![CloudResearch — powered by turkprime]

No, Americans are not gargling bleach: How bad data inflated estimates in the latest CDC report, and how to prevent this from happening in the future

Leib Litman, PhD

Overview

- Over 140 media organizations globally have reported on the recent CDC findings that 39% of Americans have participated in "high-risk" cleaning behaviors including cleaning practices such as washing food with bleach, inhaling cleaning products, and gargling disinfectants in response to the COVID-19 outbreak.
- To examine whether these results were an artifact of random responses made by inattentive participants, CloudResearch conducted a replication experiment in the form of an A/B test. A portion of respondents were routed through Sentry - a tool that identifies and routes inattentive and disengaged respondents out of surveys.
- Strikingly, the results showed that very high-risk cleaning practices were almost exclusively reported only by respondents who were identified as inattentive and disengaged, indicating that reports of very high-risk cleaning practices were largely an artifact.
- Lack of screening causes high-risk behaviors to be severely overestimated and normal cleaning practices to be underestimated.
- Best practices for collecting and interpreting survey data are discussed.

On 5 June 2020, the Centers for Disease Control and Prevention released a report titled, "Knowledge and Practices Regarding Safe Household Cleaning and Disinfection for COVID-19 Prevention — United States, May 2020" [1]. In the report, CDC researchers described an online survey in which Americans were asked about their knowledge of safe cleaning practices. Within the survey, people were also asked if they had engaged in any dangerous cleaning practices such as washing food with bleach, inhaling cleaning products, and gargling disinfectants.

According to the CDC, 39% of Americans engaged in at least one dangerous behavior during the previous month. Such behaviors included 19% of people who applied bleach to food items; 18% who used household cleaning products on their skin; 10% who misted their body with a disinfectant; 6% who inhaled vapor from a cleaning product or disinfectant; and, several people who engaged in 'very high-risk' behaviors including 4% who drank or gargled diluted bleach, 4% who drank or gargled soapy water, and 4% who drank or gargled another disinfectant. Clearly, these numbers are alarming.

So alarming, in fact, that they were quickly noticed by the media. Within five days, the CDC research was covered in 127 news reports, including by prominent organizations like the New York Times, CNN, NBC and FOX. The findings also quickly spread on social media. It is not an exaggeration to say that tens of millions of people worldwide learned about the CDC report within a matter of days. The problem is that the numbers were likely wrong.
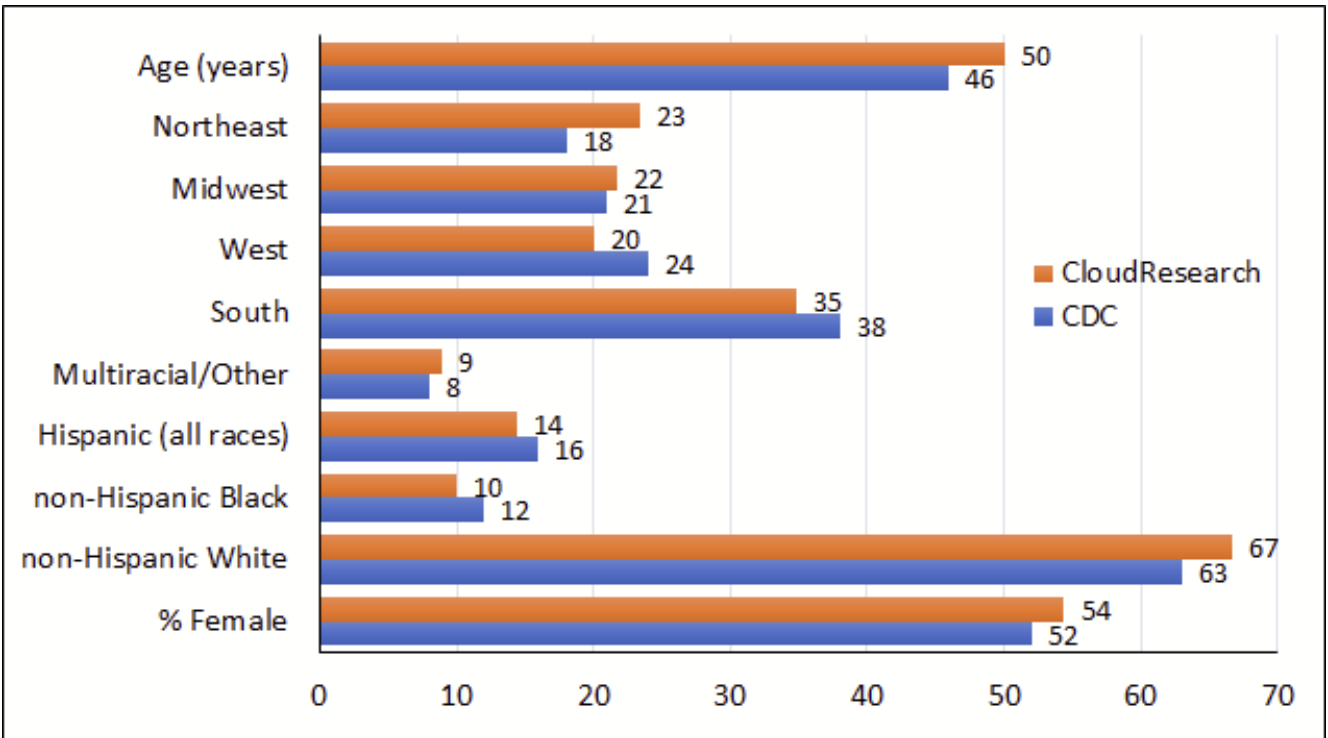
The report issued by the CDC used market research panels, which provide researchers with access to tens of millions of people around the world including millions of Americans. Online panels are among the most commonly used methods of obtaining survey responses in the social sciences, marketing research, and health-related fields. However, when researchers use such panels it is important to carefully curate the sample to eliminate inattentive respondents.

When people take surveys, multiple extraneous factors may compete for their attention including children and other family members, TVs, music, conversations, and other household distractions. These competing factors are accentuated in the time of COVID when multiple family members are likely to be home at the same time. In addition to inattentiveness, participants sometimes do not give sufficient effort to complete a survey properly or may lack the language fluency to fully comprehend all questions. These problems are not unique to online panels. Indeed, even face-to-face US Census interviews can suffer from inaccurate and fabricated responses [2].

In a recent study [3], we showed that sample curation is essential to ensure the attentiveness of respondents. In that study, we developed and validated methods for improving data quality by identifying and eliminating inattentive respondents from the sample. As a follow-up to that report, we developed a comprehensive and automated approach of sample curation that we call Sentry™. Sentry prevents inattentive and otherwise low-quality respondents from entering a survey using a combination of behavioral assessment and fraud detection technology. By routing such participants out of a survey, Sentry increases the signal to noise ratio of survey data and ensures that questions are being answered accurately.

After reading the CDC report, the team at CloudResearch launched a study to see if we could replicate the findings of the CDC and whether the findings would change when proper screening methods were applied. The goal of our inquiry was to determine what role inattentive respondents may have played in the CDC results.

![CloudResearch powered by turkprime]

FIGURE 1. A comparison of demographics reported in the CDC study and a replication study conducted by CloudResearch.

| Category	CloudResearch	CDC
Age (years)	50	46
Northeast	23	18
Midwest	22	21
West	20	24
South	35	38
Multiracial/Other	9	8
Hispanic (all races)	14	16
non-Hispanic Black	10	12
non-Hispanic White	67	63
% Female	54	52

The study conducted by CloudResearch used the same measures and sampling methodology reported by the CDC. As Figure 1 shows, the demographic and regional composition of samples was virtually the same across studies. In the CloudResearch surveys, we gathered data from over 2,500 respondents. We ran different versions of the study to be sure that the data we collected had the most accurate points for comparison, including versions where we asked people about behaviors over the past month (to replicate the exact wording of the CDC study), and about behaviors since April (since April was the month covered by the CDC study). In both cases the results were virtually identical.

We first examined the differences in self-reported cleaning practices between people who were flagged as being problematic (failed Sentry) and those who were identified as attentive (passed Sentry). As shown in Figure 2, people who failed Sentry reported engaging in high-risk cleaning behaviors much more frequently than people who passed Sentry. In fact, the differences were striking. People who reported engaging in the highest-risk behaviors (i.e. drinking/gargling household cleaner, soapy water, or diluted bleach) were almost exclusively those who failed Sentry. Specifically, 31.45% of people flagged by Sentry reported engaging in at least one of the

very high-risk cleaning behaviors (e.g., gargling bleach) while those behaviors were virtually non-existent (1.2%) among respondents cleared by Sentry.

The results were similarly striking for other high-risk behaviors such as washing food products with bleach, using cleaner or disinfectant on one's bare skin, misting the body, or inhaling vapors. For each of these behaviors, the likelihood of a 'yes' response was 17-30% higher among respondents identified as being problematic (see Figure 2).

FIGURE 2. Cleaning and disinfection practices in the previous month with the intent of preventing COVID-19 infection. Reports of engaging in very high-risk cleaning practices come almost exclusively from respondents identified as inattentive or disengaged, indicating that the effect is largely an artifact.
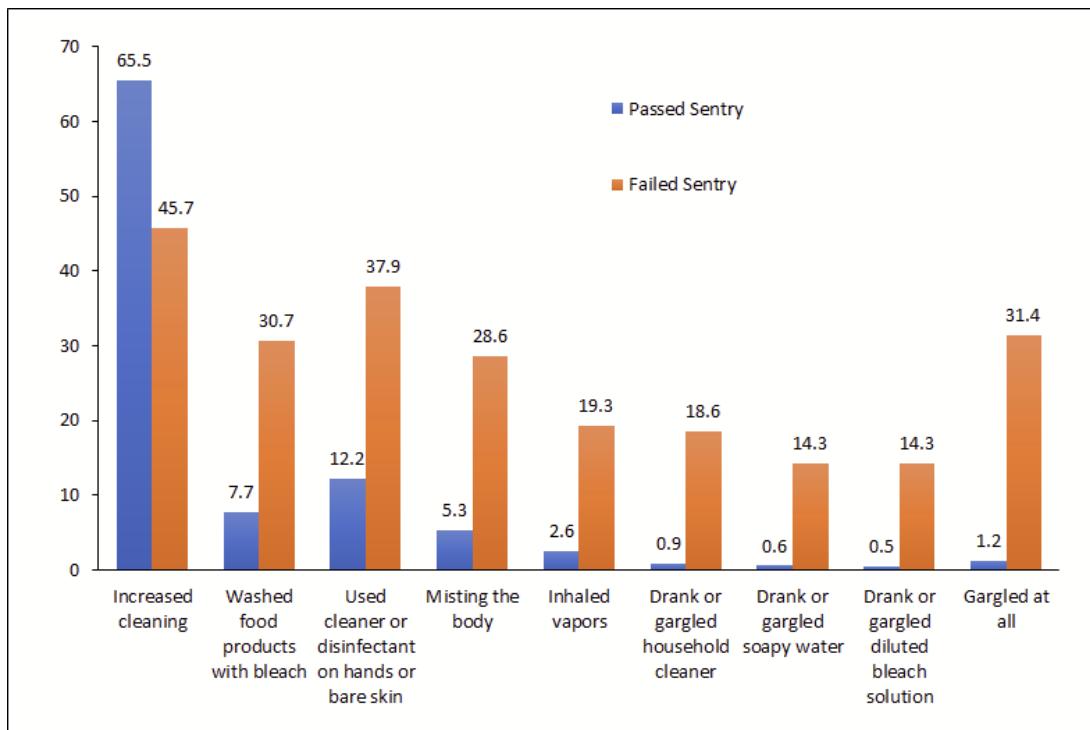
Note:

[1] Although in most cases Sentry is used to prevent low-quality respondents from entering a study, for the survey in Figure 2 we allowed all respondents to enter the survey and compared the data of people who would have been routed out of the survey by Sentry with those who passed Sentry.

[2] The study consisted of 622 people. Failed Sentry group, N = 120; Passed Sentry group, N = 410)

[3] Two versions of the study were conducted. V1 asked about behaviors over the past month. This was an exact replication of the CDC study. V2 asked about behaviors since April. April was the month covered by the CDC study. In both cases the results were virtually identical.

[4] The 'Gargled at all' bar shows the percentage of respondents who reported engaging in at least one of the three practices immediately to the left - drinking/gargling household cleaner, soapy water, or bleach solution.

Equally as important, we saw the opposite pattern for increases in regular cleaning behavior. Inattentive respondents under-reported increases in regular cleaning behavior by 20%. Thus, without screening, high-risk behaviors are severely overestimated while normal cleaning practices are underestimated.

Clearly, inattentive respondents can skew the results of surveys. But why are results affected so dramatically? The answer is simple probability. When respondents do not read questions they tend to randomly select from among the response options. Random responding decreases the signal-to-noise ratio. This noise drives all responses toward the middle of the distribution, artificially making infrequent practices seem more common and frequent practices seem less common. Importantly, low-frequency practices such as gargling household cleaners are especially vulnerable to such bias in this study because a change of even a few percentage points makes a big difference in interpretation. For this reason, very high-risk, low-frequency cleaning practices are impacted the most when inattentive and unengaged respondents are allowed to answer questions.

To demonstrate what happens when inattentive respondents are allowed to complete the CDC survey, we conducted an experiment in the form of an A/B test in which a portion of respondents were routed through Sentry and others were not. Our hypothesis was that in a sample collected without Sentry the results would look similar to those reported by the CDC. When routed through Sentry, however, we hypothesized that the incidence rate of risky behaviors would be significantly lower.

As shown in Figure 3, this is exactly what we found. Focusing only on the high-risk cleaning practices, the figure shows that people who completed the survey without Sentry provided results similar to the CDC report. When inattentive respondents were taken out of the survey by Sentry, however, the numbers were drastically reduced. Specifically, for the behaviors of drinking/gargling household cleaner, soapy water, or diluted bleach, a side-by-side comparison of the CDC data and our data with Sentry shows that the CDC numbers overestimate the percentage of people engaging in these behaviors by a factor between 4 and 8. About 4-5% of people reported doing each of these things in both the CDC and CloudResearch no-Sentry samples. Strikingly, the A/B test showed that without Sentry the likelihood of finding people who engage in at least one of these behaviors is 8.3%. However, when inattentive participants are taken out of the survey that number is reduced to just 1.2%.

FIGURE 3. Results of an A/B experimental manipulation on self-reported rates of very high-risk cleaning practices.
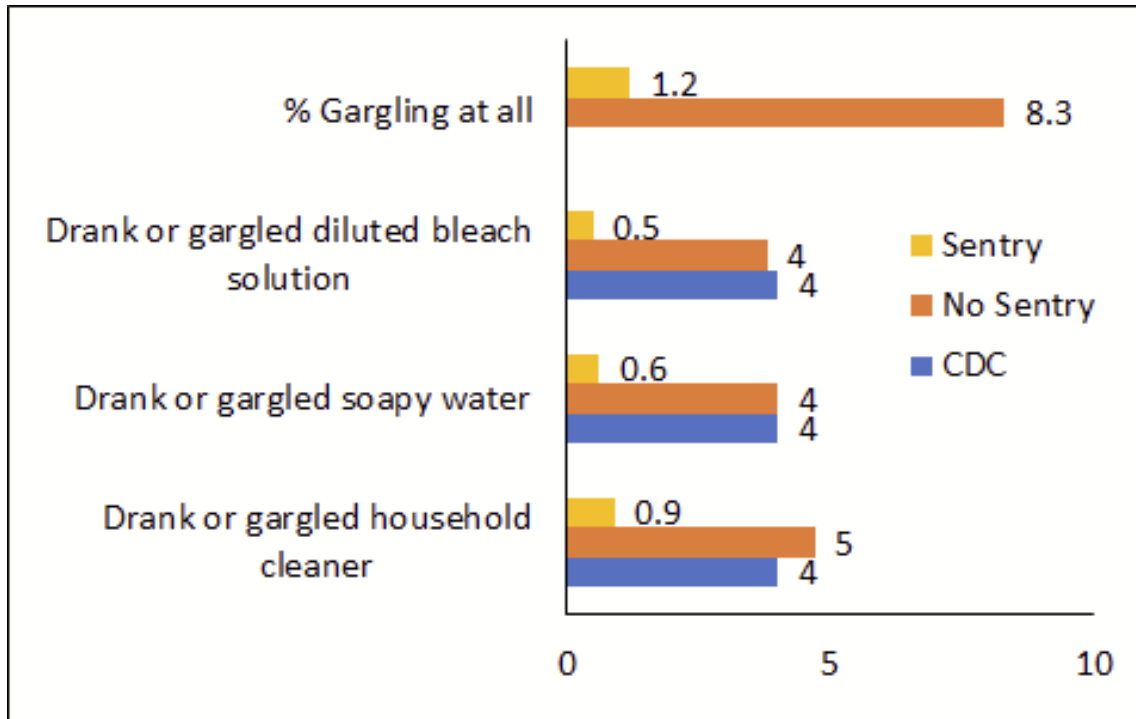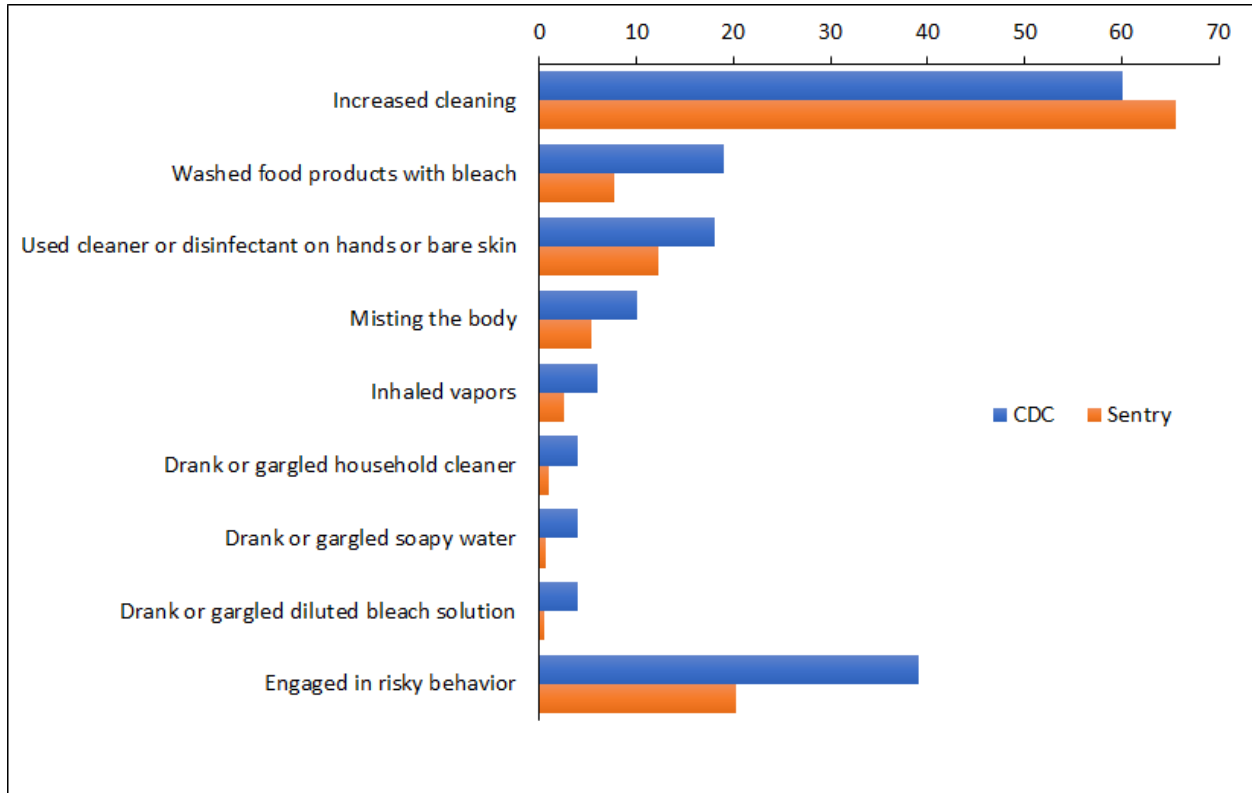

Note:
[1] The 'Gargling at all' bar shows the percentage of respondents who reported engaging in at least one of the three very high-risk practices - drinking/gargling household cleaner, soapy water, or bleach solution.
[2] The CDC data is not presented for the 'Gargling at all' bar because these data are not reported in the CDC study.
[3] Data shown in this graph for the Sentry condition is the same as in the Sentry condition in Figure 4.

For the other risky cleaning practices the discrepancy between the data reported by the CDC and the participants who passed Sentry was also dramatic (see Figure 4). Overall, we found that the percentage of people who reported engaging in any of the unsafe cleaning practices was 20%, as opposed to 39% in the CDC data. Perhaps most importantly, the more risky the behavior the higher the impact bad respondents had on the data. Cleaning practices that present the greatest health danger such as self-reports of gargling with bleach or drinking household cleaners were the least likely to be reported by attentive respondents.

FIGURE 4. A direct comparison of the incidence rate of high-risk cleaning practices as reported by the CDC and as observed by CloudResearch after preventing inattentive and disengaged respondents from taking the survey.

It should be noted that low data quality in survey research does not arise from the use of any specific online platform or panel provider. Instead, low data quality is directly proportional to the number of inattentive and disengaged respondents who take a survey. The more inattentive respondents there are in a survey, the bigger the bias is likely to be. People can be inattentive when taking surveys even in face-to-face studies and in studies that take place in labs [4]. For that reason, measures that protect surveys from inattentive respondents are a universal requirement independent of the mode of recruitment.

It should also be noted that different surveys require different levels of protection against poor data quality. Under most circumstances, inattentive and disengaged respondents add "noise" to a dataset, making it harder to find significant results. When researchers use survey data to make inferences about unusual and low-frequency behaviors such as drinking bleach, however, screening becomes even more imperative. Assessments of data quality in behavioral research should be approached from the 'fit-for-purpose' perspective: different surveys require different standards of stringency depending on each survey's measurement goals. In the present study, we used the most stringent filters because even a small bias can lead to highly misleading

results when measuring the potential occurrence of low-frequency behaviors. When the goal of a survey is not to measure low-frequency behavior, less stringent filters are almost always sufficient to produce accurate and reliable results. Overall, the data cleaning approach has to be fit to the purpose of the study, as our data clearly demonstrate.

This brings us back to the claim made in the title of this article. Are people in fact drinking bleach solution? Our data shows that when inattentive and disengaged respondents are taken out of a survey, 0.4% of people report engaging in this practice. These data may indicate that a very small percentage of the U.S. population is indeed engaging in such high-risk practices. However, we believe making inferences from such small effects does not constitute sound scientific practice. Even large nationally representative probability samples have margins of error substantially higher than 0.4%. Given that the studies conducted by both the CDC and our team gathered data via a non-probability opt-in panel, the actual rate of these behaviors cannot be determined from the current data. More importantly, no data quality tool is 100% accurate. A few problematic respondents may have eluded detection. Therefore, it is possible that even these small numbers are also an artifact of inattentiveness.

A maxim in science is that extraordinary claims require extraordinary evidence. When data suggest that people are routinely engaging in surprising and extremely dangerous behaviors—especially behaviors with important public health implications—it is critical to closely scrutinize such findings. Fortunately, online platforms make it possible to do that in a variety of ways. For example, it is possible to follow up with individual respondents to gather more detailed information. Researchers can ask respondents to describe their cleaning practices in an open-ended format, such as to provide specific examples of their behavior, to provide more context, and to describe the rationale for their practices. Researchers could even set up video interviews with select respondents in order to verify that the respondents are indeed real, that they fully understand what is being asked of them, and that their behaviors are being reported accurately. Even a handful of interviews can provide important evidence that such practices really exist. In the absence of additional evidence, however, we are left to wonder whether the few people who selected a 'yes' response are an indication of a real public health concern or an accidental click of a button by someone who was not paying much attention.

In our recent book discussing data quality in online panels [5], we described how online studies can be grouped into different categories depending on their measurement goals. The current study is what we refer to as a "population-oriented frequency study," meaning that the study's goal is to make inferences about the frequency of behavior within the population. Population-oriented frequency studies are typically conducted using probability surveys where respondents are randomly selected from the population. These studies can be conducted with non-probability opt-in panels as well. However, when researchers use non-probability samples to make population-oriented frequency claims, the goal of the study should not be to precisely measure the frequency of a behavior within the population. Instead the goal of such studies should be to

provide initial evidence that a behavior of interest occurs in a population with a substantially higher frequency than may be otherwise expected.

From this perspective, the CDC report correctly concludes that people in the US are engaging in some risky cleaning practices. While we cannot know exactly what proportion of people are engaging in these practices from the current data, the results allow us to conclude that the number is not trivial. For example, it appears some people are washing their fruits and vegetables with diluted bleach solution, using cleaner and disinfectant on their hands and bare skin, and misting their body with cleaning or alcohol spray. Even though our data suggests that such practices are not as widespread as the CDC data indicate, there is enough evidence to conclude that the percentage of people engaging in such behaviors is not trivial and is greater than zero. However, for at least three and perhaps four of the most extreme behaviors examined by the CDC—drinking or gargling bleach, soapy water, or household cleaner—our study demonstrates that there is not sufficient evidence to conclude that the frequency of these behaviors is greater than zero in the population. This does not mean, of course, that we have evidence that such behavior does not occur. Instead, it means there is not sufficient evidence to conclude that Americans actually are engaging in these behaviors from the current data.

Data collection online has become standard practice even for major institutions like the CDC, but unless care is taken to ensure respondents are honest and attentive the results can be very misleading. Given our findings that the overwhelming majority of the very high-risk cleaning practices were reported by problematic respondents, a rigorous follow-up investigation is appropriate before concluding whether and at what frequency people may be engaging in such practices.

References
1. Gharpure, R. (2020). Knowledge and Practices Regarding Safe Household Cleaning and Disinfection for COVID-19 Prevention—United States, May 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*.
2. Breiman, L. (1994). The 1991 Census Adjustment: Undercount or Bad Data?. *Statistical Science*, *9*(4), 458-475.
3. Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods*, *51*(5), 2022-2038.
4. Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, *48*(1), 400-407.
5. Litman, L., & Robinson, J. (2020). *Conducting Online Research on Amazon Mechanical Turk and Beyond* (Vol. 1). SAGE Publications, Incorporated.