



Thinking about evidence, and vice versa

[Menu](#)

[HOME](#) [ABOUT](#) [FEEDBACK POLICY](#) [TABLE OF CONTENTS](#) [SEMINAR](#)

[PAST SEMINARS](#)

[Menu](#)

[98] Evidence of Fraud in an Influential Field Experiment About Dishonesty

Posted on August 17, 2021 by Uri, Joe, & Leif

This post is co-authored with a team of researchers who have chosen to remain anonymous. They uncovered most of the evidence reported in this post. These researchers are not connected in any way to the papers described herein.

In 2012, Shu, Mazar, Gino, Ariely, and Bazerman published a three-study paper in PNAS ([.htm](#)) reporting that dishonesty can be reduced by asking people to sign a statement of honest intent *before* providing information (i.e., at the top of a document) rather than *after* providing information (i.e., at the bottom of a document). In 2020, Kristal, Whillans, and the five original authors published a follow-up in PNAS entitled, “Signing at the beginning versus at the end does not decrease dishonesty” ([.htm](#)).

They reported six studies that failed to replicate the two original lab studies, including one attempt at a direct replication and five attempts at conceptual replications.

Our focus here is on Study 3 in the 2012 paper, a field experiment ($N = 13,488$) conducted by an auto insurance company in the southeastern United States under the supervision of the fourth author. Customers were asked to report the current odometer reading of up to four cars covered by their policy. They were randomly assigned to sign a statement indicating, “I promise that the information I am providing is true” either at the top or bottom of the form. Customers assigned to the 'sign-at-the-top' condition reported driving 2,400 more miles (10.3%) than those assigned to the 'sign-at-the-bottom' condition.

The authors of the 2020 paper did not attempt to replicate that field experiment, but they did discover an anomaly in the data: a large difference in *baseline* odometer readings across conditions, even though those readings were collected long *before* – many months if not years before – participants were assigned to condition. The condition difference before random assignment (~15,000 miles) was much larger than the analyzed difference after random assignment (~2,400 miles):

This is Table 1 in Kristal et al. (2020), reporting their re-analysis of Shu et al. (2012)

	Sign-at-the-bottom, means (SD)	Sign-at-the-top, means (SD)	Two-sided t test, values
Baseline odometer reading (t_0)	75,034.50 (50,265.35)	59,692.71 (49,953.51)	$t_{(13,474)} = 17.78, P < 0.0001$
New odometer reading (t_1)	98,705.14 (51,934.76)	85,791.10 (51,701.31)	$t_{(13,475)} = 14.47, P < 0.0001$
Difference in odometer readings; i.e., miles driven ($t_1 - t_0$)*	23,670.64 (12,621.38)	26,098.40 (12,253.37)	$t_{(13,448)} = -11.331, P < 0.0001$

*This row was the outcome reported in the original paper.

In trying to understand this, the authors of the 2020 paper speculated that perhaps “the randomization failed (or may have even failed to occur as instructed) in that study” (p. 7104).

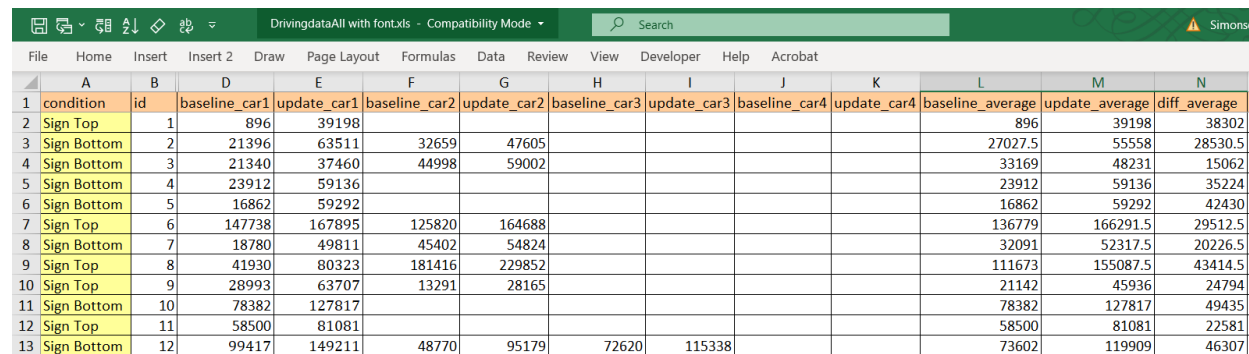
On its own, that is an interesting and important observation. But our story really starts from here, thanks to the authors of the 2020 paper, who posted the data of their replication attempts *and* the data from the original 2012 paper ([.htm](#)). A team of anonymous researchers downloaded it, and discovered that this field experiment suffers from a much bigger problem than a randomization failure: There is very strong evidence that the data were fabricated.

We’ll walk you through the evidence that we and these anonymous researchers uncovered, which comes in the form of four anomalies contained within the posted

data file. The original data, as well as all of our data and code, are available on ResearchBox ([.htm](#)).

The Data

Let's start by describing the data file. Below is a screenshot of the first 12 observations [1]:



	A	B	D	E	F	G	H	I	J	K	L	M	N
1	condition	id	baseline_car1	update_car1	baseline_car2	update_car2	baseline_car3	update_car3	baseline_car4	update_car4	baseline_average	update_average	diff_average
2	Sign Top	1	896	39198							896	39198	38302
3	Sign Bottom	2	21396	63511	32659	47605					27027.5	55558	28530.5
4	Sign Bottom	3	21340	37460	44998	59002					33169	48231	15062
5	Sign Bottom	4	23912	59136							23912	59136	35224
6	Sign Bottom	5	16862	59292							16862	59292	42430
7	Sign Top	6	147738	167895	125820	164688					136779	166291.5	29512.5
8	Sign Bottom	7	18780	49811	45402	54824					32091	52317.5	20226.5
9	Sign Top	8	41930	80323	181416	229852					111673	155087.5	43414.5
10	Sign Top	9	28993	63707	13291	28165					21142	45936	24794
11	Sign Bottom	10	78382	127817							78382	127817	49435
12	Sign Top	11	58500	81081							58500	81081	22581
13	Sign Bottom	12	99417	149211	48770	95179	72620	115338			73602	119909	46307

You can see variables representing the experimental condition, a masked policy number, and two sets of mileages for up to four cars. The “baseline_car[x]” columns contain the mileage that had been previously reported for the vehicle x (at Time 1), and the “update_car[x]” columns show the mileage reported on the form that was used in this experiment (at Time 2). The “average” columns report the average mileage of all cars in the row at Time 1 (“baseline_average”) and Time 2 (“update_average”). Finally, the last column (“diff_average”) is the dependent variable analyzed in the 2012 paper: It is the difference between the average mileage at Time 2 and the average mileage at Time 1. We will refer to this dependent variable as *miles driven*.

It is important to keep in mind that miles driven was not reported directly by customers. It was computed by subtracting their Time 1 mileage report, collected long before the experiment was conducted, from their Time 2 mileage report, collected during the experiment.

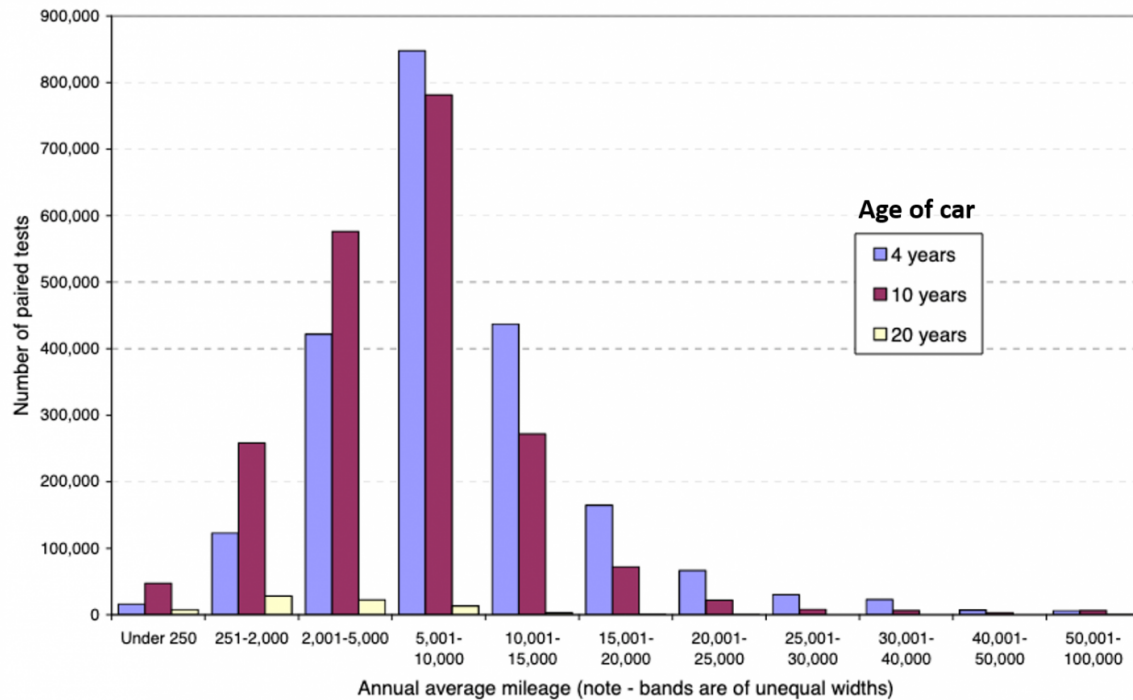
On to the anomalies.

Anomaly #1: Implausible Distribution of Miles Driven

Let's first think about what the distribution of miles driven *should* look like. If there were about a year separating the Time 1 and Time 2 mileages, we might expect something like the figure below, taken from the UK Department of Transportation ([.pdf](#)) based on similar data (two consecutive odometer readings) collected in 2010 [2]:

Figure from a UK Department of Transportation Report on Distribution of Yearly Miles Driven in 2010

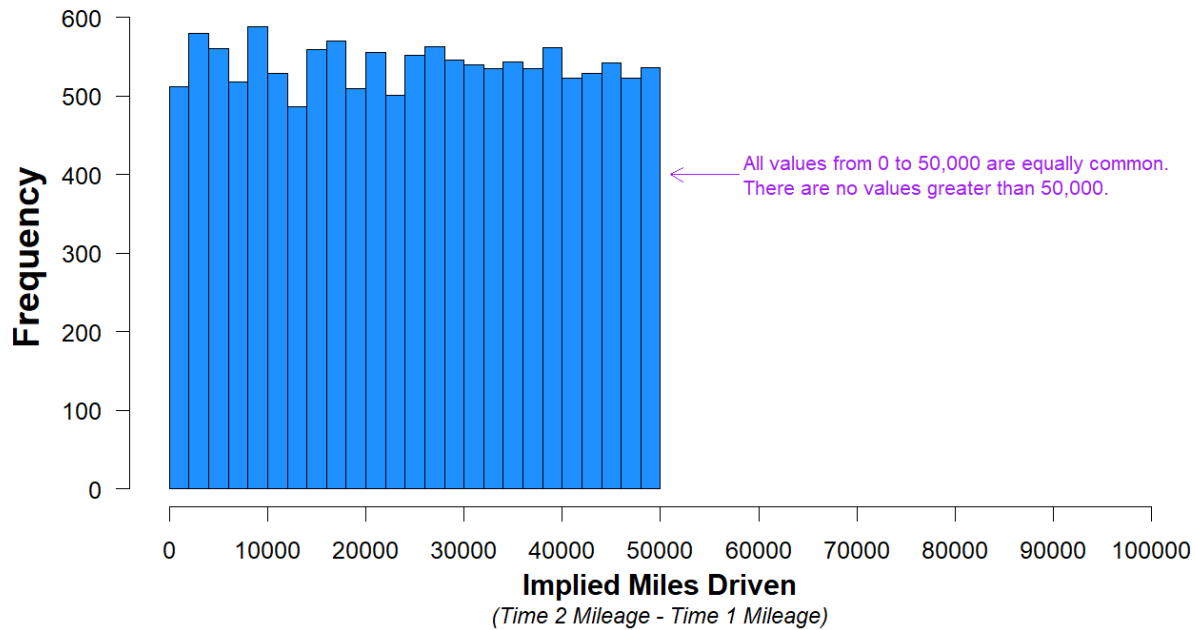
Source: <https://bit.ly/3iwSP2N>



As we might expect, we see that some people drive a whole lot, some people drive very little, and most people drive a moderate amount.

As noted by the authors of the 2012 paper, it is unknown how much time elapsed between the baseline period (Time 1) and their experiment (Time 2), and it was reportedly different for different customers [3]. For some customers the “miles driven” measure may reflect a 2-year period, while for others it may be considerably more or less than that [4]. It is therefore hard to know what the distribution of miles driven *should* look like in those data. It is not hard, however, to know what it should *not* look like. It should not look like this:

Figure 1. Histogram of Miles Driven - Car #1 (N=13,488)



This histogram shows miles driven for the first car in the dataset. There are two important features of this distribution.

First, it is visually and statistically ($p=.84$) indistinguishable from a uniform distribution ranging from 0 miles to 50,000 miles [5]. Think about what that means. Between Time 1 and Time 2, just as many people drove 40,000 miles as drove 20,000 as drove 10,000 as drove 1,000 as drove 500 miles, etc. [6]. This is not what real data look like, and we can't think of a plausible benign explanation for it.

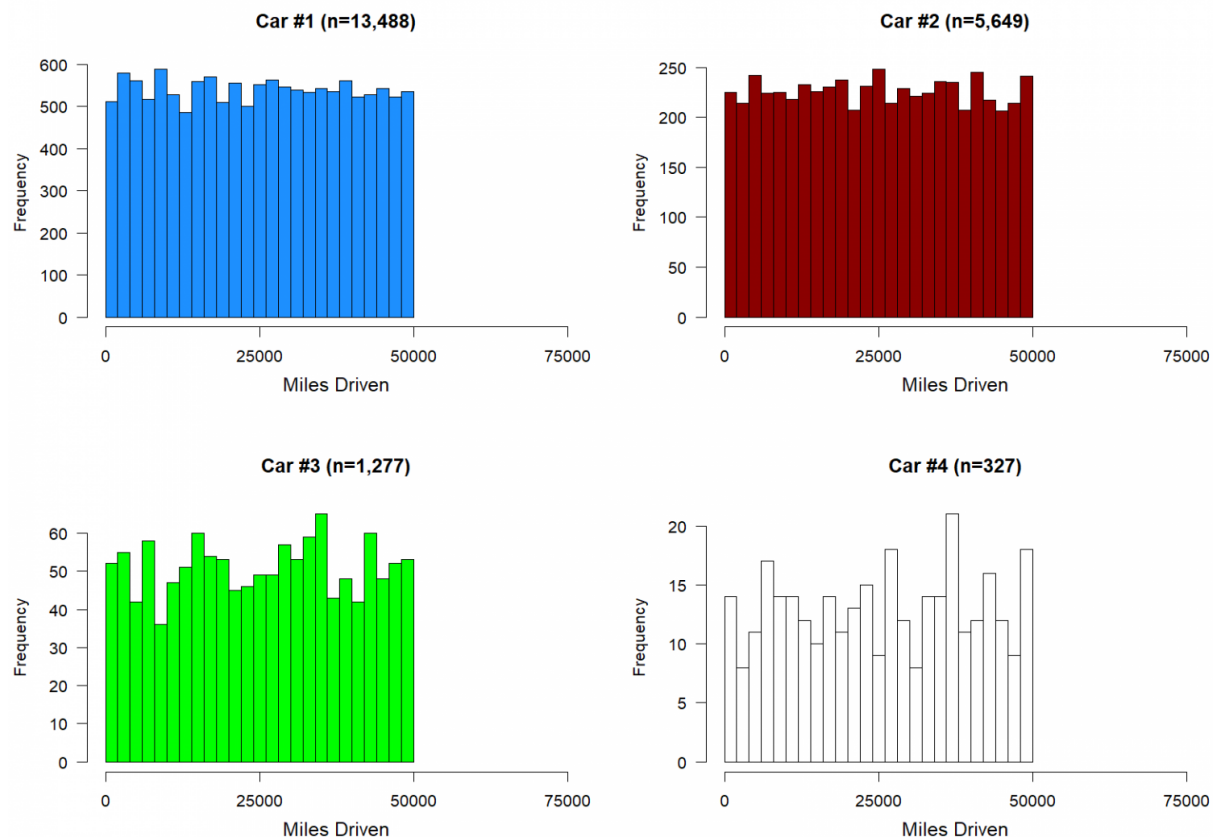
Second, you can also see that the miles driven data abruptly end at 50,000 miles. There are 1,313 customers who drove 40,000–45,000 miles, 1,339 customers who drove 45,000–50,000 miles, and *zero* customers who drove more than 50,000 miles. This is *not* because the data were winsorized at or near 50,000. The highest value in the dataset is 49,997, and it appears only once. The drop-off near 50,000 miles only makes sense if cars that were driven more than 50,000 miles between Time 2 and Time 1 were either never in this dataset or were excluded from it, either by the company or the authors. We think this is very unlikely [7].

A more likely explanation is that miles driven was generated, at least in part, by adding a uniformly distributed random number, capped at 50,000 miles, to the baseline mileage of each customer (and each car). This is easy to do in Excel (e.g., using `RANDBETWEEN(0,50000)`).

Why do we think this is what happened?

First, this uniform distribution of miles driven is not only observed for the first car, but for all four cars:

Figure 2. Miles Driven For Each Car



Once again, these distributions are visually and statistically (all p s > .78) consistent with a uniform distribution ranging from 0 to 50,000 miles.

Second, there is some weird stuff happening with rounding...

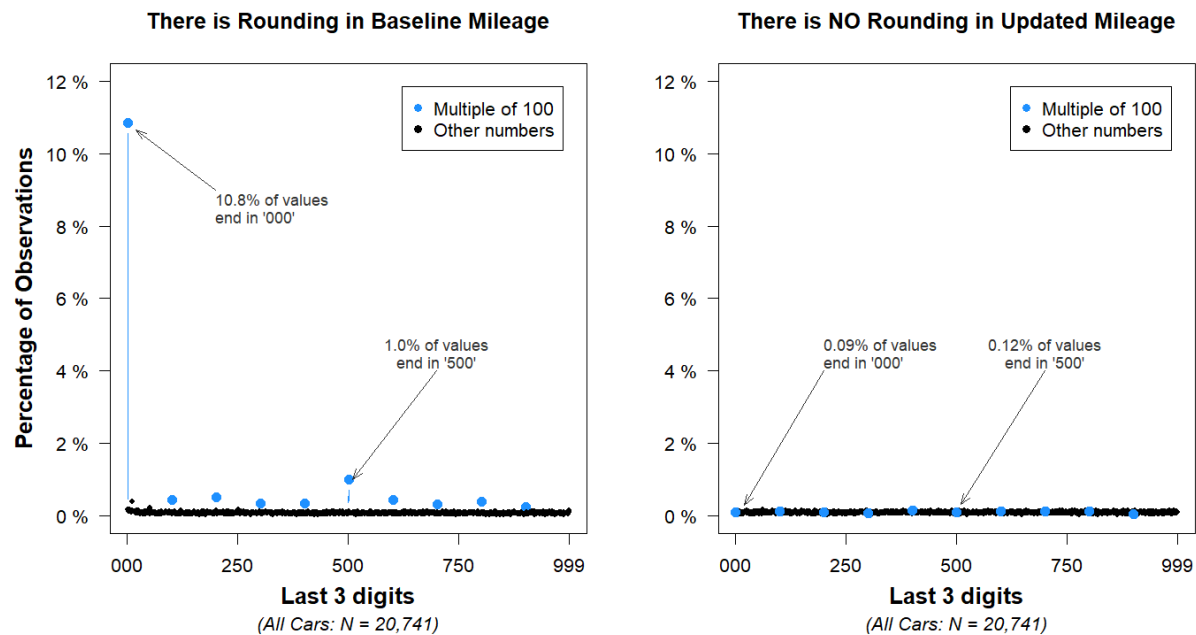
Anomaly #2: No Rounded Mileages At Time 2

The mileages reported in this experiment were just that: *reported*. They are what people wrote down on a piece of paper. And when real people report large numbers by hand, they tend to round them. Of course, in this case some customers may have looked at their odometer and reported exactly what it displayed. But undoubtedly many would have ballparked it and reported a round number. In fact, as we are about to show you, in the *baseline* (Time 1) data, there are lots of rounded values.

But random number generators don't round. And so if, as we suspect, the experimental (Time 2) data were generated with the aid of a random number generator (like `RANDBETWEEN(0,50000)`), the Time 2 mileage data would not be rounded.

Let's first look at the last three digits for every car in the dataset to see how likely people were to report mileages rounded to the nearest thousand (ending in 000) or hundred (ending in 00) [8]:

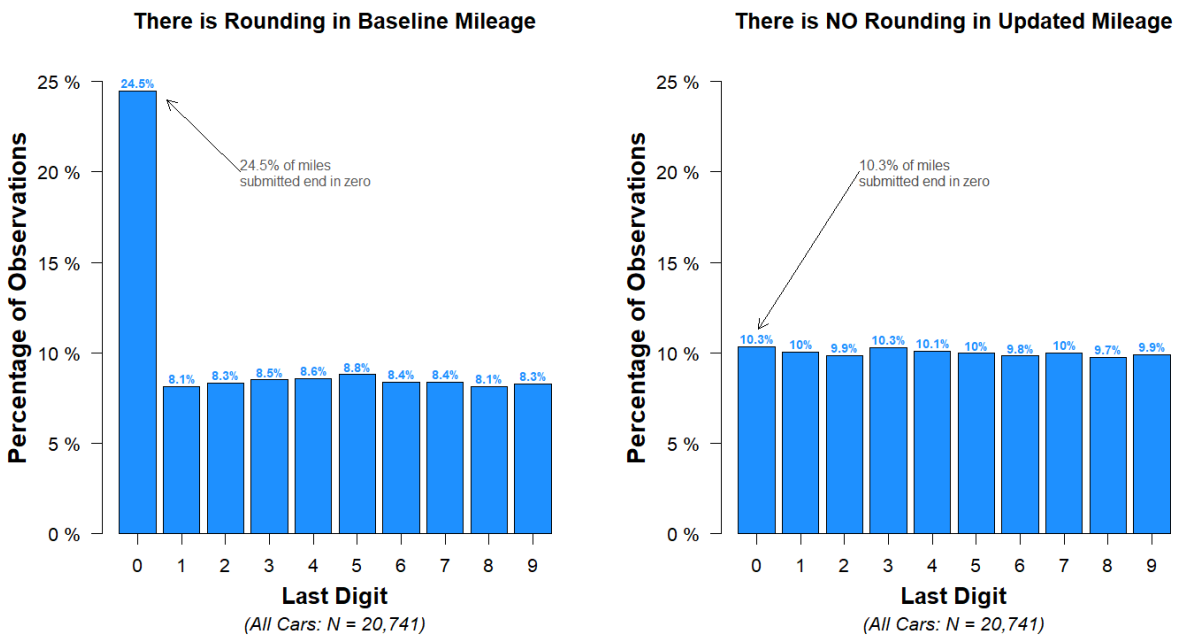
Figure 3. Last Three Digits at Baseline (Time 1) vs Updated (Time 2)



The figure shows that while multiples of 1,000 and 100 were disproportionately common in the Time 1 data, they weren't more common than other numbers in the Time 2 data. Let's consider what this implies. It implies that thousands of human beings who hand-reported their mileage data to the insurance company engaged in no rounding whatsoever. For example, it implies that a customer was *equally* likely to report an odometer reading of 17,498 miles as to report a reading of 17,500. This is not only at odds with common knowledge about how people report large numbers, but also with the Time 1 data on file at the insurance company.

You can also see this if we focus just on the last digit:

Figure 4. Last Digit at Baseline (Time 1) vs Updated (Time 2)



These data are consistent with the hypothesis that a random number generator was used to create the Time 2 data.

In the next section we will see that even the *Time 1* data were tampered with.

Interlude: Calibri and Cambria

Perhaps the most peculiar feature of the dataset is the fact that the baseline data for Car #1 in the posted Excel file appears in two different fonts. Specifically, half of the data in that column are printed in Calibri, and half are printed in Cambria. Here's a screenshot of the file again, now with a variable we added indicating which font appeared in that column. The different fonts are easier to spot if you focus on the font size, because Cambria appears larger than Calibri. For example, notice that Customers 4 and 5 both have a 5-digit number in "baseline_car1", but that the numbers are of different sizes:

DrivingdataAll with font.xlsx - Saved					
File Home Insert Insert 2 Draw Page Layout Formulas Data Review View Developer Help Acrobat					
	A	B	C	D	E
1	condition	id	font	baseline_car1	update_car1
2	Sign Top	1	Cambria	896	39198
3	Sign Bottom	2	Cambria	21396	63511
4	Sign Bottom	3	Cambria	21340	37460
5	Sign Bottom	4	Cambria	23912	59136
6	Sign Bottom	5	Calibri	16862	59292
7	Sign Top	6	Calibri	147738	167895
8	Sign Bottom	7	Calibri	18780	49811
9	Sign Top	8	Calibri	41930	80323
10	Sign Top	9	Cambria	28993	63707
11	Sign Bottom	10	Calibri	78382	127817
12	Sign Top	11	Calibri	58500	81081

The analyses we have performed on these two fonts provide evidence of a rather specific form of data tampering. We believe the dataset began with the observations in Calibri font. Those were then duplicated using Cambria font. In that process, a random number from 0 to 1,000 (e.g., `RANDBETWEEN(0,1000)`) was added to the baseline (Time 1) mileage of each car, perhaps to mask the duplication [9].

In the next two sections, we review the evidence for this particular form of data tampering [10].

Anomaly #3: Near-Duplicate Calibri and Cambria Observations


Let's start with our claim that the Calibri and Cambria observations are near-duplicates of each other. What is the evidence for that?

First, the baseline mileages for Car #1 appear in Calibri font for 6,744 customers in the dataset and Cambria font for 6,744 customers in the dataset. So exactly half are in one font, and half are in the other. For the other three cars, there is an odd number of observations, such that the split between Cambria and Calibri is off by exactly one (e.g., there are 2,825 Calibri rows and 2,824 Cambria rows for Car #2).


Second, each observation in Calibri tends to match an observation in Cambria.

To understand what we mean by "match" take a look at these two customers:


DrivingdataAll with font							
Home Insert Draw Page Layout Formulas Data Review View Tell me							
	A	B	C	D	E	F	G
1	condition	id	font	baseline_car1	baseline_car2	baseline_car3	baseline_car4
10	Sign Bottom	5938	Calibri	49675	17709	27357	64428
11	Sign Bottom	1137	Cambria	50350	18421	27714	64784




Cambria is 675
miles more
than Calibri



Cambria is 712
miles more
than Calibri



Cambria is 357
miles more
than Calibri



Cambria is 356
miles more
than Calibri

The top customer has a “baseline_car1” mileage written in Calibri, whereas the bottom’s is written in Cambria. For all four cars, these two customers have *extremely* similar baseline mileages. Indeed, in all four cases, the Cambria’s baseline mileage is (1) greater than the Calibri mileage, and (2) within 1,000 miles of the Calibri mileage. Before the experiment, these two customers were like driving twins.

Obviously, if this were the only pair of driving twins in a dataset of more than 13,000 observations, it would not be worth commenting on. But it is not the only pair. There are 22 four-car Calibri customers in the dataset [11]. *All* of them have a Cambria driving twin: a Cambria-fonted customer whose mileage for all four cars is greater than theirs by less than 1,000 miles. Here are some examples:

DrivingdataAll with font.xls - Compatibility Mode							
File Home Insert Insert 2 Draw Page Layout Formulas Data Review View Developer Help Acrobat							
	A	B	C	D	F	H	J
1	condition	id	font	baseline_car1	baseline_car2	baseline_car3	baseline_car4
2	Sign Top	1378	Calibri	128392	124477	87000	14255
3	Sign Top	12441	Cambria	128516	124659	87127	14862
4	Sign Top	6559	Calibri	11652	71000	13938	17911
5	Sign Bottom	10266	Cambria	12633	71384	13946	18711
6	Sign Bottom	8315	Calibri	14437	13640	17879	33864
7	Sign Bottom	7540	Cambria	14846	13821	18864	33985
8	Sign Top	12266	Calibri	47600	6500	15000	39000
9	Sign Top	6200	Cambria	47951	6901	15105	39364
10	Sign Bottom	5938	Calibri	49675	17709	27357	64428
11	Sign Bottom	1137	Cambria	50350	18421	27714	64784
12	Sign Top	2616	Calibri	13	130240	37910	80791
13	Sign Top	7347	Cambria	845	131045	38591	80980
14	Sign Top	10471	Calibri	57000	123663	16000	90000
15	Sign Top	11301	Cambria	57640	123666	16469	90026
16	Sign Top	9237	Calibri	8907	104849	35094	91640
17	Sign Top	7241	Cambria	9058	105406	35642	92607
18	Sign Bottom	4468	Calibri	18904	13024	103791	96954
19	Sign Bottom	4570	Cambria	19827	13425	103939	97538

Because there are so few policies with four cars, finding these twins requires minimal effort [12]. But there are twins throughout the data, and you can easily identify them for three-car, two-car, and unusual one-car customers, too.

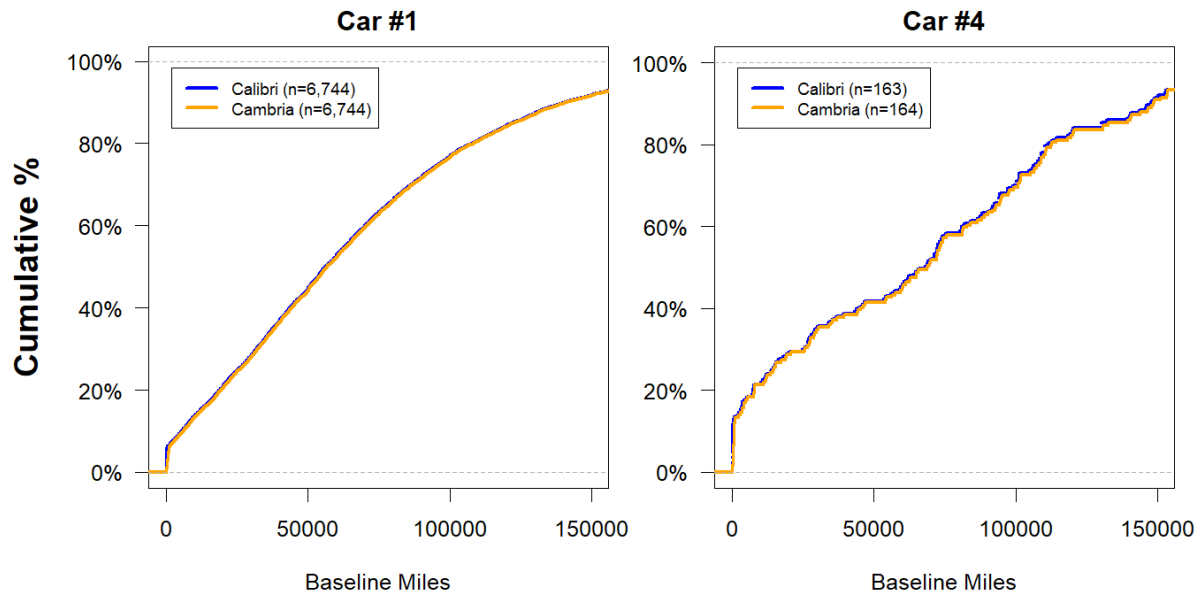
For example, it's easy to find the twins when you look at customers with extremely high baseline mileage for Car #1. Let's look at those with Car #1 mileages above 350,000.

There are 12 such policies for Calibri customers in the dataset...and 12 such policies for Cambria customers in the dataset. Once again, each Calibri observation has a Cambria twin whose baseline mileage exceeds it by less than 1,000 miles. This is again true for every car on the policy:

DrivingdataAll with font.xls - Compatibility Mode							
File Home Insert Insert 2 Draw Page Layout Formulas Data Review View Developer Help Acrobat							
	A	B	C	D	F	H	J
1	condition	id	font	baseline_car1	baseline_car2	baseline_car3	baseline_car4
2	Sign Top	12938	Cambria	983155			
3	Sign Top	13146	Calibri	982573			
4	Sign Bottom	12065	Cambria	735965	100512	163756	
5	Sign Bottom	5999	Calibri	735451	99735	163390	
6	Sign Bottom	12843	Cambria	603001	153284	130947	153254
7	Sign Bottom	5442	Calibri	602368	152327	130210	152600
8	Sign Bottom	767	Cambria	463284			
9	Sign Bottom	11557	Calibri	463090			
10	Sign Bottom	6120	Cambria	444290			
11	Sign Bottom	7357	Calibri	443920			
12	Sign Bottom	2324	Cambria	417041	48826	119477	
13	Sign Top	6297	Calibri	416537	48813	118579	
14	Sign Top	1895	Cambria	409663	31578	95013	
15	Sign Top	3821	Calibri	409515	31134	95000	
16	Sign Top	4819	Cambria	403733			
17	Sign Top	10804	Calibri	402847			
18	Sign Top	10181	Cambria	395272			
19	Sign Top	10650	Calibri	394482			
20	Sign Bottom	12845	Cambria	365387	112247	49086	
21	Sign Bottom	10362	Calibri	364774	112123	48472	
22	Sign Bottom	5117	Cambria	359700			
23	Sign Bottom	3779	Calibri	359641			
24	Sign Top	1510	Cambria	358544	112660	40845	
25	Sign Top	6129	Calibri	358236	111823	40000	

To see a fuller picture of just how similar these Calibri and Cambria customers are, take a look at Figure 5, which shows the cumulative distributions of baseline miles for Car #1 and Car #4. Within each panel, there are two lines, one for the Calibri distribution and one for the Cambria distribution. The lines are so on top of each other that it is easy to miss the fact that there are two of them:

Figure 5. CDFs of Miles Driven for Rows in Calibri vs Cambria Font



We ran 1 million simulations to determine how often this level of similarity could emerge just by chance. Under the most generous assumptions imaginable, it didn't happen once. For details, see this footnote: [\[13\]](#).

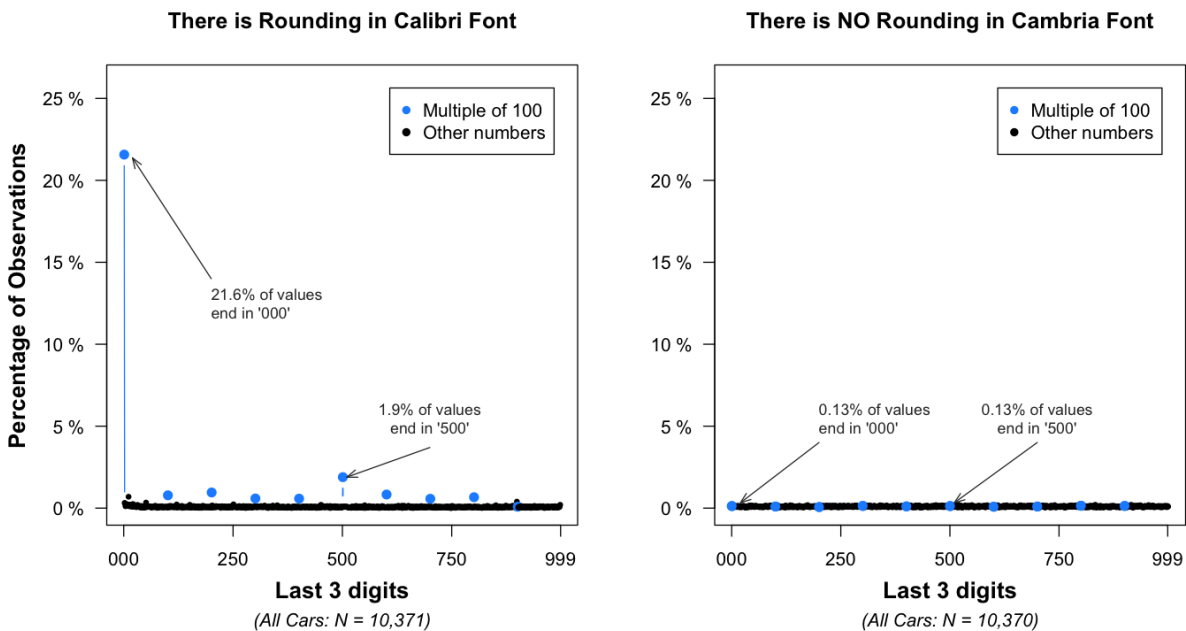
These data are not just excessively similar. They are impossibly similar.

Anomaly #4: No Rounding in Cambria Observations

As mentioned above, we believe that a random number between 0 and 1,000 was added to the Calibri baseline mileages to generate the Cambria baseline mileages. And as we have seen before, this process would predict that the Calibri mileages are rounded, but that the Cambria mileages are not.

This is indeed what we observe:

Figure 6. Last Three Digits at Baseline: Calibri vs. Cambria



We thank *The Economist* for spotting two typos in Figure 6, which we have now fixed. Original version: ([.htm](#)).

Conclusion

The evidence presented in this post indicates that the data underwent at least two forms of fabrication: (1) many Time 1 data points were duplicated and then slightly altered (using a random number generator) to create additional observations, and (2) all of the Time 2 data were created using a random number generator that capped miles driven, the key dependent variable, at 50,000 miles.

A single fraudulent dataset almost never provides enough evidence to answer all relevant questions about how that fraud was committed. And this dataset is no exception. First, it is impossible to tell from the data who fabricated it. But because the fourth author has made it clear to us that he was the only author in touch with the insurance company, there are three logical possibilities: the fourth author himself, someone in the fourth author's lab, or someone at the insurance company. This footnote contains some supporting evidence: [14]. Second, we do not yet know exactly how the data were tampered with in order to produce the condition differences. Were the condition labels generated or altered after the mileage data were created? And we also don't know the answer to other relevant questions, such as why the Calibri data were duplicated, or why the fabricator(s) generated condition differences at Time 1 [15]. Of course, we don't need to know the answer to these questions to know that the data were fabricated. We know, beyond any shadow of a doubt, that they were.

We have worked on enough fraud cases in the last decade to know that scientific fraud is more common than is convenient to believe, and that it does not happen only on the periphery of science. Addressing the problem of scientific fraud should not be left to a few anonymous (and fed up and frightened) whistleblowers and some (fed up and frightened) bloggers to root out. The consequences of fraud are experienced collectively, so eliminating it should be a collective endeavor. What can everyone do?

There will never be a perfect solution, but there is an obvious step to take: Data should be posted. The fabrication in this paper was discovered because the data were posted. If more data were posted, fraud would be easier to catch. And if fraud is easier to catch, some potential fraudsters may be more reluctant to do it. Other disciplines are already doing this. For example, many top economics journals *require* authors to post their raw data [16]. There is really no excuse. All of our journals should require data posting.

Until that day comes, all of us have a role to play. As authors (and co-authors), we should always make all of our data publicly available. And as editors and reviewers, we can ask for data during the review process, or turn down requests to review papers that do not make their data available. A field that ignores the problem of fraud, or pretends that it does not exist, risks losing its credibility. And deservedly so.



Author Feedback

We contacted all authors of the 2012 and 2020 paper, inviting them to provide feedback on earlier drafts of this post and to post responses to it. In response to feedback, we made a few minor edits to the post. In addition, four of the authors of the original paper asked us to post responses here: Nina Mazar ([.pdf](#)), Francesca Gino ([.pdf](#)), Dan Ariely ([.pdf](#)), and Max Bazerman ([.pdf](#)).

Subscribe to Blog via Email

Enter your email address to subscribe to this blog and receive notifications of new posts by email.

Footnotes.

1. To make things easier to understand, we have changed the variable names from what they were in the original posted file. [↗]
2. We added the "Age of car" legend title. [↗]
3. The authors wrote that “miles driven per car have been accumulated over varying unknown time periods” (p. 15200). [↗]
4. As the authors of the 2012 paper point out, the average reported mileage amount per car in this study (about 25,000 miles) is about twice what the average American typically drives in a single year (p. 15198). [↗]
5. The reported $p=.84$ comes from a Kolmogorov-Smirnov test comparing the miles driven data to a uniform distribution ranging from 0-50,000 [↗]
6. For example, for Car #1 there are 115 customers who drove between 0-500 miles, 136 between 5,000-5,500 miles, 134 between 10,000-10,500 miles, and 126 between 49,500-50,000 miles. [↗]
7. First, the authors of the 2012 paper did not report any such exclusions. Second, it seems very strange for a company to drop or exclude customers who reported driving more than 50,000 miles on any one car, but to retain those who reported driving much more than 50,000 miles across their many cars. This is even stranger since the time elapsed between Time 1 and Time 2 varied across customers, for it would mean, for example, that a customer who reported driving more than 50,000 over three years would be dropped while a customer who reported driving 49,998 miles over 1 year would not be dropped. It doesn't seem plausible that a company would do this. [↗]
8. These results are the same if we look at each of the four cars separately. [↗]
9. And as discussed above, we believe that for all observations, a separate random number of up to 50,000 was added to create the Time 2 data for each car. [↗]
10. It is also worth noting that all of the Time 2 mileages for Car #1 appear in Cambria font as well. Except for the two columns containing mileages for Car #1, everything else in the dataset is in Calibri font. [↗]
11. You might be wondering, “Wait, how come you are now saying there are only 44 four-car customers in the dataset when just before this you showed that there are 327 observations for Car #4?” Thanks for being so attentive. It turns out that many customers have reported mileage for Car #4 without having four cars. That is, they have values for Car #4, but not for Car #2 and/or Car #3. [↗]
12. Simply sorting policies with four cars by the number of miles reported in “baseline_car4” leads to an almost perfectly interleaved set of twins, as 20 of the 22 pairs appear in adjacent rows. The remaining two pairs of twins are separated by just one row. [↗]