Data Falsificada: The Gino Affair.

By Lorenz Goette and Guillem Riambau

[This post was written originally in English with the intent of being published in the Spanish economics blog Nada es Gratis].

Between June 17 and June 30, 2023, the Data Colada team posted a series of four posts in which they showed compelling evidence of fraud in recent published articles by (former) Professor of Business Administration Francesca Gino (Harvard Business School, HBS henceforth). Data Colada's research was carried out in 2021. Based on their concerns, HBS launched an internal investigation which resulted in Prof. Gino being put on unpaid administrative leave for two years, as her Harvard official webpage announces.

The news has had an impact beyond the usual academic and twitter circles —most notably, perhaps because the fraud was ironically conducted on studies on honest behavior. The Guardian, The New York Times, The Boston Globe, The Washington Post, The New Yok Post, Business Insider, NPR, and, less surprisingly, The Chronicle of the Higher Education (also here), VOX (the blog, not the party), or The Atlantic have reported on the issue, among many others. To the best of our knowledge, among the Spanish general media, only Expansión has echoed the news of fraud, although La Vanguardia did publish an article in 2017 on Gino's research entitled "Si te ha engañado una vez, lo volverá a hacer, según Harvard" [If they lied to you once, they'll do it again].

In an unexpected turn of the wheels, Francesca Gino announced in her Linkedin account in early August that she "had no choice but to file a lawsuit against Harvard University and members of the Data Colada group, who worked together to destroy my career and reputation" as she "[has] never, ever falsified data or engaged in research misconduct of any kind". The lawsuit is asking for no less than $25 million. (For the curious reader: it is interesting to check how supportive the vast majority of responses to her Linkedin post are, compared to the reactions we have observed in the academic world).

In what follows, given the constraints on space, we focus on two of the allegations of fraud and Francesca Gino's responses in her lawsuit. The logic of events in the other two posts is very similar, and they may be found here (Data Falsificada (Part 3): "The Cheaters Are Out of Order") and here (Data Falsificada (Part 4): "Forgetting The Words").

Data Falsificada (Part 1): "Clusterfake" (posted on June 17, 2023).

In this post, the authors discuss Study 1 in the PNAS 2012 article by Shu, Mazar, Gino, and Ariely, "Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end" (spoiler: article retracted on September 13, 2021). Study 1 was run at the University of North Carolina (UNC) in 2010. Gino, who was a professor at UNC prior to joining Harvard in 2010, was the only author involved in the data collection and analysis of Study 1.

The dataset was available at Open Science Framework. There are 101 observations in the dataset, and they are almost (but not quite) sorted by: first, the condition assignment (0 = control, 1 = sign-at-the-top, and 2 = sign-at-the-bottom), and second (within each condition assignment), a participant ID identification variable called "P#". 8 of the participant IDs are either duplicated or out-of-sequence in a suspicious manner. The Data Colada team argues that "[t]here is no way, to our knowledge, to sort the data to achieve this order. This means that these rows of data were either moved around by hand, or that the P#s were altered by hand. We will see that it is the former."

Data Colada notes that the data also include an Excel file of the same data that contains formulas. One subsidiary file that the Excel file uses to produce the spreadsheet is calcChain.xml. CalcChain "retains the order in which formulas were initially entered into the spreadsheet" regardless of where cells are ultimately moved. Using CalcChain, Data Colada shows that 6 observations that appear on top of each other in the dataset are out-of-sequence. Additionally, the P#s of the rows surrounding the places where CalcChain identifies the initial positions skip the exact position that would have been moved, further bolstering the idea that the observations were moved.

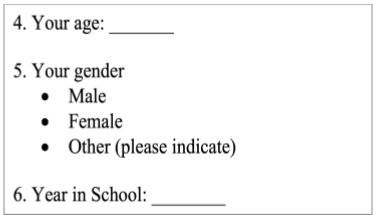
The 8 observations are critical to the paper's result, because, as Data Colada notes, "are all among the most extreme observations within their condition, and all of them in the predicted direction. All of this "strongly suggests" (they argue) that the observations were modified to yield a particular set of results: to be precise, "[w]ith just n = 8 they produce t(6) = 21.92, with a miniscule p-value".

Points 234–247 (pages 52-55) in Gino's complaint address this blog post. Most of her arguments rely on the fact that the original data was collected on paper (e.g., 247: "Data Colada knew that Study 1 was conducted on paper, with data collected on paper in 2010. Data Colada also knew that the fact the study had been conducted on paper provided a reasonable and plausible reason why the data (…) was not sorted in any particular order"). While accurate, these arguments fail to tackle the key issue pointed out by Data Colada: that the data was rearranged within the Excel spreadsheet *after* the initial input of observations.

Data Falsificada (Part 2): "My Class Year Is Harvard" (posted on June 20, 2023).

In this post, the authors discuss Study 4 in the Psychological Science 2015 article "The Moral Virtue of Authenticity: How Inauthenticity Produces Feelings of Immorality and Impurity" (Gino, Kouchaki, and Galinsky).

All participants were Harvard students. When collecting sociodemographic information, participants were asked to provide their year in school (Q6, see screenshot of the original posted materials).



4. Your age: _______

5. Your gender

- Male
- Female
- Other (please indicate)

6. Year in School: ________

Reasonable responses to Q6 are “Junior”, “junior”, “3”, “class of 2016”, “’16”, etc. What seems less reasonable as an answer is “Harvard”, which we find *up to 20 times* in the dataset. As the Data Colada authors note, “It is difficult to imagine many students independently making this highly idiosyncratic mistake (…) Moreover, and adding to the peculiarity, [these 20] students’ responses are all within 35 rows (450 through 484) of each other in the posted dataset”.

All these observations provide results in line with the predictions of the authors: those who by random assignment to treatment were predicted to give “high” responses did so, and those who were assigned to the condition associated with an expectation of “low” responses gave “low” responses. As the post notes, “the effect for the 'Harvard' observations is significantly larger than the effect for the non-Harvard observations (p < .000001). This strongly suggests that these 'Harvard' observations were altered to produce the desired effect”, or, more precisely, this suggest that this being a coincidence is very unlikely.

What does Gino have to say about this? Points 248–253 (pages 55-56) in her complaint address this blog post. Her most relevant point is 250: “Data Colada, as experienced behavioral scientists, knew that participants frequently respond to a survey to obtain payment due for their participation (as study participants) and may rush through questions, sometimes more than once to get paid, and use extreme values as their answers. It is widely known in behavioral science that participants in online studies at times provide poor-quality data by answering surveys without the attention they require.” None of Gino’s other 5 points responding to this Data Colada post addresses why 20 random participants who filled the responses virtually at random did so in the same direction, when lack of attention would predict all sorts of errors for those observations.

Gina’s legal complaint is exactly 100 pages long. It ends with the “Prayer for Relief” (pages 95 onwards) in which she requests that (page 97) “On the seventh cause of action for defamation against Defendants Simonsohn, Nelson, and Simmons, [the Data Colada team] damages of at least $25 million, in an amount to be determined at trial, including economic losses, lost career opportunities, reputational harm, emotional distress and punitive damages, costs and attorneys’ fees (…).”

We consider Gino’s decision to take this to the courts to be unfortunate. The best course of action for the academic community would be an open debate about what exactly happened with all these studies. Since this debate is now happening in the courtroom, it imposes a huge personal cost on people who raised valid questions. In particular, this is problematic because it has a chilling effect on future scrutiny of published research. This is why supporting the legal defence fund for Data Colada is an important public good, to which we should all contribute. If you agree with us, you may feel free to contribute to the Data Colada legal team through this link. Not only them, but also surely economics and science will benefit.

[We would like to note that, beyond Gino's legal complaint and the cited Data Colada posts, we have gathered much of the information from Andrew Ganato's post from August 4 entitled "Addressing the Data Analysis in Francesca Gino's Data Colada Lawsuit".]