THE TRUTH SQUAD

In its drive to expose weaknesses in science, an up-and-coming research group doesn't mind stepping on some toes

By **Erik Stokstad**, *in Tilburg, the Netherlands*
Photography by **Manon Bruininga**

In August 2011, Diederik Stapel, a prominent psychologist and a dean at Tilburg University here, confessed to faking data for dozens of papers over 15 years. As part of an internal investigation, Marcel van Assen, a psychologist in the university's Department of Methodology and Statistics, spent months looking into Stapel's data, methods, and results. The scope of the fraud was staggering, but just as alarming as the fabricated data, Van Assen says, were the flawed analyses, rife with statistical problems, that Stapel had performed. The fact that all his papers had been approved by co-authors and published in respectable journals meant psychology had a larger problem, Van Assen says. "I thought, holy shit, this is not a characteristic just of Stapel or Tilburg."

Around the same time, a psychologist with a strong interest in the same issues joined his department. Jelte Wicherts, previously an assistant professor at the University of Amsterdam, wanted to find out why "smart researchers do stupid things with statistics," as he puts it. The two hit it off, and have since created what psychologist Brian Nosek, director of the University of Virginia's Center for Open Science (COS) in Charlottesville, calls "one of the leading groups" in metascience, the study of science itself.

Metaresearchers investigate how scientists operate, and how they can slip off the rails. "We've seen things that we felt were not OK," Wicherts says. "The first way to deal with it, that's our conviction, is to study these things." They're motivated by the desire to make science better, although Van Assen is drawn to the detective work. "What I like most is to solve puzzles," he says. By scrutinizing the problems, metaresearchers aim to help scientists do more robust research. Thanks to a €2 million grant from the European Research Council (ERC), for example, the Tilburg

The metaresearch group at Tilburg University investigates how scientists operate, and how they can slip off the rails.

group is starting to build software that could help researchers explore data with less risk of bias.

That Wicherts's and Van Assen's center was built on the ruins of Stapel's deceit may seem like poetic justice, but straight-up scientific fraud is only a minor topic for the group. Its main focus is questionable research practices, such as massaging data and selective reporting of statistical tests. These misdemeanors don't get a scientist fired, but they do help explain why so many findings are hard to reproduce—the "reproducibility crisis" that has gripped not just psychology, but many areas of basic biology and clinical medicine.

For scientists who find themselves in the crosshairs, the experience can feel bruising. Several years ago, the Tilburg group—now more than a dozen faculty members and students—unveiled an algorithm, dubbed statcheck, to spot potential statistical problems in psychology studies. They ran it on tens of thousands of papers and posted the troubling results on PubPeer, a website for discussion of published papers. Some researchers felt unfairly attacked; one eminent psychologist insinuated that the group was part of a "self-appointed data police" harassing members of the research community.

Van Assen and Wicherts say it was worth stepping on some toes to get the message across, and to flag mistakes in the literature. Members of the group have become outspoken advocates for statistical honesty, publishing editorials and papers with tips for how to avoid biases, and they have won fans. "I'm amazed that they were able to build that group. It feels very progressive to me," says psychologist Simine Vazire of the University of California, Davis, a past chair of the executive committee of the Society for the Improvement of Psychological Science (SIPS).

The work by the Tilburg center and others, including SIPS and COS, is beginning to have an impact. The practice of preregistering studies—declaring a plan for the research in advance, which can lessen the chance of dodgy analyses—is growing rapidly (see story, p. 1192), as is making the data behind research papers immediately available so others can check the findings. Wicherts and others are optimistic that the perverse incentives of careerist academia, to hoard data and sacrifice rigor for headline-generating findings, will ultimately be fixed. "We created the culture," Nosek says. "We can change the culture."

TILBURG MIGHT SEEM an unlikely place for academic innovation. The city, 90 kilometers south of Amsterdam, was once a center of the Dutch textile industry; after the woolen mills shut down, insurance and transportation businesses sprung up. Tilburg University was founded in 1927 as the Roman Catholic University of Commerce, but it is now best known for its social sciences departments, which fill a 10-story concrete building. Housed on a floor near the top, the metaresearchers have an expansive view of a forested 18th century park.

One morning this May, Wicherts, an energetic and talkative 42-year-old, was making a cup of strong coffee as he related how he became involved in metascience. When he was a Ph.D. student in psychology in the mid-2000s, it was an open secret that many findings were irreproducible, he says, but scientists feared that discussing this would cast the whole field into doubt. Then in 2005, John Ioannidis, now co-director of Stanford University's Meta-research Innovation Center in Palo Alto, California, published a provocative essay, "Why Most Published Research Findings Are False." It argued that science suffers from an epidemic of small studies that try to detect modest effects, poorly designed by researchers "in chase of statistical significance." Wicherts, inspired by the paper's clarity and bravery, calls it a watershed event for psychology.

The metaresearch group, co-led by Jelte Wicherts (left), created a major stir with statcheck, an algorithm that Michèle Nuijten (right) helped develop.

PHOTO: MANON BRUININGA

WICHERTS HAD his own encounter with poor scientific practices during his Ph.D. work on the rise of intelligence scores over generations. Curious about the impact of unusual data points on statistical analyses, he and his colleagues asked the authors of 141 recent papers for their data, so that they could reanalyze them. To their surprise, 73% of the authors didn't reply or said they were not willing or able to share the data, even though the journals that published the studies stipulated they should. Wicherts dropped the study but described the experience in *American Psychologist*. The 2006 paper was an early alert about the importance of "open data," Vazire says. "We need something better than 'data available upon request.'"

Wicherts now co-leads the group with Van Assen, who had focused on cognitive and mathematical psychology before he was drawn into the Stapel investigation. Its highest profile—some would say most notorious—project is statcheck. The algorithm, developed by Michèle Nuijten, then a Ph.D. student at Tilburg, together with Sacha Epskamp of the University of Amsterdam, scours papers for statistical results reported in standardized formats, then examines them for errors, like a mathematical spell checker. When statcheck scanned 30,717 papers published between 1985 and 2013, it found a "gross inconsistency" in one out of eight. Most of these results purported to be statistically significant, but in fact

were not, Nuijten and colleagues reported in 2015. Statcheck can't distinguish between honest errors and deceit, but "it's not unimaginable that people do this on purpose," says Nuijten, now an assistant professor.

When Tilburg Ph.D. student Chris Hartgerink posted statcheck's evaluations of 50,000 psychology studies on PubPeer, some scientists were furious. The most vocal critics complained that statcheck had claimed an error when in fact it wasn't able to properly scan their statistics, which had been correct. "Statistical graffiti," one called it. In a column, Susan Fiske of Princeton University, a past president of the American Psychological Association, decried a trend of "methodological terrorism." (Fiske removed that term, which caused a tempest on social media after her draft leaked.) The German Psychological Society called for a moratorium on statcheck.

In retrospect, Nuijten says she would have written fuller explanations for the PubPeer posts, in less brusque a style. But Vazire says she handled the controversy with aplomb, showing the kind of communication skills that can "win hearts and minds" in the campaign to improve psychology. Ultimately, says Eric-Jan Wagenmakers, a statistician at the University of Amsterdam, "I think it had a really positive effect. And wouldn't have if they had done it more subtly."

Still, the episode showed how sensitive people can be to criticism. "If you make it personal," Wicherts says, "then they can't admit the errors." Rather than calling out individuals, he now believes that metaresearchers should highlight the problems, encourage best practices, and create a system where errors can be caught before publication. One sign of progress is that two psychology journals now run submissions through statcheck.

OTHER ATTEMPTS to take a hard look at psychology's practices created a backlash as well. Van Assen and two students participated in the Reproducibility Project: Psychology, a large, 4-year collaboration organized by Nosek and COS that managed to replicate only 39% of the findings in 100 studies (*Science*, 28 August 2015, p. 910). Some senior psychologists quickly pushed back; Harvard University's Daniel Gilbert and three co-authors, for instance, criticized the collaboration's methods and their "pessimistic conclusions." (A new project, published last month in *Nature Human Behaviour*, replicated 62% of experiments reported in recent papers in *Science* and *Nature*.)

Wicherts says some researchers fear such critiques could jeopardize funding or breed mistrust in science. But it's not the group's job to protect psychology's reputation, he says. And the Tilburg studies have shaken the illusion that scientists are more objective than most people, underscoring that most researchers have a poor ability to look objectively at data and overestimate the statistical power of their studies.

With his ERC grant, Wicherts plans to develop software that will help psychologists avoid the temptation to test many hypotheses and only report those that have a significant p-value. The behavior, called data dredging, or HARKing, for "hypothesizing after results are known," generates apparently well-founded results that often can't be reproduced. Following an approach used in particle physics and other fields, the software will reveal a random sample of the data that researchers have gathered, letting them explore and generate hypotheses. Then it will deliver another random selection for rigorously testing those hypotheses. "I think there could be a role for this," says psychologist Dorothy Bishop of the University of Oxford in the United Kingdom, though she suspects it will require large data sets.

> **"WE'VE SEEN THINGS THAT WE FELT WERE NOT OK. THE FIRST WAY TO DEAL WITH IT, THAT'S OUR CONVICTION, IS TO STUDY THESE THINGS."**
> —**Jelte Wicherts**, Tilburg University

Wicherts's main effort right now is leading a project in which scientists at five universities take a fresh look at the data behind 200 studies, repeating the analyses in many ways to find out whether the authors chose to report specific results that matched their hypotheses. "I think we'll find quite a lot of biases in place," Wicherts says.

SINCE WICHERTS'S DISCOVERY as a student that most psychology researchers don't share their data when asked, he and others have pushed for change. Even today, only 10% of newly published psychology papers have data available, but the "open data" ethos is gaining traction in psychology and beyond. As an incentive, 41 journals now allow authors to slap a virtual open data "badge" on a paper; after *Psychological Science* adopted the practice, the share of open-data papers rose from 3% to 39% in just over a year. (Similar badges exist for "open materials" and study preregistration.)

Like other metaresearchers, the Tilburg group has itself adopted a far-reaching open-data policy: It shares data, code, and materials, except when issues of copyright, privacy, or ownership are involved. "It's a much harder way of working—it slows you down—but it makes you more thoughtful and confident," Bishop says. Hartgerink even posted versions of chapters of his Ph.D. online as he wrote them. "I share almost everything as I do it," he says. One risk of posting entire data sets is that competitors might analyze them and come up with new findings first. Although that's arguably good for the field as a whole, some labs worry that younger scientists who have yet to make their name might lose a chance to publish a significant finding.

The reproducibility push has other potential downsides for younger researchers. Studies with respectable statistical power take major work and might fizzle. For her Ph.D., Paulette Flore, now an assistant professor at Tilburg, studied whether reminding girls of their gender hurts their performance on math tests, an effect found in many smaller studies. Flore set up the largest study of the effect ever—involving more than 2000 students at 21 Dutch high schools—only to find no evidence for it. "In earlier days, her career would have ended," Wicherts says. "Now, you do the best you can, and let the chips fall where they may. I think this is the future."

At the moment, however, negative findings "won't land you a fancy job," says Daniël Lakens, an applied statistician at Eindhoven University of Technology in the Netherlands. For that to change, science will need to put more value on good ideas, solid methods, and broadly collaborative work, and less on high-profile publications and citations. "There's a corrupting influence of our current incentive structure," Bishop says. "The pressure on younger people not to do research in a reproducible way can be quite intense."

Some of the young scientists at Tilburg are pessimistic that the situation will improve anytime soon. "At the current pace, it's going to be 2100 before things are really different," Hartgerink says. Wicherts believes avoiding bad practices in research will pay off for individual scientists in the long run. "Keep in mind that these better methods empower the truth, and that this ultimately promotes scientific progress and highlights your contributions." ■

The truth squad

Erik Stokstad

Science **361** (6408), 1189-1191.
DOI: 10.1126/science.361.6408.1189