appearance of autism and bowel disease. This was based on research results that had been published in *The Lancet* in 1998 [19] and in several other journals. After a long process, where the science journalist Brian Deer played a significant role, the article in *The Lancet* and also those in other journals were retracted [20].

Two very recent examples are provided by the evolutionary biologist Marc Hauser at Harvard University and the Dutch social psychologist Diederik Stapel at Tilburg University. Hauser was considered as a leader in animal cognition research and had among his publications a widely cited article in *Science* (2584 citations on Google Scholar in April 2013), co-authored with the renowned linguist Noam Chomsky and the evolutionary biologist W. Tecumseh Fitch [21]. However, after an investigation that started in 2007 several of his papers were questioned. Hauser was found guilty of scientific misconduct, and among his papers one published in 2002 in *Cognition* (impact factor 3.162) [22] was retracted [23]. Among those following the case was one of the authors of *Betrayers of the Truth* [8], Nicholas Wade [24].

In the case of Diederik Stapel, whistleblowing from some of his students led to the setting up of three committees to investigate his research [25]. They found instances of fraud in some 50 publications, among them dissertations that he had supervised, but also a large number of published articles. Once again, among the fraudulent publications, was a paper published recently in *Science* [26], eventually retracted by Stapel after the committee reports [27]. Most of the rest of the papers were published in the *Journal of Experimental Social Psychology* (impact factor of 2.202), *British Journal of Social Psychology* (impact factor of 1.987) and *European Journal of Social Psychology* (impact factor of 1.592) (see Appendix 4–6 in [25], and [28]). From the first of these journals, published by the American Psychological Association, 14 of the papers were withdrawn after the committee reports [29].

The above may give the impression that scientific fraud is limited to medicine and psychology. This is not the case. One spectacular case is Jan Hendrik Schön from Germany, who rose to prominence through his works on semiconductors at Bell Labs at the turn of the present century [30,31]. His findings, that transistors could be produced on a molecular scale, attracted wide attention. Again, prestige journals *Nature* and *Science* published papers that were fraudulent. More and more scholars in the field started questioning Schön's findings and a committee found a large number of cases of scientific misconduct [32]. As a result, several papers published in *Nature*, *Science*, *Physical Review* and *Applied Physical Letters* were withdrawn (see e.g. [33–36]).

It could be argued that the above examples of fraudulent behaviour are just a small group of misbehaving individuals. However, a recent systematic review of 2047 biomedical and life-science research articles that had been retracted up to May 2012 reveals that 67.4% of the retractions were attributed to misconduct [37]. The peer-review system thus failed also in a considerable number of other cases.

Errors, impact and reviewers

Errors and impact

The above examples certainly provide evidence that the fraudulent behaviour reported by William Broad and Nicholas Wade in the early 1980s [8] can be found also in subsequent decades. This is particularly worrying, since, as mentioned in the Introduction, the last quarter of a century or so has entailed a tendency that individuals are evaluated on the basis of their success in publishing in top journals. However, as we have seen above, there are reasons to be cautious in drawing the conclusion that high-impact journals publish high-quality papers. Jon Sudbø and Andrew Wakefield published articles in *The Lancet* (impact factor of 32.280) and Woo-Suk Hwang, Marc Hauser and Diederik Stapel published in *Science* (impact factor of 31.201) [37]. In addition, the *Journal of Experimental Social Psychology* (impact factor of 2.202) published as many as 14 articles by Stapel that later had to be retracted. There are therefore strong reasons to question the large emphasis on impact factors. As pointed out by Per O. Seglen [38], they are inappropriate for use in evaluating research, but are nevertheless widely used, or in the words of Kai Simons [39], misused.

Above, we have thus seen that the review system is far from perfect. This means that impact figures should be handled with much more care than is usually the case. There is no doubt that these figures with three decimals create a false impression of exactitude, which has nothing to do with quality. In addition to the deficiencies in reviewing, with both reject errors and accept errors, there are, as pointed out by Seglen [38], a number of reasons to question the use of impact factors for the evaluation of research. It is well known that editors use all sorts of methods to raise the impact factors of their journals: the publication of review articles, pushing for citations of earlier articles from the journal, privileging longer articles, etc. In addition, it is extremely important to note that distributions of citations are highly skewed: most articles, even in high-impact journals, are only cited a few times, if at all. Furthermore, comparisons across scientific fields are inappropriate since they face different conditions. Nevertheless, impact factors are often used in resource-allocation decisions and evaluations.

As mentioned above, in discussing the relationship between impact and quality, it is easier to find spectacular examples of the accept error. However, it is equally important to consider the reject error in this context. It could even be argued that, owing to reject errors, the most innovative research will be found in journals with low impact factors. The reason for this would be that the high-impact journals will be dominated by the scientific elites. These are more likely to reject innovative papers outside the mainstream. As a result, we could expect that significant papers, after rejections in the top journals, end up in lower-impact journals.

Errors and reviewers

Above, we were able to conclude that errors in the peer-review system also occur today. As a matter of fact, we may even suspect that the probabilities for such errors have increased with the passage of time. First of all, owing to the above-mentioned expansion of the system, and the strong pressure to publish, there has been and will be strong growth in the number of manuscripts that are sent to the scientific journals. Secondly, individual researchers may become less inclined to spend time reviewing manuscripts for their community, efforts for which they are not rewarded when they are assessed. This in turn appears to lead to increasing difficulties for editors in recruiting reviewers, and an increasing tendency for them to reject manuscripts without reviews (desk rejects). And, if they get reviewers, we could expect them to be slower to deliver and less careful when they deliver. Such a development may lead both to reject errors and accept errors due to lack of careful reading of manuscripts.

Some evidence regarding the tendencies described above has been obtained in correspondence with the editors of *Nature* and *The New England Journal of Medicine*, as well as the editors of four management journals.

Nature receives some 12 000 manuscripts a year with an increasing trend from 10 584 in 2008 to 12 552 in 2012. A large proportion, as much as 70–80% of these manuscripts, are rejected without review. For the remaining submissions the editor estimates that approximately 30% of those asked to write a review decline. Those who accept the task return their reviews, which are "mostly excellent [...] within 3 weeks or so" (*Nature* Editor-in-Chief, Philip Campbell, personal communication, May 2013). Nevertheless, the strong pressure on the journal in terms of submissions may of course involve risks for errors.

The New England Journal of Medicine has also experienced an increase in the number of submissions. As a result, the share of the desk rejects has increased from 39% in 2002 to 62% in 2012. Nevertheless, the number of reviews has increased from 3870 in 2002 to 6058 in 2012. In the same period, the turn-around performance has increased considerably: 48% of the reviews were returned after 2 weeks and 78% after 3 weeks in 2002, whereas the corresponding figures in 2012 were 67% and 93% respectively (Assistant to the Editor, *The New England Journal of Medicine*, Pamela Miller, personal communication, May 2013). Again, we can see evidence of an increasing flow of manuscripts and higher shares of desk rejections. In addition, we also see increasing pressures to process the manuscripts under review faster.

The increasing flow of manuscripts is also confirmed by the four editors of management journals (personal communications, February 2013). This, in turn, has implied that for all four, approximately one-third of the manuscripts are desk rejected:

> "*On average, one-third of those submissions will be desk rejected.*" (U.S. editor, personal communication, February 2013)

> "[The percentage of desk reject] *was quite high throughout my tenure (papers falling outside the scope of the journal is one category; but there is a lot of rubbish circulating around the world, and most editors will tell you the same story – submissions from outside Europe and North America too often fall into this category.*" (European editor 1, personal communication, February 2013)

> "*At least 30% – my criterion for desk reject was a paper that I would be annoyed receiving for review myself.*" (European editor 2, personal communication, February 2013)

> "*This was fairly constant at around 35%. But some years saw an increase. This usually coincided with the Research Assessment Exercise in the U.K. when there would be a flood of mediocre papers to beat the census date (not all papers were poor quality but the majority were). This would take the desk rejection rate up to 39/40%. But these were 'blips' in an otherwise constant 35% rejection rate.*" (European editor 3, personal communication, February 2013)

The perceived need to desk reject certainly implies considerable power for editors, and this in turn could increase the probability of reject errors.

As the European editor 2 points out, the desk reject is important in order not to annoy reviewers. It is also essential in order to decrease the need for reviewers, since, as mentioned above, it is difficult to get their assistance:

> "*This number has grown quite strikingly over the years. We use three reviewers per manuscript. About five years ago we would have to ask 3.3 people to get 3 reviewers. Today, that number is very close to 4 requests for three reviewers. I think that academics today are much more inclined to refuse a review than past generations.*" (U.S. editor, personal communication, February 2013)

> "*Our problem was reviewers never responding* […] *Often I had to contact 6–7 people to get 3, and too often I had to contend with 2 reviewers.*" (European editor 1, personal communication, February 2013)

> "*30%, on average one of the three assigned reviewers for each paper would be replaced in the process.*" (European editor 2, personal communication, February 2013)

> "*This was my greatest headache. I think I coincided as Editor with a time when academics worldwide were becoming increasingly busy in their own institutions and (voluntary) reviewing came fairly low down the priority list for many. I would estimate that out of every three requests for review, one would always decline, so a minimum of 33% declining to review. Annoyingly, some sat on the paper for a few weeks and said nothing (so I assumed they were reviewing the paper). Then they would contact me and say they had to decline. It was at this point that I widened the pool of reviewers to include some junior (but good) faculty. I would use two senior reviewers and one less senior. The standard 'set' of reviewers would comprise one editorial board member, one external senior specialist in the field and one less senior (often general) reviewer. This worked well both in terms of quality and in terms of getting three good reviews for each paper.*" (European editor 3, personal communication, February 2013)

And, when the assigned reviewers accept the assignment, they take more time than editors like:

> "*We ask our reviewers to provide their comments within 30 days. We have, over the years, developed a broad range of mechanisms to ensure that we hold*

to this number. The logic is this. If we want to get the best manuscripts we have to ensure that authors get prompt but high quality feedback. So, when a reviewer becomes unresponsive or very late with their review, I encourage my reviewers to go with only two reviews if there is consistency in their opinion on the manuscript. If there is not, I have designated some reviewers to be 'on call'. These tend to be very good reviewers who understand that they will sometimes be asked to render an opinion in an attenuated time frame." (U.S. editor, personal communication, February 2013)

"*Chasing reviewers was cumbersome* [...] *My experience is that this is getting harder and harder to get people to deliver on time (for us: within 5–6 weeks).*" (European editor 1, personal communication)

"*They were given about six weeks but most took three months.*" (European editor 2, personal communication, February 2013)

"*We would demand a 12-week turnaround, but more often than not you would have to wait around 16 weeks to get three reviews for a paper. There were stunning variations however and the range was quite wide. For example I had two or three papers which had all 3 reviews received in 4 weeks. On the other hand, two or three papers had review times of nearly 20 weeks. I guess 16 weeks would be around the average.*" (European editor 3, personal communication, February 2013)

In the present chapter, we may note that the U.S. editor is pushing harder than his European colleagues. This could even be interpreted as an effort to use turnaround time of manuscripts as an asset for a particular journal. However, speed is not enough; quality is also very important in the reviews, and in this respect, editors appear so far to be fairly content with their reviewers:

"*Quality of reviews is of utmost importance. The associate editors rate the quality of each review on a five point scale, with space for additional comments. If reviewers excel in both quality and timeliness, they will be elevated to the editorial review board. Unlike many journals, thus,* [our journal] *is a meritocracy built on your skill in reviewing. We also acknowledge our outstanding reviewers at the* [...] *meetings every year. We have just started to put on seminars* [...] *on how to write a high-quality review.*" (U.S. editor, personal communication, February 2013)

"*Quality of reviews varies a lot: junior scholars often do a thorough job, while some big names end up on a black list after doing an incredibly poor job, or forgetting about having promised to review* [...] *At the end of the day, editors end up with a bunch of reliable people who you then bother too much* [...] *An afterthought: I think people should be rewarded by their home universities for reviewing! This is something that is too often forgotten in the contemporary system, which celebrates publishing.*" (European editor 1, personal communication, February 2013)

"*On the whole, the reviews were good. Very seldom did I have to 'reject' a review because of quality or inappropriate advice (but it did happen). While*

> *being good, they were also quite often not in mutual agreement. As editor, I had the privilege of mediating between contradictory reviews without alienating any of them.*" (European editor 2, personal communication, February 2013)

> "*Mostly very good. I would say 80–90% of reviews were of good quality. This was a time when I cleared out some of the less active of the Editorial Board (who tended to hand in pedestrian and descriptive reviews which were of little help to me or the authors). Things got better after that in terms of quality of reviews and it was quite rare to receive a poor quality review.*" (European editor 3, personal communication, February 2013)

From the above comments it is particularly interesting to note how the U.S. editor mentions how the professional association is strengthening reviewing through acknowledgment and through seminars. In the same spirit, the European editor 1 suggests that "people should be rewarded by their home universities for reviewing."

All in all, the editors of the four management journals are thus fairly content with the reviews they have received, although they agree that, first, there is a need to reject about one-third of the manuscripts themselves due to poor quality, secondly, a considerable share of reviewers decline, and thirdly, those who agree take too much time to deliver. To what extent the review processes have reject errors cannot be determined. So far, the editors in question have not been obliged to face accept errors and retractions.

The examination of scientific output is nowadays not limited to editors and their reviewers, however. As far as plagiarism is concerned, modern information technology facilitates more advanced monitoring. Evidence of this is provided by a number of German politicians who have been found to be plagiarists after their dissertations were run through the program VroniPlag. Among them were the Minister of Defence Karl-Theodor zu Guttenberg, the Minister for Education and Research Annette Schavan, as well as the European Parliament members Georgios Chatzimarkakis and Silvana Koch-Merin [40].

The German cases point to the increasing role of the group of actors labelled above as scrutinizers in the governance of academia. Other examples are provided by the work of the science journalists William Broad, Brian Deer and Nicholas Wade, as mentioned above. However, the most powerful mechanism for detecting scientific fraud is always the critical examination by close colleagues and, as was the case regarding Stapel, students.

Conclusions

The point of departure for the present chapter has been the increasing tendency in various disciplines to outsource quality assessments to journal editors and their reviewers. Far too often it seems that actual reading of submitted material is replaced by counting the number of publications in what are considered to be top journals. And these assessments are based on calculations of impact factors, which in turn are based on citations.

Scholars are taking these shortcuts in evaluations at the same time as many of them complain about the power positions that bibliometrics has attained in the past decades. Against this backdrop the paper has provided the following arguments:

1. Editorial decisions include two errors: the reject error and the accept error. The first type is constituted by the rejection of papers that deserve publication, whereas the second type occurs when bad papers are published.
2. Although it is difficult to determine the extent to which these two types of error occur, a number of examples have been provided to show that they have indeed occurred. In this context, it is particularly worth noting that the scandalous accept error has happened in several prestigious high-impact journals, and that these incidents do not seem to harm their reputation.
3. Under the above-mentioned circumstances, it is apparent that it is far from reliable to use impact factors as measures of quality and even less reliable to use them to assess individual researchers.
4. Journal editors seem to face an increasing flow of manuscripts (particularly during times of research assessments) as well as a resistance from potential reviewers, who, as a result of the strong focus on publishing, prefer to publish more of their own papers instead of doing reviews. In addition, those who accept reviewing assignments take more time for their assessments.
5. Finally, it has been pointed out that modern information technology offers new opportunities to reveal plagiarism.

All of this means that it is now urgent more than ever to critically examine academic studies, and not outsource quality assessments to journal editors and their reviewers. As pointed out by Sydney Brenner [41], "we should remind ourselves that what matters absolutely is the scientific content of a paper and that *nothing will substitute for either knowing it or reading it*" [Author's italics].

References

1. Wallis, W.A. and Robert, H.V. (1956) *Statistics: A New Approach*. Free Press, New York
2. Peters, D.P. and Ceci, S.J. (1982) Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behavioral and Brain Sciences* **5**, 187–255
3. Yalow, R.S. (1992) Radioimmunoassay: a probe for fine structure of biologic systems. In *Nobel Lectures, Including Presentation Speeches and Laureates' Biographies 1971–1980* (Lindsten, J.E., ed.), pp. 447–468, World Scientific, Singapore
4. Borrell, B. (2010) Nature rejects Krebs's paper, 1937. In *The Scientist*. 1 March 2010, http://www.the-scientist.com/?articles.view/articleNo/28819/title/Nature-rejects-Krebs-s-paper--1937/
5. Sokal, A. (1996) Transgressing the boundaries: toward a transformative hermeneutics of quantum gravity. *Social Text* **14**, 217–252
6. Sokal, A. (1996) A physicist experiments with cultural studies. *Lingua Franca* **6**, 62–64
7. Sokal, A. (2008) *Beyond the Hoax: Science, Philosophy and Culture*. Oxford University Press, Oxford

8. Broad, W. and Wade, N. (1982) *Betrayers of the Truth*. Simon and Schuster, New York
9. Broad, W.J. (1980) Would-be academician pirates papers: five of his published papers are demonstrable plagiarisms, and more than 55 others are suspect. *Science* **208**, 1438–1440
10. Sudbø, J., Lippman, S.M., Lee, J.J. et al. (2004) The influence of resection an aneuploidy on mortality in oral leukoplakia. *New England Journal of Medicine* **350**, 1405–1413
11. Sudbø, J., Samuelsson, R., Risberg, B. et al. (2005) Risk markers of oral cancer in clinically normal mucosa as an aid in smoking cessation counseling. *Journal of Clinical Oncology* **23**, 1927–1933
12. Sudbø, J., Lee, J.J., Lippman, S.M. et al. (2005) Non-steroidal anti-inflammatory drugs and the risk of oral cancer: a nested case-control study. *The Lancet* **366**, 1359–1366
13. Vastag, B. (2006) Cancer fraud case stuns research community, prompts reflection on peer review process. *Journal of the National Cancer Institute* **98**, 374–376
14. Normile, D. (2009) Hwang convicted but dodges jail; stem cell research has moved on. *Science* **326**, 650–651
15. Hwang, W.S., Ryu, Y.J., Park, J.H. et al. (2004) Evidence of a pluripotent human embryonic stem cell line derived from a cloned blastocyst. *Science* **303**, 1669–1674
16. Hwang, W.S., Roh, S.I., Lee, B.C. et al. (2005) Patient-specific embryonic stem cells derived from human SCNT blastocysts. *Science* **308**, 1777–1783
17. Kennedy, D. (2006) Editorial retraction. *Science* **311**, 335
18. Google Scholar search result: Woo-Suk Hwang. http://scholar.google.se/scholar?hl=sv&q=Woo-Suk+Hwang&btnG= (Accessed April 2013)
19. Wakefield, A.J., Murch, S.H., Anthony, A. et al. (1998) Ileal lymphoid nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet* **351**, 637–641
20. Godlee, F., Smith, J. and Marcovitch, H. (2011) Wakefield's article linking MMR vaccine and autism was fraudulent. *British Journal of Medicine* **342**, c7452
21. Hauser, M.D., Chomsky, N. and Fitch, W.T. (2002) The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579
22. Hauser, M.D., Weiss, D. and Marcus, G. (2002) Rule learning by cotton-top tamarins. *Cognition* **86**, B15–B22
23. Retraction notice. (2010) *Cognition* **117**, 106
24. Shaw, K. (2010), Harvard professor found guilty of scientific fraud. In *Ars Technica*. 23 August 2010, http://arstechnica.com/science/2010/08/harvard-professor-found-guilty-of-scientific-misconduct/ (Accessed April 2013)
25. Levelt Committee, Noort Committee and Drenth Committee (2012) *Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel* (Translation of *Falende Wetenschap: De Frauduleuze Onderzoekspraktijken van Social-Psycholoog Diederik Stapel*), Tilburg University, Tilburg
26. Stapel, D.A. and Lindenberg, S. (2011) Coping with chaos: how disordered contexts promote stereotyping and discrimination. *Science* **332**, 251–253
27. Stapel, D.A. and Lindenberg, S. (2011) Retraction. *Science* **332**, 2 December
28. Impact factors of psychology journals. http://psychology.wikia.com/wiki/Impact_factors_of_psychology_journals (Accessed April 2013)
29. Retraction watch. http://retractionwatch.com/?s=stapel (Accessed April 2013)
30. Reich, E.S. (2009) *Plastic Fantastic. How the Biggest Fraud in Physics Shook the Scientific World*. Palgrave Macmillan, New York
31. Goodstein, D. (2010) *On Fact and Fraud: Cautionary Tales from the Frontline of Science*. Princeton University Press, Princeton
32. Beasley, M.R., Datta, S., Kogelnik, H., Kroemer, H. and Monroe, D. (2002) Report of the Investigation Committee on the Possibility of Scientific Misconduct in the Work of Hendrik Schön and Coauthors. The Lucent Technologies Report. Available at: http://www.engineering.utoronto.ca/Assets/AppSci+Digital+Assets/pdf/GradStudents+Ethics/Schoen_Full+Report.pdf
33. Retractions (2003) *Nature* **422**, 92. Also available at: http://www.nature.com/nature/journal/v422/n6927/pdf/nature01463.pdf
34. Bao, Z., Batlogg, B., Berg, S. et al. (2002) Retraction. *Science* **298**, 961. Also available at: http://www.sciencemag.org/content/298/5595/961.2.full
35. Schön, J.H., Kloc, C. and Batlogg, B. (2002) Retraction: universal crossover from band to hopping conduction in molecular organic semiconductors [Phys. Rev. Lett. 86, 3843 (2001)]. *Physical Review Letter* **89**, 289902
36. Errata (2002) *Physical Review B* **66**(24), 15 December 2002 (see also http://journals.aps.org/prb/issues/66/24#sect-errata

37. Fang, F.C., Steen, R.G. and Casadevall, A. (2012) Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences United States of America* **110**, 17028–17033
38. Seglen, P.O. (1997) Why the impact factor of journals should not be used for evaluating research. *British Medical Journal* **314**, 498–502
39. Simons, K. (2008) The misused impact factor. *Science* **322**, 165
40. http://www.vroniplag.de/ (Accessed April 2013)
41. Brenner, S. (1995) Loose end. *Current Biology* **5**, 568