

Stanford Report, November 16, 2015

# Stanford researchers uncover patterns in how scientists lie about their data

When scientists falsify data, they try to cover it up by writing differently in their published works. A pair of Stanford researchers have devised a way of identifying these written clues.

BY BJORN CAREY

Andrey Popov/[Shutterstock](#)



Stanford communication scholars have devised an 'obfuscation index' that can help catch falsified scientific research before it is published.

Even the best poker players have "tells" that give away when they're bluffing with a weak hand. Scientists who commit fraud have similar, but even more subtle, tells, and a pair of Stanford

researchers have cracked the writing patterns of scientists who attempt to pass along falsified data.

The work, published in the *Journal of Language and Social Psychology*, could eventually help scientists identify falsified research before it is published.

There is a fair amount of research dedicated to understanding the ways liars lie. Studies have shown that liars generally tend to express more negative emotion terms and use fewer first-person pronouns. Fraudulent financial reports typically display higher levels of linguistic obfuscation – phrasing that is meant to distract from or conceal the fake data – than accurate reports.

To see if similar patterns exist in scientific academia, Jeff Hancock, a professor of communication at Stanford, and graduate student David Markowitz searched the archives of PubMed, a database of life sciences journals, from 1973 to 2013 for retracted papers. They identified 253, primarily from biomedical journals, that were retracted for documented fraud and compared the writing in these to unretracted papers from the same journals and publication years, and covering the same topics.

They then rated the level of fraud of each paper using a customized "obfuscation index," which rated the degree to which the authors attempted to mask their false results. This was achieved through a summary score of causal terms, abstract language, jargon, positive emotion terms and a standardized ease of reading score.

"We believe the underlying idea behind obfuscation is to muddle the truth," said Markowitz, the lead author on the paper. "Scientists faking data know that they are committing a misconduct and do not want to get caught. Therefore, one strategy to evade this may be to obscure parts of the paper. We suggest that language can be one of many variables to differentiate between fraudulent and genuine science."

The results showed that fraudulent retracted papers scored significantly higher on the obfuscation index than papers retracted

significantly higher on the obfuscation index than papers retracted for other reasons. For example, fraudulent papers contained approximately 1.5 percent more jargon than unretracted papers.

"Fraudulent papers had about 60 more jargon-like words per paper compared to unretracted papers," Markowitz said. "This is a non-trivial amount."

The researchers say that scientists might commit data fraud for a variety of reasons. Previous research points to a "publish or perish" mentality that may motivate researchers to manipulate their findings or fake studies altogether. But the change the researchers found in the writing, however, is directly related to the author's goals of covering up lies through the manipulation of language. For instance, a fraudulent author may use fewer positive emotion terms to curb praise for the data, for fear of triggering inquiry.

In the future, a computerized system based on this work might be able to flag a submitted paper so that editors could give it a more critical review before publication, depending on the journal's threshold for obfuscated language. But the authors warn that this approach isn't currently feasible given the false-positive rate.

"Science fraud is of increasing concern in academia, and automatic tools for identifying fraud might be useful," Hancock said. "But much more research is needed before considering this kind of approach. Obviously, there is a very high error rate that would need to be improved, but also science is based on trust, and introducing a 'fraud detection' tool into the publication process might undermine that trust."

## MEDIA CONTACT

Bjorn Carey, Stanford News Service: (650) 725-1944,  
[bccarey@stanford.edu](mailto:bccarey@stanford.edu)