This week, our understanding of the impact journalists can make on public sentiment changed forever with the publication of an article in Science by Harvard Professor Gary King demonstrating that collaborations of even small outlets can change public sentiment by x% [embargoed].

That we can now measure the impact of jouranlism is the big story. But just as important as what the research showed was how the research was done. This experiment marks the first time ever that quantitative scientists were able to create a randomized experiment with journalists. My organization, the Media Consortium, helped develop and manage those experiments.

According to King, “Much of the work leading up to this study involved finding a way to bridge the cultural divide between journalism and science. Through years of conversations, much trial and error, and a partnership with Media Consortium Executive Director Jo Ellen Green Kaiser, we learned to understand journalistic standards and practices, and the journalists learned to understand our scientific requirements. What ultimately made it all work was a novel research design we developed that satisfied both camps.”

I remember sitting down with Gary on October 12, 2011, at Pizzaiolo restaurant in Oakland to figure out how we could make such an experiment happen.

We had been brought together by John Schwartz, leader of the Voqal Fund and a numbers guy who cares deeply about independent media. As the founder of Free Speech TV—a Media Consortium member—John was frustrated that he couldn’t find a quantitative method to measure its impact. He had researched the then-nascent metrics field and decided that analyzing changes in sentiment on social media would be the best proxy for how to measure changes in public opinion as a result of a news story. He reached out to Gary, the preeminent researcher in sentiment analysis, and in a series of phone calls, brought Gary and me together to see if we could devise an experiment using Media Consortium members to produce real numbers on impact.

When John first called me, in the summer of 2011, I had just started at the Media Consortium. I didn’t know much about metrics, but I knew a lot about independent media, having worked as managing editor at Tikkun and served in various consultant roles at the Independent Press Association. I was eager to support the project if it would help my members.

Frankly, however, my biggest concern coming into the meeting was whether the experiment would backfire and hurt our members. Media Consortium member outlets range in size from Democracy Now! and Mother Jones to Cascadia Times and People Power Media. Within the United States media ecosystem, even our largest outlets might be considered mid-sized, and our smallest outlets are very small indeed. I was worried that the smallest outlets—which were the most likely to participate—would not move the needle when measured against all posts on Twitter, the massive scale that Gary proposed.

So I came to the meeting with a proposal. Instead of measuring the impact of one outlet, why not measure the impact of a collaboration of at least three outlets?

I had just read a ground-breaking piece in the Standford Social Innovation Review by John Kania and Mark Kramer putting forward a [theory of collective impact](http://ssir.org/articles/entry/collective_impact). In this [approach](http://www.collaborationforimpact.com/collective-impact/), “multiple organisations or entities from different sectors … abandon their own agenda in favour of a common agenda, shared measurement and alignment of effort” supported by a backbone organization. The Media Consortium would act as that backbone organization, while multiple members of the Media Consortium would participate in publishing collaboratively around a shared set of topics, with their impact measured by the researchers.

Using this theory, I suggested to John and Gary that the impact of these smaller outlets would be greater than the sum of their user numbers would suggest. So instead of measuring sentiment change caused by a story published only by one outlet, I proposed that we measure sentiment change caused by the co-publication of a story by at least three outlets.

Gary and John quickly agreed, appetizers were eaten, and I thought we were well on our way. As it turned out, however, the hardest aspect of the experiment was yet to be discussed.

Over pizza, Gary explained sentiment analysis to me and what he would need from Media Consortium members. To make the analysis valid, he explained, he would have to create at least 30 experiments, and those experiments would have to have a random element. If the experiments weren’t randomized, then the researchers couldn’t be sure that they were measuring the impact of journalism. For example, most journalists time stories to break with breaking events, but then researchers might be measuring the effect of the event, rather than the journalism.

I got it. I had taken Bio 101. My dad was a scientist. Yes, we had to randomize the experiment. But then Gary launched into ways to randomize, and I stopped eating.

In biology experiments, scientists randomize by choosing what content will go into the petri dish; in medicine, they randomize by choosing what medicine to give the patient; and so, Gary suggested, we should randomize by choosing what content the news outlets should deliver to their audiences.

Gulp.

No wonder scientists had never been able to work with journalists! I explained that journalistic integrity is sacred. Editors must be free to decide what they will publish. Editors and journalists must be free to decide how they will tell the story. Telling news outlets what to publish was a non-starter.

Then I got an idea. If Gary wasn’t interested in measuring the effect of breaking stories—in fact, if he wanted to stay away from breaking stories—then we could create an experiment that would randomize when outlets publish.

I explained that all news operations assign feature stories, profiles, and other content that specifically don’t need to run on certain days. Editors use those stories as a safety net to fill in on days when the news is slow. While they wouldn’t want to hold those stories forever, it wouldn’t be a problem to ask them to hold such a story an extra week.

There was our answer. After spending almost two years to refine the concept, including running a number of tests, Gary’s team decided that they could randomize by asking outlets to pick two weeks during which they would be willing to run a story. Once those two weeks were picked, the researchers would flip a coin. Outlets had to run the stories on the week they picked—preferably on Tuesday, so the researchers could measure the impact of the whole week. The other week, either before or after, became the control week.

The next day, Gary gave a talk at the Media Consortium’s annual meeting about this project. Our members were excited about the possibility of finally being able to prove the impact of their work, but were concerned about how the experiment would work. Gary asked if outlets might be willing to hold stories so that investigators could randomize the outcome. The response was positive, although outlets suggested that they were worried that participating might lead them to incur additional costs as editors tried to match their pieces to our schedule. Voqal generously agreed to provide a small fund to allow us to compensate participating outlets for their time, and we were set. The project was on.

Of course, it took us quite a while to explain how the experiment would work to the journalists. We had to toss out several potential experimental instances that we worked hard to pull together because outlets didn’t follow the researchers rules—they published before the chosen week or waited to publish until the following week. In one or two cases, a story being held for the experiment became relevant to a breaking story, and the outlet had to publish it ahead of schedule. In a couple of other cases, the participating outlet had a story fall through. In all, Media Consortium staff created 15 instances that the researchers were not able to use, in addition to the 35 instances that were used in the actual experiment.

We also found that we needed to hire a project manager to run the project—and were fortunate to be able to bring on Manolia Charlotin to organize the collaborations (for details on the process we used read this Mediashift article). The actual collaborations ran from March 2015 to March 2016.

Patience, it turns out, is another quality journalists need to learn in working with researchers. The work we began on October 12, 2011 at Pizzaiolo did not come to life until 2015-2016 and did not really wrap up until this week, November 2017, with the publication of the researcher’s findings. And to tell the truth, this research is just the first step. There are so many more questions we want to explore now that we have these initial findings. Now that we understand the researchers’ pace, however, we will be ready for them. First step: order lots more pizza.