Replication, replication, replication

Stuart J. Ritchie, Richard Wiseman and **Christopher C. French** with the opening contribution to a special on one of the cornerstones of scientific progress

Last year, Cornell social psychologist Daryl Bem had a paper published in the prestigious *Journal of Personality and Social Psychology* (JPSP) entitled 'Feeling the future' (Bem, 2011b). According to the nine studies described in the paper, participants could reliably – though unconsciously – predict future events using extrasensory perception. The findings proved eye-catching, with many major media outlets covering the story; Bem even discussed his work on the popular American TV show *The Colbert Report*.

The wide-ranging discussion of Bem's paper has raised questions regarding the limits of science, our current statistical paradigm, the policies of academic journal publishing, and what exactly a scientist needs to do to convince the world that a surprising finding is true. In this article we outline the 'Feeling the future' controversy, our part in it, and highlight these important questions about scientific psychology.

Some recent parapsychological research projects have taken a somewhat idiosyncratic approach to extrasensory perception by examining, for example, whether zebra finches can see into the future (Alvarez, 2010). In contrast, Bem adopted a more back-to-basics approach, taking well-worn psychological phenomena and 'time-reversing' them to place the causes after the effects. By far the largest effect size was obtained in Bem's final experiment, which investigated the 'retroactive facilitation of recall'. In this procedure, participants were shown a serial list of words, which they then had to type into a computer from memory in a surprise free recall test. After the test, the computer randomly selected half of the words from the list and showed them again to the participants. Bem's results appeared to show that this post-test practice had worked backwards in time to help his participants to remember the selected words – in the recall test they had remembered more of the words they were about to (randomly) see again.

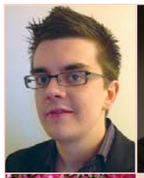
If these results are true, the implications for psychology – and society – are huge. In principle, experimental results could be confounded by participants obtaining information from the future, and studying for an exam after it has finished could improve your grade!

As several commentators have pointed out, Bem's (2011b) experiments were far from watertight – for instance, Alcock (2011) and Yarkoni (2011) have outlined numerous experimental flaws in the design. We won't describe these various issues here as they have been widely discussed in the blogosphere and elsewhere (see, for instance, Bem's, 2011a, response to Alcock). While many of these methodological problems are worrying, we don't think any of them completely undermine what appears to be an impressive dataset.

The 'Feeling the future' study has become a test case for proponents of Bayesian theory in psychology, with some commentators (e.g. Rouder & Morey, 2011) suggesting that Bem's seemingly extraordinary results are an inevitable consequence of psychology's love for null-hypothesis significance testing. Indeed, Wagenmakers et al. (2011a) suggest that had Bayesian analyses been employed, with appropriate priors, most of Bem's effects would have been reduced to a credibility level no higher than anecdotal evidence. Given that casinos are not going bankrupt across the world, argued the authors, our prior level of scepticism about the existence of precognitive psychic powers should be high.

Bem and colleagues responded (2011), suggesting a selection of priors which were in their view more reasonable, and which were in our view illustrative of the problem with Bayesian analyses, especially in a controversial area like parapsychology: Your Bayesian prior will depend on where you stand on the previous evidence. Do you, unlike most scientists, take seriously the positive results that are regularly published in parapsychology journals like the *Journal of the Society for Psychical Research*, or the *Journal of Parapsychology*? Or do you only accept those that occasionally appear in orthodox journals, like the recent meta-analysis of 'ganzfeld' telepathy studies in *Psychological Bulletin* (Storm et al., 2010)? Do you consider the real world – full as it is of the aforementioned successful casinos – as automatic evidence against the existence of any psychic powers? Your answers to these questions will inform your priors and, consequently, the results of your Bayesian analyses (see Wagenmakers et al., 2011b, for a response to Bem et al., 2011).

We reasoned that the first step towards discovering whether Bem's alleged effects were genuine was to see if they would replicate. As one of us has pointed out previously (Wiseman, 2010), the only definitive way of doing this is to carry out exact replications of the procedure in the original experiment. Otherwise, any experimental differences muddy the waters and – if the replications fail – allow for alternative interpretations and 'get-outs' from the original proponents. Recently, this argument was taken up with direct reference to Bem's experiment by LeBel and Peters (2011), who strongly argued in favour of more exact replications.



"the file drawer problem has been discussed for decades, but still no solid solutions appear to be forthcoming"

- **Stuart J. Ritchie** *(above left) is at the University of Edinburgh S.Ritchie-5@sms.ed.ac.uk*
- **Richard Wiseman** *(above) is at the University of Hertfordshire*
- **Christopher C. French** *(left) is at Goldsmiths, University of London*

Admittedly, carrying out exact replications of someone else's work is hardly the most glamorous way to spend your time as a scientist. But we are often reminded – most recently by an excellent article in the *APS Observer* (Roediger, 2012, and this issue) – that replication is one of the cornerstones of scientific progress. Keeping this in mind, the three of us each repeated the procedure for Bem's 'retroactive facilitation of recall' experiment in our respective psychology departments, using Bem's instructions, involving the same number of undergraduate participants (50) as he used, and – crucially – using Bem's computer program (with only some minor modifications, such as anglicising a few of the words). Either surprisingly or unsurprisingly, depending on your priors, all three replication attempts were abject failures. Our participants were no better at remembering the words they were about to see again than the words they would not, and thus none of our three studies yielded evidence for psychic powers.

We duly wrote up our findings and sent them off to the *JPSP*. The editor's response came very quickly, and was friendly, but negative. The journal's policy, the editor wrote, is not to publish replication attempts, either successful or unsuccessful. Add something new to the study (a 'replication-and-extension' study), he told us, and we may consider it. We replied, arguing that, since Bem's precognitive effect would be of such clear importance for psychology, it would surely be critical to check whether it exists in the first place, before going on to look at it in different contexts. The editor politely declined once more, as described by Aldous (2011), and Goldacre (2011).

While exact replications are useful for science, they're clearly not very interesting for top journals like JPSP, which will only publish findings that make new theoretical or empirical contributions. We are not arguing that our paper should automatically have been published; for all we knew at this point, it may have suffered from some unidentified flaw. We would, however, like to raise the question of whether journals should be so fast to reject without review exact replications of work they have previously published, especially in the age of online publishing, where saving paper is no longer a priority (Srivastava, 2011).

references

Alcock, J.E. (2011, 6 January). Back from the future: Parapsychology and the Bem affair. *Skeptical Inquirer*. Retrieved 6 March 2012 from http://tinyurl.com/5wtrh9q

Aldhous, P. (2011, 5 May). Journal rejects studies contradicting precognition. *New Scientist*. Retrieved 6 March, 2012 from http://tinyurl.com/3rsb8hs

Alvarez, F. (2010). Higher anticipatory response at 13.5 ± 1 H local sidereal time in zebra finches. *Journal of Parapsychology, 74*(2), 323–334.

Bem, D.J. (2011a, 6 January). Response to Alcock's 'Back from the future: comments on Bem'. *Skeptical Inquirer*. Retrieved 6 March 2012, from http://tinyurl.com/chhtgpm

Bem, D.J. (2011b). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. doi: 10.1037/a0021524

Bem, D.J., Utts, J. & Johnson, W.O. (2011). Must psychologists change the way they analyse their data? A response to Wagenmakers, Wetzels, Borsboom & van der Maas (2011). *Journal of Personality and Social Psychology, 101*(4), 716–719.

Goldacre, B. (2011, 23 April). Backwards step on looking into the future. *The Guardian*. Retrieved 16 March 2012 from http://tinyurl.com/3d9o65e

LeBel, E.P., & Peters, K.R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*(4), 371–379. doi: 10.1037/a0025172

Ritchie, S.J., Wiseman, R., & French, C.C. (2012). Failing the future: Three unsuccessful replications of Bem's 'retroactive facilitation of recall' effect. *PLoS ONE, 7*(3), e33423.

Roediger, H.L., III (2012). Psychology's woes and a partial cure: The value of replication. *Observer*. Retrieved 16 March 2012 from http://tinyurl.com/d4lfnwu

Rosenthal, R. (1966). *Experimenter effects in behavioural research*. New York: Appleton-Century-Crofts.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. doi: 10.1037/0033-2909.86.3.638

Rouder, J.N. & Morey, R.D. (2011). A Bayes-factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682–689.

Schlitz, M., Wiseman, R., Watt, C., & Radin, D. (2006). Of two minds: Sceptic-proponent collaboration within parapsychology. *British Journal of Psychology, 97*, 313–322. doi: 10.1348/000712605X80704

Srivastava, S. (2011, May 10). How should journals handle replication studies? [Web log post]. Retrieved 6 March 2012 from http://tinyurl.com/crb24a8

Storm, L., Tressoldi, P. & Di Risio, L. (2010). Meta-analysis of free response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136*(4), 471–485. doi: 10.1037/a001945

Wagenmakers, E.-J., Wetzels, R., Borsboom, D. & van der Maas, H.L.J. (2011a). Why psychologists must change the way they analyse their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432. doi: 10.1037/a0022790

After a couple of other submission attempts rapidly failed, we submitted our paper to the *British Journal of Psychology* (BJP). They have no automatic rejection policy for replication studies, and to our relief, sent our paper out for review. After a long waiting period, we heard back from two reviewers, the first of whom was very positive, and urged the editor to publish our paper. Conversely, the second reviewer – who found no technical faults with our procedures – recommended against publishing our paper as it was, because of one problem: the experimenter effect.

We know, argued the reviewer, that experimenter expectations can influence obtained results (e.g. Rosenthal, 1966), and this problem is worse in parapsychology, where sceptical experimenters might somehow communicate their scepticism to participants, thus affecting task performance. We found this implausible, given both the very brief and simple nature of Bem's experiment, and the lack of evidence that performance on a psychic memory task is influenced by prior scepticism. But one explanation for parapsychological experimenter effects is that sceptics might inadvertently use their own psychic powers to nullify the powers of their participants (e.g. Wiseman & Schlitz, 1998), rendering them unable to 'feel the future' for the duration of the study. Perhaps the editors found this explanation convincing, because they agreed with the reviewer, asking us to run a fourth study where a believer in psychic powers ran the experiments.

This latter explanation seemed to us to beg the question, to say the least. In trying to assess whether or not psychic powers exist, it is surely jumping the gun somewhat to expect that those unverified powers are influencing the assessment itself! Indeed, one of us had previously tested this exact phenomenon in parapsychology, asking whether believers obtain more positive results than sceptics who use the same methodology. The first experiments, on remote detection of staring, seemed to show that this was the case (Wiseman & Schlitz, 1998). However, the most recent study of this phenomenon (Schlitz et al., 2006) – the one with the largest sample size and the tightest experimental set-up, published in the BJP – showed no experimenter effects, with sceptic and believer both finding null results. Not exactly stunning evidence for the existence of unconscious bias, or indeed psychic interference, on the part of sceptics.

Most importantly, however, any experimenter effects should not have mattered – Bem, in his original paper, pointed out that his experimental setup reduced experimenter effects to a minimum (Bem, 2011b, p.16), as the computer program ran and scored the entire procedure itself, and the experimenter only greeted and debriefed participants. In two of our replication attempts we, like Bem, had research assistants do all the required jobs, and had no contact with the participants ourselves. This reviewer, then, seemed to have missed Bem's point – these were specifically intended to be replicable experiments, which could demonstrate precognitive effects to sceptics everywhere.

Since we didn't agree with the logic behind the believer-as-experimenter condition (we wonder – should being criticised for not believing in a particular phenomenon be dubbed the 'Tinkerbell effect'?), we withdrew our paper from the BJP, and decided to have one final try at submitting elsewhere. Happily for us, *PLoS ONE* accepted our article for publication (Ritchie et al., 2012), and the article is now available on their open-access website.

We would be the first to state that, even though we have had three failed replication attempts published, this does not rule out Bem's precognitive effects. Most obviously, we have only attempted to replicate one of Bem's nine experiments; much work is yet to be done. We've been made aware of a few other replication attempts through a study registry set up by Wiseman and Caroline Watt (tinyurl.com/bemreplic). Like trial registries in clinical medicine, researchers were asked to record their Bem replication attempts here, for their results to be included in a meta-analysis, which is currently in progress. The ideal, we believe, would be a prospective meta-analysis: researchers sit down together and plan, say, 10 studies in different laboratories of one effect, with sample sizes and analyses set in stone before they start. When the studies are complete, the data is pooled and analysed, and conclusions can be drawn that are (hopefully) acceptable to everyone involved.

While our experience relates to a rather outlandish area of psychology, the controversies, questions, and lessons we can draw from it apply to all publishing scientists (see Box 1). How many other researchers, we wonder, have tried and failed to replicate more sublunary effects and had their papers rejected, or not even attempted to publish them in the first place? Scientists' and scientific journals' well-known aversion to publishing null or negative results – the file-drawer problem (e.g. Rosenthal, 1979) – has been discussed for decades, but still no solid solutions appear to be forthcoming. We have a feeling the future will hold further debate over these vexed and important questions.

Questions arising

- To quote the title of Bem et al.'s (2011) response to Wagenmakers et al. (2011a): 'Do psychologists need to change the way they analyse their data?' Do we need to consider becoming Bayesians?
- How do we deal with experimenter effects in psychology laboratories?
- Should journals accept papers reporting replication attempts, successful or failed, when they themselves have published the original effect?
- Where should journals publish replication attempts? Internet-only, with article abstracts in the paper copy?
- Who should carry out replication studies? Should scientists be required to replicate their own findings?
- If a scientist chose to carry out many replications of other people's work, how would this impact his or her career?
- Should more outstanding and controversial scientific questions be subject to prospective meta-analyses?

Wagenmakers, E.-J., Wetzels, R., Borsboom, D. & van der Maas, H.L.J. (2011b). *Yes, psychologists must change the way they analyse their data: Clarifications for Bem, Utts, and Johnson (2011).* Unpublished manuscript.

Wiseman, R. (2010). 'Heads I win, tails you lose': How parapsychologists nullify null results. *Skeptical Inquirer, 34*(1), 36–39.

Wiseman, R. & Schlitz, M. (1998). Experimenter effects and the remote detection of staring. *Journal of Parapsychology, 61*(3), 197–208.

Yarkoni, T. (2011, 10 January). The psychology of parapsychology, or why good researchers publishing good articles in good journals can still get it totally wrong [Web log post]. Retrieved 6 March 2012 from http://tinyurl.com/694ycam

Replication: Where do we go from here?

A stellar cast of contributors offer their personal take on replication and possible progress

The need for new incentives

Scientists hope other laboratories will replicate their findings – any competent researcher should be able to re-do an experiment and produce the same effect, and independent replication is the benchmark of cumulative science. For good reason, we all dread publishing findings that fail to replicate. But replication failures happen for many reasons.

Statistically, replication failures *should* happen some of the time, even for direct replications of a real effect. Individual studies are samples of reality, noisy estimates of the truth. But such estimates vary, sometimes overestimating and other times underestimating the true effect. For underpowered studies – those testing small effects using relatively few participants – replication failures are statistically more likely. But so are false positives.

Given the existing bias shared by authors, reviewers and journals to publish only significant results (see Fanelli, 2012, for evidence that the problem is worsening and not restricted to psychology), coupled with investigator degrees of freedom that inflate significance levels artificially (e.g. Simmons et al., 2011), most striking new findings from underpowered studies are likely to overestimate the actual effect size. Some will be false positives, with little or no underlying effect. Consequently, a similarly powered replication attempt is likely to find a smaller, possibly non-significant result.

As a field, we could reduce the likelihood of false positives and enhance our ability to detect them if journals were more willing to publish both successful and failed replications. Perhaps more importantly, we need to rethink the publishing incentives in psychology to encourage multiple replication attempts for important new findings. Otherwise, science cannot adequately correct its mistakes or lead to a cumulative estimate of the true effect. The present publication model provides strong disincentives to replications. Some journals outright refuse to publish direct replication attempts, and a failed replication often incurs the wrath of the original researcher.

"if researchers know replication attempts will follow, they will be more cautious about publishing dubious data"

Daniel J. Simons *is at University of Illinois replicationjournal@gmail.com.*

My favoured solution to the incentive problem would be a journal of replication attempts. In my view, the peer-review process for such a journal should occur *before* a replication is attempted, based on a detailed method section and analysis plan. Once the design and analysis meet the approval of the original authors (or someone they designate), the results would be published regardless of the outcome. A crucial advantage of this approach would be a more constructive review process – because the original researchers want to see their findings replicated, they have an incentive to make sure such replications are conducted competently and accurately. And researchers would be guaranteed a publication should they choose to replicate another's research. For important findings, the field would benefit because many labs could contribute direct replication attempts, leading to a more accurate estimate of the true effect and to a more cumulative science (the measured effect from each replication attempt could be catalogued on a site like www.psychfiledrawer.org). Finally, this approach would carry an indirect benefit to the public perception of psychology – flashy findings based on underpowered studies garner tremendous coverage in the media. Such effects are less likely to replicate, and if researchers know replication attempts will follow, they will be more cautious about publishing dubious data in the first place.

The goal of scientific psychology should be to obtain an accurate estimate of the actual size of the effects we measure. The more direct replication attempts, the better the estimate of the true effect. In light of recent evidence for publication bias, investigator degrees of freedom in analysis, and the risk of false positives, any individual finding, especially one from an underpowered study, should be viewed with as much skepticism as a single failure to replicate. What the field needs are many direct replication attempts for each important finding. Only then can we be confident that an intriguing finding is more than a false positive or that a replication failure is more than a false negative.

Interested in a new journal like the one I have proposed? Please e-mail me.

Anderson, I., Pilling, S., Barnes, A. et al. (2009). *Clinical Practice Guideline No.90: Update: Depression in adults in primary and secondary care*. London: Gaskell/British Psychological Society.

Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 426–432.

Chalmers, I. (2002). Lessons for research ethics committees. *Lancet, 359*, 174.

Chan, A.W., Hrobjartsson, A., Haahr, M.T. et al. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association, 291*, 2457–2465.

Egger, M., Davey Smith, G. & Altman, D.G. (eds.) (2001). *Systematic reviews in health care: Meta-analysis in context* (2nd edn). London: BMJ Publishing.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904. doi: 10.1007/s11192-011-0494-7.

Fellows, L.K. & Farah, M.J. (2005). Is anterior cingulate cortex necessary for cognitive control? *Brain, 128*, 788–796.

Hartshorne, J.K. & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6*, 8.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. doi: 10.1371/journal.pmed.0020124

Isen, A.M. & Levin, P.F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and*

Twist, bend and hammer your effect

At least in my little corner of the world of psychological science, I see replications all the time. Often, for cognitive psychologists, replications of experiments are required for publication by editors in our most prestigious journals. To those who argue that a robust level of statistical significance is all one needs to assure replicability, I recall the aphorism (attributed to Confucius) that 'One replication is worth a thousand *t*-tests'.

Researchers should, whenever possible, replicate a pattern of results before publishing it. The phenomenon of interest should be twisted, bent and hammered to see if it will survive. If the basic effect is replicated under the exact conditions as in the original study, but it disappears when conditions are changed a bit, then the effect is real but brittle; the boundary conditions for obtaining the effect are rather narrow. That is not ideal, but is certainly worth knowing.

In the mid-1990s, Kathleen McDermott and I were collaborating on research, and we tried two rather risky experiments, ones that seemed likely to fail but that were worth trying. To our surprise, we found startling patterns of data in both procedures.

One case involved a technique for studying false memories in a list-learning situation in which the illusory memories seemed to occur nearly immediately and to be remarkably strong. After a first classroom pilot experiment, we conducted a proper second experiment that confirmed and strengthened our initial results. We started to write up the two experiments. However, we were still a bit worried about the robustness of the effects, so we continued experimenting while we wrote. We were able to confirm the results in new experiments (employing various twists), so that by the time the paper was published in the *Journal of Experimental Psychology* in 1995, we had several more replications and extensions ready to be written. Papers by other researchers, replicating and extending the effect, were also quickly published – no problem in getting replications published in this instance – and thus, within two years of its initial publication, anyone in my field who cared could know that our effect was genuine.

"Replicate your own work prior to publication. Don't let others find out that you are wrong"

Henry L. Roediger, III *is at Washington University in St. Louis. roediger@wustl.edu*

The second experiment we were excited about at that time did not have so happy a fate. After feeling confident enough to submit the research to be presented as a conference talk, we decided we needed to replicate and extend the effect before submitting it for publication. Altogether, we tried several more times over the next few years to replicate the effect. Sometimes we got results that hinted at the effect in our new experiments, but more often the results glared out at us, dull and lifeless, telling us our pet idea was just wrong. We gave up.

McDermott and I might well have published our initial single initial experiment as a short report. After all, it was well conducted, the result was novel, we could tell a good story, and the initial statistics were convincing. I would bet strongly we could have had the paper accepted. Luckily, we did not pollute the literature with our unreplicable data – but only because we required replication ourselves (even if the editors probably would not have – brief reports do not encourage and sometimes do not permit replication).

The moral of the story is obvious: Replicate your own work prior to publication. Don't let others find out that you are wrong or that your work is tightly constrained by boundary conditions. If there were a way to retract conference papers, we would have retracted that one. Most people don't count conference presentations as 'real' for the scientific literature, and our case provides another good reason for that attitude. At least we found out that our effect was not replicable before we published it.

The recent critical examination of our field, though painful, may lead us to come out stronger on the other side. Of course, failures to replicate and the other problems (fraud, the rush to publish) are not unique to psychology. Far from it. A recent issue of *Science* (2 December 2011) contained a section on 'Data replication and reproducibility' that covered issues in many different fields. In addition, an article in the *Wall Street Journal* ('Scientists' elusive goal: Reproducing study results', 2 December 2011) covered failures to replicate in medical research. So, failures to replicate are not only a problem in psychology. Somehow, though, when an issue of fraud or a failure-to-replicate occurs in (say) field biology, journalists do not create headlines attacking field biology or even all of biology. It seems that psychology is special that way.

This contribution is an edited version of a column for the APS Observer: see tinyurl.com/6vju949

The importance of replication in the field

A common retort to complaints about the infrequency of replication in psychology is to point out that, while 'strict replications' may be rare, 'conceptual replications' are not (see Nussbaum, this issue). Conceptual replications build on prior studies by employing common methods or similar operationalising of variables in new settings to examine a theory's limits, but many of these new settings involve artificial relations created in the laboratory or in an online experiment.

"psychological subfields differ markedly in the degree to which their laboratory results hold up in the field"

Gregory Mitchell, *School of Law, University in Virginia. greg_mitchell@virginia.edu*

What can we learn from a laboratory replication of an earlier laboratory study? We may learn that the first study contained errors, we may learn that a result is more or less strong than previously thought, or we may learn that a result generalises to a new laboratory situation. What we *cannot* learn is whether the result holds up in the field setting of ultimate interest to the theorist.

In a recent study that compiled meta-analyses comparing effect sizes found in the laboratory to those found in the field for a wide variety of psychological phenomena, I found that psychological subfields differ markedly in the degree to which their laboratory results hold up in the field (Mitchell, 2012). The subfield traditionally most concerned about the external validity of laboratory studies, industrial-organisational psychology, performed remarkably well: lab and field results were highly correlated (r = .89), and the magnitude of effects was

The role of conceptual replication

There is no substitute for direct replication – if you cannot reproduce the same result using the same methods then you cannot have a cumulative science. But *conceptual replication* also has a very important role to play in psychological science. What is conceptual replication? It's when instead of replicating the exact same experiment in exactly the same way, we test the experiment's underlying hypothesis using different methods.

One reason conceptual replication is important is that psychologists aren't always dealing with objectively defined quantities like 2mg of magnesium; instead we have to operationalise conceptual variables in a concrete manner. For instance, if we want to test the effects of good mood on helping behaviour, what counts as a good mood and what counts as helping behaviour? These are not objectively defined quantities, so we have to decide on something reasonable. For example, in the 1970s Alice Isen found that people were far more likely to help someone who had dropped some papers after they had found a dime in a phone booth (Isen & Levin, 1972). But we have even more confidence in this result now that it's been conceptually replicated: helping has been found to increase not only after finding money, but after reflecting on happy memories, doing well on a test, or receiving cookies. Numerous different ways of measuring helping behaviour have been used as well. In other words, even if nobody had ever tried to directly replicate the original research, the conceptual replications give us confidence that the underlying hypothesis – that a positive mood increases helping behaviour – is correct.

"conceptual replications give us confidence that the underlying hypothesis is correct"

Dave Nussbaum *is Adjunct Assistant Professor of Behavioral Science at the Booth School of Business at the University of Chicago davenussbaum@gmail.com*

In the recent debate (see tinyurl.com/cfkl2gk and tinyurl.com/7ffztux) about the failure to replicate John Bargh's finding that priming the elderly stereotype leads people to later walk more slowly as they leave the experiment, there has been a great deal of misunderstanding because of confusion between direct replication and conceptual replication. While it is true that there have been few direct replications of that finding, the underlying hypothesis that 'activating a stereotype in the perceiver without the perceiver's knowledge would increase the perceiver's own behavioural tendencies to act in line with the content of that stereotype' has been replicated many times. It was even conceptually replicated in the very same series of published studies in which the original 'slow walking' study was published, as well as by independent researchers at different universities around the world. Taking these conceptual replications into account, most psychologists are not nearly as troubled by a single failure to replicate the result as it may appear that they should be.

Social Psychology, 21, 384–388.
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F. & Baker, C.I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience 12(5)*, 535–540.
Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*, 109–117.
Moonesinghe, R., Khoury, M.J. & Janssens, C.J.W. (2007). Most published research findings are false – but a little replication goes a long way. *PLoS Medicine, 4*, e28.
Poldrack, R.A., Fletcher, P.C., Henson, R.N. et al. (2008). 'Guidelines for reporting an fMRI study. *NeuroImage 40(2)*, 409–414.
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*, e33423.
Schulz, K.F., Altman, D.G. & Moher, D. (2010) , CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c332.
Shallice, T. & Cooper, R.P. (2011). *The organization of mind*. Oxford: Oxford University Press.
Simmons, J., Nelson, L.D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.
Song, F., Parekh-Bhurke, S., Hooper, L. et al. (2009). Extent of publication bias in different categories of research cohorts: A meta-analysis of empirical studies *BMC Medical Research Methodology, 79*. doi: 10.1186/1471-2288-9-79.
Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – Or vice versa. *Journal of the American Statistical Association, 54(285)*, 30–34.
Sterling T.D., Rosenbaum W.L. & Weinkam, J.J. (1995). Publication decisions revisited – The effect of the outcome of statistical tests on the decision to publish and vice-versa. *The American Statistician, 49*, 108–112.
Turner, E.H., Matthews, A.M., Linardatos, E. et al. (2008) Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252–260.
Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105–110.
Wicherts, J.M., Bakker, M. & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE, 6(11)*, e26828. doi:10.1371/journal.pone.0026828

opinion

similar in the lab and field. Social psychology, on the other hand, performed much more poorly: over 20 per cent of the results from the laboratory changed directions in the field, the correlation of lab and field results was significantly lower (r = .53) , and there were larger disparities in effect sizes between the lab and field. In short, if we consider the record of a theory using only laboratory tests of that theory, theories from I-O psychology will look less impressive than they should, and theories from social psychology will look more impressive than they should.

Another noteworthy result from my study was the relative dearth of meta-analyses comparing laboratory and field effects outside the areas of I-O and social psychology. Although it is possible that field studies are common in other psychological subfields but have escaped quantitative synthesis, the more likely explanation is that field replications are relatively rare for many areas of psychology.

We should not pin our hopes for a mature science on conceptual replications in the lab. As I found for a number of theories in social psychology, successful laboratory replications may be positively misleading about the size and direction of an effect. Replication in the field, not the laboratory, is crucial to the development of reliable theory.

Longstanding misunderstandings about replication

Ritchie, Wiseman and French's failed attempt to replicate one of my nine experiments on precognition (see p.356) has been widely interpreted by the popular media and some psychologists as convincingly falsifying my claim that I have produced evidence of psi (ESP). This coverage has revealed many longstanding misunderstandings about replication – held even by those who should know better.

The first misunderstanding is the sheer overestimation of how likely it is that any replication attempt will be successful, even if the claimed effect is genuine. Tversky and Kahneman (1971) posed the following problem to their colleagues at meetings of the Mathematical Psychology Group and the American Psychological Association: 'Suppose you have run an experiment on 20 subjects and have obtained a significant result which confirms your theory (z = 2.23, p < .05, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?' The median estimate was .85, with 9 out of 10 respondents providing an estimate greater than .60. The correct answer is less than .5 (approximately .48).

"it takes a long time for enough replications to accumulate to draw any firm conclusions"

Daryl J. Bem *is Professor Emeritus of Psychology at Cornell University*
d.bem@cornell.edu

Second, it takes a long time for enough replications to accumulate to draw any firm conclusions. Wiseman set up an online registry for those planning to replicate any of my experiments. As he noted: 'We will carry out a meta-analysis of all registered studies…that have been completed by 1 December 2011.' The deadline was only a few months after my article appeared, and by then only three experiments other than those by Ritchie et al. had been reported. Two of them had successfully reproduced my original findings at statistically significant levels, a fact known to Ritchie et al., but not mentioned in the literature review section of their report. In any case, firm conclusions are clearly somewhat premature at this point.

In mainstream psychology it takes several years and many experiments to determine which variables influence the success of replications. Consider, for example, the well-known 'mere exposure effect', first brought to the attention of psychologists by Robert Zajonc in 1968: Across a wide range of contexts, the more frequently humans or other animals are exposed to a particular stimulus, the more they come to like it. Twenty years later, a meta-analysis of over 200 mere exposure experiments was published, confirming the reality of the effect. But that same meta-analysis reveals that the effect fails to replicate on simple stimuli if other, more complex stimuli are presented in the same session. It fails to replicate if too many exposures are used, if the exposure duration is too long, if the interval between exposure and the assessment of liking is too short, or if participants are prone to boredom. As a result, the meta-analysis included many failures to replicate the effect; several of them actually produced results in the direction opposite to prediction. A virtually identical situation has now arisen in the context of currently popular 'priming' experiments in cognitive and social psychology.

Finally, I believe that some major variables determining the success or failure of replications are likely to be the experimenters' expectations about, and attitudes toward, the experimental hypothesis. Psychologists seem to have forgotten Robert Rosenthal's extensive and convincing demonstrations of this in mainstream psychology during the 1960s. The same effect has been observed in psi experiments as well. Ironically, Wiseman, a psi-skeptic, has himself participated in a test of the experimenter effect in a series of three psi experiments in which he and psi-proponent Marilyn Schlitz used the same subject pool, identical procedures, and were randomly assigned to sessions. Schlitz obtained a significant psi effect in two of the three experiments whereas Wiseman failed to obtain an effect in any of the three. Thus, it may be pertinent that Ritchie, Wiseman and French are well-known as psi sceptics, whereas I and the investigators of the two successful replications are at least neutral with respect to the existence of psi.

The existence of such experimenter effects does not imply that psi results are unverifiable by independent investigators, but that we must begin to systematically include the experimenters' attributes, expectations and attitudes as variables.

The case of neuroimaging

The field of neuroimaging deserves to be placed in a special class where questions of replication are concerned.

The most common current form, fMRI, has only been in widespread use for about 15 years, and methods are still in development. Somewhat understandably, therefore, many older papers include what we now acknowledge are clear statistical flaws, and the community is debating how we should deal with such a residue of potentially false results (e.g. see bit.ly/H80N5S and bit.ly/Ha0Bcg). Of more concern, though, is that while on average the field is improving, with nascent guidelines agreed upon (Poldrack et al., 2008), common flaws still regularly arise (Kriegeskorte et al., 2009).

I recently wrote a blog on these matters (bit.ly/HCxDLM), with the comments section hosting a constructive debate between many leading neuroscientists, demonstrating that the neuroimaging community are broadly aware of these problems, and keen to move forward.

There are several issues specific to neuroimaging. Unlike many behavioural psychology experiments, neuroimaging researchers rarely replicate their own study prior to publication. Of course, we all should be doing this, but when a standard fMRI study may total £50,000 (including salaries) and involve 6–12 months work, many believe that internal replication is too great a burden to bear.

A second problem concerns the extent of knowledge required, both to generate a sound fMRI design, and particularly to analyse the results. Although an entire year of study of the techniques would make considerable sense, the intense time pressures in academia make this option unpalatable. In addition, only the largest neuroimaging departments have the infrastructure to support regular courses and a decent collective understanding of current methods. Smaller centres are thus more likely to publish flawed neuroimaging papers, because of more limited access to technical knowledge.

A third issue relates again to the complexity of fMRI, though in a more disturbing way. Because of the lack of maturity of such complex techniques applied to massive datasets, as well many varieties of analysis, it is worryingly easy

A view from the media

When there's a failure to replicate a study that was initially covered in the press, should that failure to replicate be reported in the media too? There are two difficulties with doing so. One is that a failure to replicate is very different from deliberate fraud. In *All in the Mind* we covered Stapel's research on stereotyping in one series and then, in the next, after it had been discovered to be fraudulent, discussed the case and how the discipline of psychology handles fraud. Fraud is an issue that's interesting even if a listener hadn't heard of the specific research before. Failure to replicate implies no wrong-doing; so, although it interests me, you could argue that to the general audience at which a programme like mine is aimed, it's less newsworthy.

The second difficulty is that we could report that a previous finding might be incorrect – and in the recent case of the failure to replicate John Bargh's work we'd be reporting that it's possible that reading words associated with age doesn't make you walk more slowly – but this would be reporting that something *hasn't* happened. This is tricky unless it the finding was very well known in the first place. Outside the world of psychology most people will have never have heard of Bargh's priming experiment. I can think of very few studies in the history of psychology that are well known enough for a failure to replicate to get a lot of coverage. Perhaps if someone demonstrated that dogs can't be conditioned to salivate at the sound of bell then we might see 'Pavlov got it wrong' headlines,

"very few studies in the history of psychology are well known enough for a failure to replicate to get a lot of coverage"

Claudia Hammond *is presenter of All in the Mind on BBC Radio 4 and author of Time Warped: Unlocking the Mysteries of Time Perception. claudia.hammond@bbc.co.uk*

but it would need to be a study at that level of fame. For less famous studies the space afforded by blogs seems perfect for this kind of discussion.

Then there's the suggestion that journalists should wait until a study has been replicated before they report it in the first place. But with the scarcity of publications of replications you could wait a very long time for this happen, by which time the findings wouldn't be new, which is after all what 'news' is supposed to be. This leaves journalists having to trust in the peer review process, but the best journalists do stress that a study is the first of its kind and should examine it critically. To be fair, I have the luxury of working on a specialist programme where there's time to discuss the limitations of a study, and unlike journalists working on general programmes I don't have to compete for air-time with every other subject on earth to get an editor's attention.

It's true that the more staggering a result is, the more likely it is to get reported. The media know that we all enjoy a discovery that's counterintuitive. These are the stories that people retweet and tell their friends about. They're the same studies that journals send out to the media, knowing they'll get coverage. And of course psychologists are interested in the counterintuitive too. If lay theories about human behaviour were all correct, much psychological research would be redundant.

But could some of the most surprising and therefore newsworthy findings also turn out to be those that are hardest to replicate? Without a full database of replications it's hard to know. But in recent years psychology has been very successful at increasing the coverage of research in the media, often pushing the stories that are most entertaining and curious. Here it's not so much a lack of replications that's the problem, but the tendency that I've observed for two views of psychologists to predominate – therapists who know what you're thinking or people who produce research that's entertaining yet largely trivial. For psychology to contribute to society in all the ways that it could, the next challenge is to find a way to communicate the whole range of research out there.

for neuroimagers, who should know better, to bend the rules, and publish striking results, which aren't statistically valid – but in a way that could be hidden from the reader, and subtle enough to get past the average reviewer. Although there are many solid neuroimaging papers, even now others get published where, to the trained eye, these tricks are relatively apparent.

"it is worryingly easy for neuroimagers to bend the rules, and publish striking results, which aren't statistically valid"

Daniel Bor *is at the Sackler Centre for Consciousness Science at the University of Sussex. D.bor@sussex.ac.uk*

Consequently, I discount the majority of neuroimaging papers I read. There was a reasonable consensus for this perspective on my blog, with Nancy Kanwisher claiming that 'way less than 50%' and Matthew Brett estimating that only 30 per cent of neuroimaging papers would replicate.

In most cases, such publications are detrimental to the field: they imply invalid techniques are acceptable, prolong wrong theories, and may waste considerable time and money, as other scientists fail to replicate a result that was never real in the first place. If the study is clinically relevant, the damage may be far more critical, for instance leading to wrong medical advice or inappropriate treatment.

So how do we improve the situation? Better education should be a priority: many researchers need to learn more methods and common pitfalls, while large neuroimaging departments should further help educate those within and outside their borders, with courses and online material.

There should be a culture of greater transparency, with any neuroimaging paper refraining from any selectivity about methods or results. Ideally, papers should publicly release all raw imaging data, so that another lab can validate the results using different analyses. There are steps towards this (e.g. www.fmridc.org), but far more could be achieved.

Finally, there should be a second cultural shift, so that the community values rigour above dramatic results. The gatekeepers – the manuscript referees, editors and journals themselves – have the most vital role to play here. For certain issues, there is a broad consensus about what steps are invalid, so that a checklist of minimum standards could be written by a few key neuroimaging experts, and this could be adopted by journals, and upheld by their editors and those participating in peer review.

Undoubtedly the neuroimaging field is gradually gaining sophistication and rigour. But as a community, we need to be embarrassed by our slow progress and do all we can to improve matters.

The journal editor's view

It is commonplace to observe that the norms of scientific publishing prioritise striking, novel findings. The problem with this emphasis is well known. Given the potential for bias in the interpretation of experimental results, the 'file-draw problem', and simply the nature of probability, a large number of published scientific findings – perhaps a majority (Ioannidis, 2005) – will be false. A simple remedy for this problem is also well known. If a published scientific finding can be replicated, it is much more likely to be true (Moonesinghe et al., 2007).

To some extent, developments in the field of scientific publishing are addressing this problem. Internet-only journals – without limited printed pages to constrain the number of articles that can be accepted – can play a big role. For instance, *PLoS ONE* recently published Ritchie et al.'s (2012) failure to replicate one of the studies reported in Bem's (2011) highly surprising paper on precognition. Ritchie et al.'s paper was previously rejected from other journals on the basis of its status as a replication attempt. But the editorial criteria of *PLoS ONE* are clear. It will publish any paper that appropriately describes an experiment performed to a high technical standard, described in sufficient detail, and drawing appropriate conclusions that are supported by the data. That is not to say that *PLoS ONE* will publish anything. All submissions to *PLoS ONE* are subject to peer review, as with any other journal. Authors are offered a waiver if they are unable to pay the publication fee, but editors and reviewers are blind to the fee status of each submission. However, highly subjective considerations of novelty, impact and importance are not considered in making editorial decisions, and no manuscript would be rejected on the grounds of being a 'mere' replication. Alongside journals such as *PLoS ONE*, internet innovations such as PsychFileDrawer.org provide another repository for the publication of replication attempts.

"internet-only journals – without limited printed pages to constrain the number of articles – can play a big role"

Sam Gilbert *is research fellow at UCL's Institute of Cognitive Neuroscience and an academic editor at PLoS ONE sam.gilbert@ucl.ac.uk*

These developments in scientific publishing certainly help. But even with greater availability of outlets for the publication of replication attempts, scientists must still be willing and able to conduct the relevant experiments. And with the current reward structure of science, most researchers have little incentive to spend time on replication studies, most of which will generate relatively little benefit in career progression. This is perhaps a more intractable problem.

If we are to conduct replication studies, how precisely should we attempt to repeat the earlier work? While attempts to replicate an earlier study as closely as possible may be relatively rare, perhaps the majority of published studies involve an incremental advance drawing on previous work. Studies of this kind that corroborate as well as extend earlier results might be thought of as 'conceptual replications'.

It could be argued, however, that such conceptual replications have limited value. Insofar as earlier results are not supported, this may be attributed to methodological differences with the earlier study. But insofar as such conceptual replications do support earlier findings, typically they only do so in a rather narrow way, based on a limited set of tasks and methodologies that tend to be repeated from one study to the next. This is particularly problematic in a field such as cognitive neuroscience, which has diverse intellectual origins and is still undergoing rapid methodological development. Shallice and Cooper (2011) describe 'a golden rule of cognitive neuroscience – that all empirical brain-based findings must currently rely on somewhat flaky inference procedures as far as cognitive conclusions are concerned' (p.156). Given this golden rule, one can be much more confident in a conclusion that has been derived from several methodologies, each with their own distinct sets of assumptions. Consider, for example, the enormous functional imaging literature on anterior cingulate cortex and error detection/conflict monitoring, against the striking failure to detect an impairment in such processes in patients with damage to this brain region (e.g. Fellows & Farah, 2005). I would argue that cognitive neuroscience would benefit from both an increase in attempts at direct replication, attempting to replicate a study as precisely as possible, and also an increase in bold conceptual replications involving as diverse a set of tasks and methodologies as possible. Both of these might be more informative than methodologically narrow conceptual replications.

Where's the data?

Recently, researchers in a number of fields have been raising the alarm, worried that little of what appears in the pages of our journals is actually replicable (i.e. true). This is not the first time such concerns have been raised, and editorials calling for improved methodological and statistical techniques in order to increase the proportion of reported results that are true go back decades. On the flip side, though they are less vocal about it, it appears that many in our field are sceptical about the pervasiveness of the problem, taking the view that certainly some results do not replicate, but probably not very many, and in the long run, science is self-correcting.

What is remarkable about this discussion is how very nearly data-free it is. A handful of intrepid researchers – e.g. John Ioannidis and colleagues, looking at the medical sciences – have managed to demonstrate worryingly low replicability rates in some targeted sub-literatures, and recent surveys (e.g. Hartshorne & Schachner, 2012) hint that this may extend to psychology, but no comprehensive data exist. If a reform of our practices were to be enacted, we would remain ignorant as to whether it improved replicability rates or even harmed them.

"By making it possible to calculate a quantitative replicability index for use alongside impact factor, journals that successfully ensure reliability will get credit for their efforts"

Joshua K. Hartshorne *and* **Adena Schachner** *are at Harvard University gameswithwords@gmail.com*

The current state of affairs is so familiar that it may bear reminding of just how strange it is. As a field, we demand precision in our theories and rigour in their empirical validation. The fact that the data needed to test the theory do not exist is not an accepted reason for deciding the matter in their absence. The claim that current methodologies are or are not sufficient for guaranteeing reliability is an empirical claim, and there is nobody in a better position to address the validity of these claims than us. Evaluating empirical claims is our business. Nor is there anybody to whom it matters more: low replicability threatens our ability to build on previous findings and make cumulative scientific progress.

We have argued that the first, key step is to systematically track which studies have and have not been replicated, and have discussed how this might be done (see PsychFileDrawer.org, for an alternative but related approach). First, this provides the raw material for any systematic study of the factors that lead to more replicable research. Just as important, it changes the incentive structure. Instead of languishing in a file drawer or in an obscure corner of some journal, replication attempts will be noted, highlighted and used, making it more worthwhile to conduct and report replication attempts. By making it possible to calculate a quantitative replicability index for use alongside impact factor, journals that successfully ensure reliability of reported results will get credit for their efforts.

This is not a trivial task: conducting, reporting and indexing replication attempts will require changes in our research practices, not to mention time and money. We hope our proposal leads to discussion and progress toward ensuring replicability of findings. The difficulty of the problem does not mitigate the necessity of solving it, and we believe that the scientific community can take action to resolve questions of replicability.

Making it quick and easy to report replications

What proportion of the statistically significant empirical findings reported in psychological journals are not real? That is, for how many is there actually no phenomenon in the world resembling the one reported?

The intention of our basic statistical practices is to place a small cap on the proportion of errors we are willing to tolerate in our literature, namely the alpha level we use in our statistics – typically 5 per cent. The alpha level is the maximum probability for a published study that an effect at least as large as the one measured could have occurred if there were actually no effect present. This ought to place an upper bound on the proportion of spurious results appearing in our literature. But it doesn't work.

One of the biggest reasons is *publication bias*: the fact that scientists are much more likely to publish significant findings than they are to report null effects. There are a host of reasons for this practice, some of them reflecting confusions about statistics, some reflecting editorial tastes that may have some merit. We also know that researchers who find non-significant results tend not to publish them, probably because they are both difficult to publish and there is little reward for doing so.

To see how this publication bias can lead to a literature infested with error, imagine that 1000 investigators each conduct a study looking for a difference, and a difference actually exists in, say, 10 per cent of these studies, or 100 cases. If the investigations have a power of .5, then 50 of these differences will be discovered.

Of the 900 studies performed looking for effects that do *not* exist, 5 per cent or 45 will succeed. The result, then, will be that 45 out of 95 significant results (47 per cent) will be type 1 errors. As Joe Simmons and colleagues (2011) recently pointed out in *Psychological Science*, hidden flexibility in data analysis and presentation is likely to inflate the rate of type 1 errors still further.

The harm done by publication bias has been recognised since at least 1959, when Theodore Sterling canvassed 294 psychology papers and found that 286 of them reported a positive result. Thirty-seven years later, Sterling re-evaluated the literature and concluded that little had changed. Indeed, the problem appears to be getting worse, and not just in psychology. The title of a recent paper in *Scientometrics* by Fanelli (2012) declares 'Negative results are disappearing from most disciplines and countries' based on an analysis of various scientific fields.

The problem is so bad that after a series of calculations involving some additional considerations, John Ioannidis concluded in a 2005 paper that 'most published research findings are false'. Although the estimate of error depends on several unknowns regarding the proportion of effects being looked for which actually exist, statistical power, and so forth, Ioannidis' conclusion is, unfortunately and disturbingly, quite reasonable.

Given that the incentives for publishing null results are modest, we must make it easy and quick for scientists to report them. The traditional system of journal article cover-letter writing, submission, rejection, repeat until sent out for review, wait for reviews, revise, and write rejoinders to reviewers is far too consuming of researchers' time.

To provide a quick way to report replication studies, we have created, together with Bobbie Spellman and Sean Kang, a new website called PsychFileDrawer.org. The site is designed specifically for replications of previously published studies, as this allows the reporting process to be quick. In the case of exact replications, for instance, researchers can simply indicate that their methodology was identical to the published study. When their method differs somewhat, they can report only the differences. Researchers are invited to report the results in as much detail as they can, but we believe even those who simply report the main effect and associated statistics are making a valuable contribution to the researcher community.

In addition to allowing researchers to report their method and results, a forum is provided for users to discuss each report. The website further allows users to vote on 'What are important studies that your field of Psychology gives credence to, but which – as far as you know – have not been replicated in any published follow-up work?' Each registered user is allowed up to three votes. As of this writing, the study with the most votes is entitled 'Improving fluid intelligence with training on working memory', which was published in *PNAS* in 2008. As the list solidifies, we hope it will encourage investigators to conduct replications of these studies and report the results on the site.

"The site is designed specifically for replications of previously published studies"

Alex O. Holcombe *is in the School of Psychology, University of Sydney. alex.holcombe@sydney.edu.au*
Hal Pashler *is in the Department of Psychology, University of California San Diego*

The most novel feature of PsychFileDrawer is an 'article-specific networking tool' designed for users who are timid about posting their findings or who simply wish to connect with others interested in a particular published finding about which nothing is yet posted to PsychFileDrawer. With this feature, users register their interest in learning about unpublished replication attempts relating to a particular study; whenever other users express interest in the same study, the website automatically puts these users in touch with each other via e-mail so they can discuss their experiences and hopefully post their results on the site.

The website is still in beta testing and we continue to add new features. We hope readers will visit and provide suggestions for how it might be improved.

Share your data

Researchers typically hold strong beliefs, can be quite ambitious, and seldom respond dispassionately to *p*-values from their own analyses. Moreover, statistical analysis of psychological data involves numerous decisions (e.g. on how to deal with outliers, operationalise dependent variables, select covariates, etc.) that are not always carved in stone. This provides researchers with considerable room to manoeuvre when analysing their data. These issues are ignored in most textbooks.

Marjan Bakker and I recently documented an alarmingly high prevalence of errors in the reporting of statistical results in psychology. In nearly half of papers we scrutinised we encountered at least one inconsistent statistical result. Errors typically aligned with the researcher's hypothesis and so may introduce bias. We employed a superficial test of statistical accuracy, so it would be interesting to conduct a full reanalysis of the raw data. Unfortunately quite a few researchers are reluctant to share their data for independent replication of their analyses, even when they are ethically obliged to do so. My colleagues and I also found that researchers who were unwilling to share their data report more inconsistent statistical results and more results that appear contentious. It is tempting to accuse these researchers of a lack of integrity, but we should not forget that data archiving too is not part of most textbooks.

"researchers who were unwilling to share their data report more results that appear contentious"

Jelte M. Wicherts *is an associate professor at the Department of Methods and Statistics at Tilburg University. J.M.Wicherts@uvt.nl*

As researchers, we are accustomed to dealing with observer biases by blinding procedures during data collection. It is rather naive to pretend that statistical analyses are completely immune to similar biases. Whenever feasible (ethically) we should start publishing the data to supplement research papers. This enables replication of statistical outcomes (even by sceptics) and superior means for detecting misconduct. It also opens the door to future use of the data and debates on what science is all about.

Evaluating psychological treatments

The past 50 years have seen a transformation in the evaluation of psychological interventions with the randomised controlled trial establishing psychological interventions on a par with pharmacological treatments in the treatment of mental disorders (e.g. Anderson et al., 2009).

Replication of the results of early clinical trials has been and remains central to the proper evaluation of all healthcare interventions. Good scientific practice demands replication, but there are challenges in achieving this. These include, the ethical behaviour of health professionals, publication bias, conflicts of interest, trial design and reporting, and the systematic evaluation of datasets of specific interventions.

Chalmers (e.g. 2002) has argued that health service professionals conducting research have an overriding ethical responsibility to any research participant in line with those governing professionals when treating patients. This imposes responsibilities not only for the conduct of clinical trials but also for their publishing and reporting. Chalmers argues that a rigorous application and monitoring of these ethical responsibilities is one of the best ways to ensure effective research practice. Readers may want to hold this in mind when considering the issues discussed below.

Publication bias remains a major problem, and the consequences can be significant leading to inappropriate and harmful treatment. This is perhaps most obvious in pharmacological interventions (e.g. Turner et al., 2008), but it is wrong to assume that the problem does not apply to psychological interventions. Evidence suggests that this problem lies not with the decisions of journals (they publish just as many negative as positive trials – Song et al., 2009) but with researchers who do not seek to publish their work. Lack of transparency about conflicts of interests may be crucial here. Conflicts of interest may also impact on the accurate reporting of the outcomes, with evidence that over 50 per cent of trials do not report the initial primary outcome (Chan et al., 2004). Again this can lead to the use of inappropriate and harmful treatments.

"Replication of early clinical trials has been and remains central to the proper evaluation of all healthcare interventions"

Stephen Pilling *is the Director of CORE (Centre for Outcomes Research & Effectiveness) at University College London, and the Director of the National Collaborating Centre for Mental Health s.pilling@ucl.ac.uk*

Considerable efforts have been made to address these problems, including the development of procedures for trial reporting (the CONSORT Statement: Shultz et al., 2010) and the pre-registration of trials protocols (including the predetermined primary outcome) without which publication of trials will no longer be possible. The use of systematic reviews and meta-analysis (Egger et al., 2001) to increase precision about the effects of treatment can also help address some of these problems, particularly where careful attention is devoted to the quality of included studies and the comparators used in trials.

Clinical trials are costly and make significant demands on participants; badly conducted or unpublished studies not only waste resources, they also do harm. A number of methodological developments have emerged to address this issue, but they do not remove the responsibility of all health professionals to uphold the highest standards not only in their clinical practice but also in their research practice.

Opinion special – Have your say

As ever, we would like to hear your views on this topic. E-mail your letters for publication to psychologist@bps.org.uk or post to the Leicester office address.

I would also like to know what you think of the 'opinion special' more generally. This is the first we have tried in this format: a large collection of brief and personal responses to recent events, accompanying a main article. I think this allows us to respond in a timely fashion, building on our role as a forum for communication, discussion and controversy while fulfilling the main object of the Society's Royal Charter: 'to promote the advancement and diffusion of a knowledge of psychology pure and applied'. But do you agree? Let me know on jon.sutton@bps.org.uk.

If feedback is positive, we would look to include more of these opinion specials, and would welcome your suggestions for topics likely to engage and inform our large and diverse audience.

Dr Jon Sutton Managing Editor