Or: Why We Believe That Social Role Models Influence Mood (“Social Priming”) But Not That Homeopathy Or Telepathy Works

When we say something is “scientifically proven”, we usually mean that a series of conditions have been met, at least the following:

1. A phenomenon must be documented using methods that meet current scientific standards. It is best if this is an experimental method, i.e. one in which the researcher has created a control condition and shown that the effect of interest occurs in an experimental condition but not in the control condition. This is what animal, clinical or psychological experimental studies do, for example. Rats or mice, or humans, are then given something that controls do not receive, and a variable is measured from which the effect can be read.
2. The study must be published and thus available and testable for everyone. This term is not easy to define, especially today, where everyone can publish everything quickly and cheaply on the internet. As a rule, the operational definition of “published” is: a scientific result has been submitted to a peer-reviewed journal. The journal’s reviewers have examined the manuscript and found that the style, content and method conform to currently accepted standards and practices; that the method and results are appropriate to support the authors’ conclusions; and the journal’s editors felt the result was interesting enough for the journal’s readers. Furthermore, the journal must be judged by its standards by bibliographical professionals to be listed in relevant databases, and thus the articles published there.
   This enables people looking for scientific results to find them. A finding published somewhere that appears in a journal, for example, where no experts check the content, is not necessarily considered “scientifically published”, and mere availability on the internet is also not a criterion for “published” in a scientific sense. However, the reverse is also true: just because something has been published in a scientific organ does not necessarily mean that it is correct, acceptable or even proven. Peer review, as many studies have shown, can be deceptive. It is really nothing more than a social filter and can at best say what is currently acceptable, communicable and interesting and understandable to others. No more and no less [1]. It is particularly important to bear in mind that a whole forest of supposedly peer-reviewed online journals has sprung up in the last few years, which are nothing more than money printing machines for the publishers. This is because the authors sometimes pay considerable sums for the publication. Peer review exists only on paper, and it is easy to publish anything there, even obviously wrong results, as a study has shown [2]. The American librarian Beall maintains a list of such editors and journals on the

internet.

3. The claimed phenomenon must be robust and thus more than a random fluctuation in the sea of our collective perception. We know this from our everyday life: we think we have seen something special and then say to our partner: “Look, there is an ibex sitting in the grass”. Because we don’t have our glasses with us, we mistake a stone for an ibex. Our partner has better eyes and can clear up the perception error. It’s often like that in science, too. A research group or a researcher discover something, perhaps even with accepted methods according to the latest standard. He knows the method well and is therefore sure that it is not a mistake, an artefact or a perceptual illusion. His colleagues, who review the findings for a scientific journal, also agree with him. The study is published. Is it therefore already “scientifically proven”? Of course not. This just opens the discourse. Someone says: “Look, I have something interesting”. Just as I said to my wife in the summer: “Look, there’s an ibex sitting in the meadow”. But how does this first, perhaps even unique, finding become a scientific fact? That is what we want to deal with today. Because only through replication, if possible independent or even under extended and more difficult conditions, and through the subsequent communication processes in the scientific literature and the scientific community does a finding become a scientific fact. We will save the social side for another time. This time it is mainly about replication.

“Replication” means something like: another person is looking with his or her eyes at the same phenomenon. If he or she also sees an ibex and more other people do too, then most likely there really is an ibex sitting there. If I alone think I see an ibex, and everyone else says: that’s just a normal stone or a tree stump, then I’m probably wrong and have to go to the eye doctor. The founder of Gestalt therapy, Fritz Perls, used to say: “If one person says ‘You are a monkey’, you can ignore it. If two or three say it, then it’s time you bought a pack of peanuts.” [3] So that’s replication: the repeated stating of a fact, ideally using similar methods, but from different perspectives.

In science, replication means something very different depending on the discipline. In physics, for example, which has long since passed its crisis in this regard, experimental results are only taken seriously when they have been multiply replicated and tested, confirmed by several working groups and backed up in a collaborative process. Only when something has “5 sigma”, i.e. 5 times more than the standard error of a measurement, usually proven by multiple measurements, does one begin to recognize something as actually existing. Unfortunately, this is not the case in other fields like medicine and psychology. Just recently, Richard Horton, the editor of the Lancet medical journal, complained bitterly that far too many chance findings are traded as “scientific” in medicine: “What is medicine’s 5 sigma?” he asked, echoing the results of an international workshop

that addressed the replication problem [4].

With this, Horton has taken up an issue that has been simmering in medicine for a while: a few years ago, the Greek epidemiologist Ioannidis, who teaches at Stanford, caused a stir with an essay “Why most published research findings are false” [5]. The article has been viewed 1.4 million times and cited 1728 times so far (for comparison: the most cited article in the same journal in the same year has been cited 436 times). I consider this essay one of the most significant methodological contributions I have come across in a long time. In it, Ioannidis builds a simple argument: Authors want to be remembered primarily for positive discoveries. Research results that do not meet expectations are not prepared for publication, or are rejected by journal editors. Very often, there are also too few fully independent replications, and those that are available are done by the same team or by people repeating systematic mistakes that were made before.

Together with the tendency to leave negative results unpublished, which has meanwhile been proven by the example of various industry-sponsored studies – see part 8 and part 14 of my methodological series – this results in an explosive mixture: first, positive findings are trumpeted in a big way, the press helps a lot, because they are also interested in news and success stories (see part 7 of the series). Negative findings then have a very hard time being taken seriously or published, and if they are published, they usually appear in second- and third-tier journals that are read less often. Often they are even actively withheld. As a result, most people, even the experts, often only have the well known first, positive results in mind, ignoring the rest.

This is partly because we are all Bayesians (see Series Part 5): Such strong initial findings change our preconception, which subsequently acts like a filter. Therefore, we see mainly what we know and expect, and ignore the rest. This is the same in normal life as it is in science. For this reason, the Cochrane Collaboration, a network of interested scientists who systematically compile scientific knowledge in the clinical field, has also stipulated that a review should, if possible, include all so-called “grey” literature, including unpublished or poorly published literature. This includes diploma theses, master’s theses, internal reports, doctoral theses and the like, i.e. everything that is not covered by the citation databases. When this is done, there is often little left of the supposed clarity and scientific evidence. In 2007, for example, El Dib and colleagues examined 1016 randomly selected Cochrane reviews with the question of what we now really know with scientific certainty [6]. If one assumes that the scientists in the Cochrane Collaboration only address the urgent questions, then this finding is probably representative – and perhaps even flattering – for the overall state of scientific knowledge in the clinical-medical sector. Only 3.4% of all these reviews came to a really clear conclusion whether the intervention studied worked or not. In 2% it was clear that the intervention was harmful, in 1.4% of the interventions studied it

was clear that it was excellent. That’s just 14 of the 1016 interventions studied! The ratio of useful and effective to clearly harmful is unfavourable. And the rest? For 48% of the interventions studied, we still know too little and further research is needed. For another 5%, we can assume that further research will prove that the intervention is harmful. And for 43% – still – we can assume that the intervention is probably helpful, but just too little clear evidence.

Now, it is precisely this large grey area that is at stake. For here, a clearly negative replication or a study that came out negatively and is not published would be the element of knowledge that could change our view of things. The chances of there being unpublished results with a “negative” outcome in one area or another are relatively high, but impossible to measure. This is precisely why replications, and especially independent replications, are important. What does “replication” mean here and what “independent”? A few years ago, my colleague Stefan Schmidt took up this topic and wrote a review [7]. He found that although everyone talks about replication, everyone demands it, everyone thinks it exists, very few fields, at least in psychology and the social sciences, are really well replicated – mainly because replications are unpopular with researchers and unloved by journal editors. Moreover, research funders are not eager to fund replication either; they would rather be known for having helped bring new findings to light [8]. Replication means: a group of researchers takes a published experiment or another study, rebuilds the methodology and everything else, and tries to find another group’s results roughly as they were reported.

This can take various forms: In the narrowest sense, a replication can be exactly as reported. Hardly anyone does that. Because that is rather boring. So-called “conceptual replications” are more common. This means that the basic principle is replicated. So if someone reports that they were able to reduce headaches with a certain substance, let’s say with the administration of water for headaches [9], then in a follow-up study you try to pick up on this finding and perhaps make the design a little more careful: for example, to measure better, to observe longer, to characterize the patients’ illness better, to control the intervention better – for example, not just to say “drink more”, but to really give the patients more water and also to control that they drink more. Then it is often the case, as in our example, that the follow-up study produces much smaller or even different effects than the initial study [10]. In the clinical case, one then sees that an extension of the concept does not work so well. However, one does not necessarily know then whether the “negative” result is related to the fact that perhaps some crucial but perhaps not yet discovered parameter was changed in the replication study. In our example, for example, it could be that in the first study there happened to be more people who were naturally low drinkers and for whom the instruction to drink more actually had a positive effect, whereas in a

larger follow-up study this baseline difference disappears and so does the effect.

For this reason alone, replications are important, especially conceptual ones, to test the robustness of assumptions and concepts. But, as I said, replications are unpopular. You don’t get a Nobel Prize for confirming what others have found, and certainly not for disproving other people’s results. Prizes are awarded for the discovery of new findings. Of course, in order for them to be generally accepted, they have to have been replicated by other researchers. And only when they have really been replicated many times and confirmed as robust do they become generally accepted. At least that is the case in theory and often in practice. But Nobel Prizes in economics have also been awarded for models that are ingenious but then fail to prove themselves, as the last crisis in the financial sector has shown.

But let’s return to clinical research. There, replications are actually also necessary for us to accept an intervention as “effective”. Why is that? After all, it could be that an initial trial is simply a chance finding, a kind of random fluctuation. If that were so, then the next time we tried to detect the effect, the statistical variation would have to swing the other way, and we would see a null effect or even a negative effect. And on the third attempt, perhaps nothing at all, so that across all studies there is actually a zero effect at the end. Let’s assume that the effect found in the first study is a systematic positive effect that proves that the intervention – water drinking for migraine in our example – is successful. Then a follow-up study would have to show such a positive effect again, and a third study again. And even if there were a negative or only very weakly positive study, a positive and clearly different effect from zero would have to emerge across all studies. This could be demonstrated in a meta-analysis, such as the one published by the Cochrane Collaboration.

You can see immediately from this example that as soon as a negative study is suppressed, we distort the picture. That is why it is so important that all, but really all, findings, especially the negative ones, are published. Because we usually learn more from the negative findings than from the positive ones. As can be seen from the Cochrane Collaboration data briefly referenced above, this happens far less than one might think.

What Ioannidis argues theoretically for medicine [5], Horton has recently reiterated [4] is not empirically studied in the context of medicine. But psychology has definitely had a serious replication problem recently. For it is now empirically clear that less than half of all experimental findings published in psychology are even remotely replicable as reported in the literature [11]. Psychologist Brian Nosek became aware of the problem some time ago. He inspired a very large group of colleagues to replicate data from 100 studies from the most recent years of the most important psychological journals (“Psychological Science”, “Journal of Experimental Psychology”, “Journal of Personality and Social Psychology”) using

exactly the same methods. To do this, he sought out working groups that were qualified in the areas in question and were familiar with the method. They then contacted the first authors and had the method explained to them in detail and other details that perhaps could not be presented in the publication. The researchers really spared no effort to replicate the original findings as faithfully as possible. The effect sizes found were only half those originally published [12]. 97% of the original studies reported significant effects. But only 35% of the replications could find significant effects. Only 47% of the original effect sizes, or less than half, were within the confidence interval [13] of the replications.

In other words, less than half of the original reported data were statistically compatible with the replication results. Only 39% of the studies were subjectively judged by the researchers to be successful replications. The mean effect size of the original reported studies was r = 0.4, while the replicated one was r = 0.197, just half as large. The original p-values correlated negatively with the replicated ones, with r = -.327. This means: the larger and statistically more significant the original reported effects were, the more likely it was that they could not be replicated. Overall, a clear negative publication bias is visible. This means: in the original studies, negative results were not published, or the researchers kept trying until they had positive findings, or only presented those aspects from a study that made the whole finding appear positive.

What does this mean? Apparently, the tendency to suppress negative findings is also widespread in psychology. This is especially easy with experimental or cross-sectional studies. They are not very complicated to carry out in psychology – but also in medicine, pharmacology or basic biological research – if you know a method and have everything at hand. All psychology students have to collect hours as test subjects in order to be admitted to their exams and thus serve as human guinea pigs for their lecturers, who can try out one thing or another in this way. Then they can quickly do a new little experiment because they have a good idea for once. If nothing comes of it, the data is ignored. If you have a positive result – perhaps by chance or because you were sloppy – you shout hurray, open the champagne and send a manuscript to a journal. Another research group feels inspired, wants to replicate the result, but has a negative finding. This negative replication is rarely published. And so a hodgepodge of apparently positive scientific findings accumulates in the literature. As shown above: almost all original findings report effects found, correlations or meaningful differences.

So you see, “scientifically proven” does not mean “somewhere a positive result has been published.” It must also be guaranteed that this result has been replicated, ideally by another research group using their method. And the more controversial the finding, the

more stable the replication must be. Note the comparative “more controversial”: if data fit into a currently accepted model of thought or framework of theory, one or two replications will be considered sufficient evidence for the accuracy of the original finding.

Since the original finding is already very likely a selection from all possible data, including negative data, even in the case of “generally accepted” findings there is a high probability that one is making a mistake when one says something is “scientifically” proven. This is also the case with Ioannidis critical comment. However, when it comes to controversial areas, the research community will really expect very robust evidence, i.e. multiply replicated and, above all, independently replicated findings.

And this is also the real reason why *we* (meaning the mainstream of society and science) assume that antidepressants work, but homeopathy does not, and why we believe that “social priming” exists but not telepathy. From a purely objective point of view, the meta-analytically determined difference between homeopathic preparations and placebo preparations across all known, published and unpublished studies is statistically different from zero with an odds ratio of 1.53 or OR = 1.98 for the reliable studies. This means that a patient treated with homeopathy has twice the chance or at least 150% the chance of being cured as one treated with placebo [14]. I discussed this in my previous post in the series (Part 15).

But the theoretical understanding of what happens in homeopathy is not scientifically settled. “Social priming” fits into the current mainstream of social psychology. What is meant by this is that someone who is seen as a socially important person influences not only the behaviour of others through their behaviour, but also their feelings and cognitions [15]. The best-known study had someone walk past a group of participants at a tired pace and visibly exerting themselves, and subsequently found the participants to also walk more slowly, feeling tired and depressed. This fits into the paradigm of subliminal cognitive control, which is currently being studied everywhere.

The problem: The finding turned out to be unreplicable, and this experience prompted, among other things, Brian Nosek’s “Open Science Collaboration Project”. However, these negative findings never saw the light of day because journal editors doubted the competence of the experimenters or simply did not want to publish them. Because the finding is compatible with mainstream opinion, therefore, “we” collectively still believe in the possibility of social priming in a very broad sense, but are very sceptical about the possibility of homeopathic remedies having an effect, even though there are far more and far more robust findings.

Only: also with homeopathy, independent replication looks worse than Mathie’s meta-analysis suggests. This is because in the meta-analysis, many independent data were

processed together. The few attempts to perform replications within a research model were rarely successful. This is the reason why I myself came to the conclusion years ago that whatever works here does not work in a classical causal sense. One has to use a number of tricks, with which we also succeeded in the end in isolating homeopathic effects from those of placebo in experimental studies, several times [16]. Now others need to replicate the findings so that they can potentially contribute to making homeopathy look “scientifically proven”. But before this can happen, a plausible theory must also be found.

Similarly, with telepathy and psychokinesis. There are many positive findings [17]. There is also at least one theoretical model [18] and we have recently replicated a promising experimental paradigm, which we will report on. The findings are at least as statistically robust as many of the mainstream effects. But first, the theory is not widely accepted and acceptable [19], and second, independent, on-target replications of an experimental paradigm have not succeeded to date [20].

So what is “scientifically proven” is very elusive. Publications alone are not enough. They also often tend to be falsely positive, and we should always apply a fair amount of criticism. But even if positive data are available and published, there is always the question: are they replicated and replicable? And if this question is answered positively, we still have to ask: are the results socially acceptable, within the framework of the currently valid theories and models of thought? And only when all three are given is something “scientifically proven”. As we have seen, such judgements, especially when it comes to mainstream theories and effects, are often made hastily. And therefore something can be “scientifically proven” and still be wrong. And something else can be considered “scientifically questionable” and still end up being right.

To counteract publication bias, studies should be registered before they are conducted. This makes it easier to search afterwards for work that has been done but not published. This is now mandatory in clinical research, and most journals no longer accept publications of clinical trials that have not been registered in a common trial registry. In parapsychology, this policy of publishing all studies, including negative ones, has been implemented for at least 20 years [19]. But in experimental research, both in psychology and in medicine or pharmacology, there are only very cursory approaches to this.

Sources and references

1. Henderson, M. (2010). End of the peer review show. British Medical Journal, 340,

738-740.
Hopewell, S., Collins, G. S., Boutron, I., Yu, L.-M., Cook, J., Shanyinde, M., et al. (2014). Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. British Medical Journal, 349, g4145.
Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American society for Information Science and Technology, 64, 2-17.
Ritter, J. M. (2011). Impact, orthodoxy and peer review British Journal of Clinical Pharmacology, 72, 367-368.

2. Bohannon, J. (2013). Who’s afraid of peer review? Science, 342(6154), 60-65.
   Walach, H. (2015). Die Schrott-Schwemme und fünf Gründe, warum wir nicht dazugehören. Forschende Komplementärmedizin, 22, 152-154. http://www.karger.com/Article/Pdf/434665
3. Das ist ein überliefertes Diktum, das meine Gestaltlehrer John und Judith Brown, die Fritz Perls noch selber erlebt hatten, kolportiert haben.
4. Horton, R. (2015). Offline: What is medicine’s 5 sigma? Lancet, 385, 1380. http://dx.doi.org/10.1016/S0140-6736(15)60696-1
5. Ioannidis, J. P. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
6. El Dib, R. P., Atallah, A. N., & Andriolo, R. B. (2007). Mapping the Cochrane evidence for decision making in health care. Journal of Evaluation in Clinical Practice, 13, 689-692.
7. Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the Social Sciences. Review of General Psychology, 13, 90-100.
8. *We once submitted a proposal together with a Norwegian team, which was supposed to be about the replication of findings from placebo research. We did this against the background that our own results contradicted the known data (e.g. Walach, H., Schmidt, S., Bihr, Y.-M., & Wiesch, S. (2001). The effects of a caffeine placebo and experimenter expectation on blood pressure, heart rate, well-being, and cognitive performance. European Psychologist, 6, 15-25; and Walach, H., Schmidt, S., Dirhold, T., & Nosch, S. (2002). The effects of a caffeine placebo and suggestion on blood pressure, heart rate, well-being and cognitive performance. International Journal of Psychophysiology, 43, 247-260.) and placebo effects with placebo caffeine were much smaller than reported in the literature. The reviewers said, serious-minded: the project was not worthy of funding because it was merely a replication.*
9. Spigt, M. G., Kuijper, E. C., van Schayck, C. P., Troost, J., Knipschild, P. G., Linssen, V. M., et al. (2005). Increasing the daily water intake for the prophylactic treatment of headache: a pilot trial. European Journal of Neurology, 12, 715-718.

10. Spigt, M., Weerkamp, N., Troost, J., van Schayck, C. P., & Knottnerus, J. A. (2012). A randomized trial on the effects of regular water intake in patients with recurrent headaches. Family Practice, 29, 370-375.
11. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.
12. *I explained the term “effect sizes” in detail in my methodology Blog part 13. Briefly, it is a measure of the absolute size of an effect across studies that is comparable. There are measures for correlation, the correlation coefficient, for the difference between groups in continuous masses (“d”, or “g”) and for the difference between groups in dichotomous masses (e.g. odds ratio). They can all be transformed into each other. Here, a correlation coefficient was used ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). Normally, in social research, a correlation of variables in the range between r = .3 and .5 is to be expected: the correlation between intelligence and income lies roughly in this range, or between school grades and future income.*
13. *I have also explained the confidence interval in earlier chapters. It is an estimate of whether a value found lies within or outside a statistical range of variation within which, for example, 95% of values are expected to lie. Technically, confidence intervals are determined as follows: one calculates the standard error of the mean (SEM). This is the range of variation in the mean estimate that can be expected due to the statistical uncertainty of the estimate. It is defined as the standard deviation calculated around the empirical mean divided by the square root of the total number of observations. If this is very large, the standard error becomes small; if it is small, it is large. This therefore expresses the certainty of the estimate. This standard error of the mean can now be interpreted as a standard measure of a normal distribution constructed around the empirically found mean, with the standard error as a measure of dispersion. Then the confidence interval of the mean is defined as ±1.96*SEM. Incidentally, in this way one can also back-calculate the standard deviation from data in published studies if this is not specified.*
14. Mathie, R. T., Lloyd, S. M., Legg, L. A., Clausen, J., Moss, S., Davidson, J. R., et al. (2014). Randomised placebo-controlled trials of individualised homoeopathic treatment: sytematic review and meta-analysis. Systematic Reviews, 3(142).
15. Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. Journal of Personality and Social Psychology, 81, 1014-1027.
16. Möllinger, H., Schneider, R., Löffel, M., & Walach, H. (2004). A double-blind, randomized, homeopathic pathogenetic trial with healthy persons: Comparing two high potencies. Forschende Komplementärmedizin und Klassische Naturheilkunde, 11, 274-280.

Walach, H., Sherr, J., Schneider, R., Shabi, R., Bond, A., & Rieberer, G. (2004). Homeopathic proving symptoms: result of a local,non-local, or placebo process? A blinded, placebo-controlled pilot study. Homeopathy, 93, 179-185.
Möllinger, H., Schneider, R., & Walach, H. (2009). Homeopathic pathogenetic trials produce symptoms different from placebo. Forschende Komplementärmedizin, 16, 105-110.
Walach, H., Möllinger, H., Sherr, J., & Schneider, R. (2008). Homeopathic pathogenetic trials produce more specific than non-specific symptoms: Results from two double-blind placebo controlled trials. Journal of Psychopharmacology, 22, 543-552.
Eine Zusammenfassung bietet Walach, H., & Teut, M. (2015). Scientific provings of ultra high dilutions in humans. Homeopathy, in print.

17. Schmidt, S. (2014). Experimentelle Parapsychologie – Eine Einführung. Würzburg: Ergon.
18. Walach, H., von Ludacou, W., & Römer, H. (2014). Parapsychological phenomena as examples of generalized non-local correlations – A theoretical framework. Journal of Scientific Exploration, 28, 605-631.
Lucadou, W. v. (2015). The Model of Pragmatic Information (MPI). In E. C. May & S. Marwaha (Eds.), Extrasensory Perception: Support, Skepticism, and Science: Vol. 2: Theories and the Future of the Field (pp. 221-242). Santa Barbara, Ca: Praeger
19. *The philosopher Daniel Dennett once told Dick Bierman, a parapsychology researcher: if it turned out that such effects actually existed, he would commit suicide. This is, of course, to be taken rather facetiously, but it shows how high the emotional waves run*. Quote as personal communication, in Bierman, D. J. (2001). On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective world of physics? In P. Van Loocke (Ed.), The Physical Nature of Consciousness (pp. 269–292). New York: Benjamins, p. 269.
20. 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. This series of positive findings could not be replicated Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: three unsuccessful attempts to replicate Bem’s ‘retroactive facilitation of recall’ effect. PLoS One, 7(3), e33423. And a large replication of the PEAR Lab’s long-running experiments also failed: Jahn, R. G., Dunne, B. J., Bradish, G. J., Dobyns, Y. H., Lettieri, A., Nelson, R. D., et al. (2000). Mind/machine interaction consortium: PortREG replication experiments. Journal of Scientific Exploration, 14, 499-555.