OPINION, **Q&A**

How to safeguard online data collection against fraud

BY **GRACE HUCKINS**

30 MARCH 2021

The COVID-19 pandemic has forced large numbers of researchers to move their studies from the **laboratory to the internet**. In some ways, this has been a good thing — decades ago, a global pandemic would have shut down research entirely. Now, scientists can use social media, online registries and crowdsourcing tools such as **Amazon Mechanical Turk** to recruit participants to take surveys or perform tasks, without investing much time or money.

But as helpful as these online tools are, they are not without risks, says **Clare Harrop**, assistant professor of allied health sciences at the University of North Carolina at Chapel Hill. When she and her team tried to recruit participants for a study in 2019 by sharing a link to a survey on social media, they received a torrent of fraudulent responses — presumably from someone who wanted the $5 gift cards her team offered respondents as compensation.

After the pandemic began and so much research moved online, Harrop and her colleagues knew that they needed to spread the word about what had happened to them, so they wrote a **letter to the editor**, published in *Autism Research* in January. Harrop spoke with *Spectrum* about her experience and shared some tips on how other researchers can protect themselves from fraud.

***Spectrum*: Tell me about the project you were working on when you ran into this issue.**

Clare Harrop: This data collection was part of a large outcome measurement grant led by **Brian Boyd** to create a measure of behavioral inflexibility for autistic children and then extend that to children with neurogenetic conditions — Down syndrome, Prader-Willi syndrome and fragile X syndrome. Most of my research focuses on sex differences, and the majority of my work has been done in person, but for this project we had used a magnitude of methods: focus groups, online surveys and in-person assessments.

S: What did the process of online data collection look like?

CH: We had an **online survey** led by **Luc Lecavalier** at Ohio State University that we used to collect data from nearly 1,000 parents of autistic children to validate the behavior inflexibility scale we created. For that survey, we worked with the **Interactive Autism Network**'s databases of parents who had consented to be contacted for research. And then we have since used other registries to extend the survey to parents of children with fragile X or Down syndrome.

We also meant to include Prader-Willi syndrome, but that survey dataset was victim to fraud. So that data never got collected.

S: Why was the situation so different with Prader-Willi?

CH: What happened is we worked with a national society, and they don't have a mailing list like the other groups do. So we launched the survey via social media, and it was shared. And that's where the issue came up. Initially, we had a very slow response, which is what you would expect. Prader-Willi is a rare genetic syndrome — you wouldn't expect 300 parents to complete it in two days. That population isn't readily accessible.

S: How did you find out that some of your data were fraudulent?

CH: Around day three, I remember I was in a meeting, and I had my laptop open and was answering all these emails from our data management team. And they said, "The survey is at capacity; it's all filled out." And I was like, "No, no, that's not normal," — our capacity was about 150 or 200 — "that's not right; there is no way that we've gone from about 20 participants to all of this. There's no way all these people filled it out." And our data management team pushed back and said, "Oh, someone really popular must have shared it, like a parent that knows a lot of people." And I was like, "No, no, it isn't." So we shut it down.

And then we had to go through the painstaking process of combing through the data, which took days. It was very clear that it was fraudulent data. The way we figured it out is that these surveys were started within minutes of each other, so it's almost like someone had them open on multiple servers, and the surveys were done very quickly. We tell parents, "Oh, this survey might take between 30 minutes and an hour." So when all your surveys are filled out within two minutes, you know that there's probably a problem. We collected names in the survey to send gift cards, and they were famous names. There were a lot of baseball and basketball players.

S: How did this affect the rest of your research?

CH: We had to go to all our other datasets to check that the same thing hadn't happened. And fortunately, it hadn't. We looked at response rates; we looked for duplicate responses. And we found only a handful — I think we had 10 or so. Some of them may have been intentional, but for others it was clearly obvious that both of the parents have filled out the form because they both got the email.

It was awful; it was really stressful. It took hours to do it.

S: What are you doing to protect yourself from fraudulent data going forward?

CH: We have a study looking at extending the behavioral inflexibility scale to attention deficit hyperactivity disorder that started just after this happened. We sent unique links to each individual we contacted, which is a lot more work on our part.

One of the things that I've noticed helps, because I've completed a few online surveys during COVID myself, is to throw in test questions — so suddenly reversing the order of scoring, or just a random question, like "What is three plus four?" And if it's a bot, or it's someone going through really quickly, you can catch people that way.

Another option is to use simple tests called CAPTCHAS. And definitely collect paradata, such as how long it takes a participant to complete their answers. That's really important, as is knowing the typical response rate. Other researchers sometimes use lottery compensation as opposed to compensation for all.

S: Why did you decide to share this experience in a letter to the editor?

CH: We wrote this letter to the editor over a year ago, but we were pushing forward other papers on the project. And then once COVID hit, I just turned around to the co-authors, and I was like, "Look, we're collecting all this data remotely now. And I feel like we need to get that letter out there, because this is going to be more of an issue with regard to COVID." So then we turned it around very quickly and submitted it in November.

S: Outside of the COVID-19 context, are there benefits to collecting data online, despite the risk of fraud?

CH: We know that autistic adults have a preference for screen-based media. So social media is great for recruiting adults in a way that's comfortable for them, rather than having them come into the lab and doing intensive interviews. It also allows us to increase our catchment areas and get a representative sample. And when we're thinking about rare genetic conditions, those participants are otherwise especially hard to get. I work with autistic females, and they're not that easy to get into the research clinic, either. So social media does give us a way to get bigger samples.

It's hard because we know that people use social media a lot, particularly for work with autistic adults. I don't think we should stop using it. But I think there are ways to make sure it's not fraudulent.