The Quest For Replicability

July 30, 2015

Roberta Klatzky

If it weren't for an attempted replication, Michael LaCour might have gotten away with it. LaCour, who is alleged to have fabricated data for a groundbreaking study on how canvassers can change attitudes toward gay marriage published in — and now retracted from — the journal *Science* in December 2014, was caught when David Broockman and Joshua Kalla, graduate students at the University of California, Berkeley, found red-flag irregularities in the results as they tried to replicate and extend LaCour's compelling findings.

However, outright fraud is by far one of the least common reasons a study may not be reproducible; failed replications have myriad causes, including how the replication is run and interpreted. But psychological scientists can also take steps in their original empirical studies to increase the reproducibility of their results. Researchers dedicated to improving reproducibility gathered at the 27th 2015 APS Annual Convention in New York City to discuss the obstacles to making psychological research more reproducible and brainstorm how to best circumvent those obstacles.

APS Treasurer Roberta Klatzky of Carnegie Mellon University discussed the many causes of data variability and failures to replicate published findings. These issues can arise from sampling variability and measurement noise just as readily as from sloppiness or questionable research practices. Often, the root cause of this variability is at least partly attributable to stochastic factors, which must be considered in assessments of a given phenomenon's replicability.

Simine Vazire

Understanding whether an individual replication failed because of one of these inevitable or external factors or because of an actual methodological flaw is especially vital to properly contextualizing the alluring and counterintuitive findings that comprise much of the most controversial psychological science literature today.

Simine Vazire, an associate professor of psychology at the University of California, Davis, noted that because psychological findings in particular are often so compelling and exciting, we *want* to believe them to be true, but this enthusiasm can ultimately damage the field's credibility. Vazire reviewed recent meta-analytic findings indicating that many false positive results exist within the literature of psychological science and presented several suggestions to address this issue. She offered philosophical recommendations for scientists based around the general idea of increased skepticism, such as treating all results as preliminary, being prepared to "unbelieve" even the most cherished findings in the literature (especially one's own), and keeping in mind that confidence in preliminary results should be rare.

On the more practical side, Vazire advocated for reducing motivations to *p* hack, the (often unconscious) bias that compels researchers to tweak data selection or statistical analyses until a result reaches statistical significance. This bias can be reduced by preregistering experiments as well as increasing disclosure about measures not analyzed or reported in a manuscript. She believes that the scientific community should incentivize more transparent science and distinguish between preplanned analyses and exploratory analyses — both of which have value — along with treating exploratory results with a healthy amount of skepticism.

Sean MacKinnon

On a logistical level, a valid and reliable statistical analysis is a vital component of reproducible research. For researchers who conduct binary hypothesis testing, statistical power — the probability that the hypothesis test correctly rejects the null hypothesis when the alternative hypothesis is true — is key to interpreting and assessing confidence in experimental data. Power is based on a complex equation incorporating several parameters, and most researchers tend to focus on one in particular: sample size. But, Sean Mackinnon explained, there are other ways to increase power.

Mackinnon, an instructor at Dalhousie University in Canada, showed how reducing the mean square error and/or increasing the variance of the predictor variable can increase statistical power; like sample size, these elements can be manipulated through the experimental design. Mackinnon explained how controlling for confounding variables and using a repeated measures design can help reduce the mean square error. He also explained how increasing the variance of the predictor variable in correlational designs can improve statistical power.

Zoltan Dienes

Zoltan Dienes of the University of Sussex, United Kingdom, addressed another issue related to binary hypothesis testing: interpreting nonsignificant results, typically defined as those having a *p* value of

greater than 0.05. It may seem intuitive to interpret a nonsignificant result as support for the null hypothesis, but Dienes noted that a *p* value of greater than 0.05 can actually mean that there is either no effect *or* simply no evidence for or against the effect in this particular test. He then showed how researchers can use the Bayes factor — a value used in Bayesian analysis that serves as an alternative to hypothesis testing — to distinguish evidence supporting the null hypothesis from evidence of nothing at all. Dienes argued that, despite common belief, determining power does not achieve this goal.

The speakers' diversity of approaches to increasing the reproducibility of findings revealed that there are changes at virtually every stage of research, from experimental design to data analysis to publication, which can be adopted to improve reproducibility. Implementing these changes across the entire domain of psychological science is undoubtedly a daunting task, but Vazire, for one, is optimistic about the direction that things are moving. "We're asking a lot of our field," she said. "We're asking people to change their individual behavior, we're asking journals to change their submission guidelines, and things are happening faster than I expected."