GESIS Guides to Digital Behavioral Data #17

Expert Insights into Crowdsourcing and Studying Human Coordination

An Interview with Mark E. Whiting

Publication date: July 22, 2024

Crowdsourcing is increasingly used for data coding tasks and also – and excitingly from a methodological (computational) social science perspective – for studying human behavior and the coordination of very large groups. We talked to Mark E. Whiting about this growing area of research and practice.

Mark is a senior computational social scientist at the CSSLab at the University of Pennsylvania working with Duncan J. Watts, in affiliation with Computer & Information Science in Engineering and Applied Science and Operations, Information and Decisions at Wharton. He was previously a postdoc under Michael S. Bernstein in the HCI group in Computer Science at Stanford University. Mark builds systems to study how people behave and coordinate at scale. Among his papers you will find a wide range of aspects covered from team dynamics, the impact of deep fakes to fair working conditions in crowd work.

The interview was conducted by Leon Fröhling and Indira Sen during the International Conference on Computational Social Science (IC2S2-23) in Copenhagen on July 18th, 2023. The interview has been edited for clarity and length.

Keywords: crowdsourcing, human computer interaction (HCI), data coding, experiments at scale, collective values and norms, collective intelligence, Empirica, research ethics

GESIS: Hello Mark, thank you for introducing us to your research with this interview! What are you working on at the moment, especially in the context of studying human behavior and coordination?

Mark Whiting: I work on a really wide range of different kinds of projects in that envelope. Pretty much everything is around some kind of human coordination or at least human understanding. On one hand, I do things around common sense, which is all about the degree to which there are shared beliefs, and groups that have consistent sets of shared beliefs in society. I use crowdsourcing and similar tools for those kinds of studies. On the other hand, I do a lot of work in designing experiments to understand team synergy or team performance, what leads to more performing in teams. Most of that is done in a lab setting, again using online participants from crowdsourcing platforms.

Another line of work is around applying some of these similar experimental design ideas in the field, working closely with companies to design experiments that try to get at aspects of causal inference or to build dynamic experimental environments that update based on things that are happening at the company in real time. We can systematically learn things while also improving the outcomes for a company. All tend to revolve around the same vision of making teams and other groups of humans more effective and productive.

GESIS: How did you first enter that field of expertise? How did you start looking into human coordination?

> ***[As] a human population one of our biggest challenges is making sure that we do not fight too much and that we manage to succeed at some large-scale collective goals.***

Mark Whiting: I spent a lot of time studying design. My undergraduate and master's studies were in industrial design. Then I did a PhD in mechanical engineering with an eye toward formalizing any of the concepts from design. I was frustrated by the sort of design literature that was built around rules of thumb and informal studies. In engineering design, the subfield of mechanical engineering that I was working in, it is common to model human preferences and to think about how you can deal with those at a large scale. This is when I realized we really need better ways to measure what people are thinking and how social contingencies impact that thinking.

That really got me in this very specific area of social computing and crowdsourcing, thinking about remote sensing on humans and trying to understand what people are doing and thinking at large scales. This relates to an ongoing interest that I have always had, that as a human population one of our biggest challenges is making sure that we do not fight too much and that we manage to succeed at some large-scale collective goals. We have not done a very good job of marching towards that outcome. We have settled on things like democracy, but I think we could go a lot further. I am really motivated by that big picture.

GESIS: Since you first started working on these topics, were there any major turning points that changed your perspective?

> ***There was a time when you could run a small study,publish it, and have a great impact. Nowadays we have different expectations in science.***

Mark Whiting: When I started working in this field, it was more centered around HCI (human-computer interaction) and using computational systems to study humans. I have shifted into the broader spectrum of computational social science now. I have worked with Duncan Watts for a while and found his lens of looking at the very big scale, the challenge of understanding not just how people are using some system, but the societal effects and the large-scale dynamics around a particular concept or shift of social practices really fascinating. I have seen that as a major shift in the way I think about and deal with this kind of work.

Another thing that has changed over time quite a lot is the way we think about what good evidence is. There was a time when you could run a small study, publish it, and have a great impact. Nowadays we have different expectations in science. This is a change that is affecting everybody, but it has been particularly a change for people who deal with methodological questions. I have been working a lot on designing better experiments, asking how to design experiments that help you know things that were traditionally very hard to know, while dealing with all sorts of complexities and challenges.

GESIS: Moving towards some of the methodology around crowdsourcing, could you briefly explain why you think that crowdsourcing platforms are such an interesting resource for computational social science?

> ***There are two distinct uses of crowdsourcing that are most relevant to computational social science. […] One is for data coding, the other is as experimental participants.***

Mark Whiting: There are two distinct uses of crowdsourcing that are most relevant to computational social science. The first is getting participants for studies where your measurements are about the people who are participating in the study. The other is getting participants who help you measure things about the stimuli, the content or the concepts that you are studying. One is for data coding, the other is as experimental participants.

Crowdsourcing has strengths in both these areas, strengths that before crowdsourcing were close to impossible to achieve. Instead of doing lab experiments where we invite a bunch of students from the local university to participate in something, we can now invite a much wider range of people. We can now even know more about them, because we can track them over time in different ways and we can integrate that knowledge in reasonable ways. It also means that we now can run experiments at a much larger scale that integrate a much wider range of conditions. All of that is tremendously useful for computational social science, especially in keeping up with our expectations for evidence, which are getting stronger and more severe. However, crowdsourcing has some limitations that arise from these same aspects. It is still the easy way to recruit participants. The new convenience sample is Amazon Mechanical Turk [1]. We are increasingly seeing people try more field studies or observational studies that try to learn accurate points about the world, without engaging with crowdsourcing directly. Since our goals and our expectations of how we build scientific knowledge are constantly changing, it will be interesting to see how crowdsourcing and the use of crowdsourcing evolves over the next few years in light of these constant changes.

GESIS: You already mentioned that you are involved in the creation of the Empirica platform [2], which is a virtual lab for online experiments. Could you tell us a bit more about that and explain how it relates to crowdsourcing?

> *[In] the crowdsourcing world we can run studies that scale up effectively. We can do a pilot with 10 people and 10 minutes later we can do an actual experiment with 10,000 people.*

Mark Whiting: There are other people who have been much more centrally involved with Empirica here, including Abdullah Almaatouq — the driving force behind the project — and James Houghton, among various others. Essentially, Empirica is a wrapper that can be applied to something like crowdsourcing, but you could just as well also use it in an experiment in a lab or an experiment in a classroom. The key point about it is that it makes very repeatable and parameterized experiments possible. And that facilitates a type of research that is outside of the scope of traditional experiments. Instead of studying single conditions at a time, with the Empricia platform that lets you run these very consistent experiments, you can start to run many different conditions at a time, because you have a lot of certainty about the consistency of the experience that participants are getting. That is married very nicely to crowdsourcing, because in the crowdsourcing world we can run studies that scale up effectively. We can do a pilot with 10 people and 10 minutes later we can do an actual experiment with 10,000 people. That is very conducive to thinking about better ways to run these kinds of experiments.

That is where I would see Empirica fitting in. There are two papers that I recommend checking out around that. One is the methodological paper that introduced Empirica [3], and in the other one we talk a lot about this new experimental method that we think is very exciting and coming [4].

GESIS: How do you think about when to use crowdsourcing and when not to use it to recruit for your experiments? Are there some taxonomy research questions for which you would and would not use crowdsourcing?

Mark Whiting: We often use crowdsourcing as a way to get people to participate in experiments, even if they are very big. We do not see them as alternative mechanisms, but more complementary. I do think there are situations where we need to think very differently about methods. In parameterizable experiments, which we also call integrative experiments, you might have a lot of experiments that you want to run to answer a series of somehow related questions about an underlying concepts.

One of the challenges is that the number of conditions for which you need to do this may be extremely large. Let us say you have a five-dimensional design space, with five dimensions that contribute to which experiment you are running. Those dimensions can be very simple, like whether or not the participants are cognitively diverse or whether the participants are doing a task that allows their effort to be combined in a reasonable way.

You quickly realize that just with a small number of dimensions, the number of possible experiments you could run may exceed the population of the Earth. You have this curse of dimensionality, and it quickly unfolds.

You quickly realize that just with a small number of dimensions, the number of possible experiments you could run may exceed the population of the Earth. You have this curse of dimensionality, and it quickly unfolds. Another component is how to integrate modeling and sampling techniques to avoid them from becoming too complicated and messy.

When thinking about whether to use crowdsourcing, you also have to think about its limits. One of the limits is that it costs money and takes time, and that people do not always enjoy certain kinds of activities. It is thus very interesting to consider whether there are ways to train a model that can reliably predict what a crowdsourced individual might do in a certain situation, so that you may even avoid running certain conditions in an experiment. A training model is one way to do this, but there are many other things that people do. You can build systems and behavior, you can build agents, we see all sorts of approaches. Crowdsourcing gives you all these strengths, but it also has these very specific weaknesses that are interesting to exploit with other methods.

GESIS: Recently there are all these ideas around large language models and how they might be affecting various aspects of computational social science research. There are, for example, papers [5] that talk about silicon samples and about replacing crowd workers with large language models. How do you see this development play out? What are some potentials and pitfalls?

Mark Whiting: It is extremely problem-specific. There are particular problems that language models and other new AI technologies are going to solve really well, and there are other types of problems that they just do not solve so well. This brings us back to the initial example of crowds as participants and crowds as data coders. Data coding often is a very expensive step in our studies, and a very hard one. It can take hours and hours of training and a lot of work to do it well, especially with very sophisticated data codes. It would be nice if we could use language models in that setting, but in our experience, even subtle changes in question types or subtle changes in context lead to significantly different answers. To us, this suggests that although the potential is there, it is not yet realizable without extensive testing.

> ***Something like bias in training data would be dramatically reinforced by assuming that you understand people based on a silicon sample, because you simply do not even know what the biases are.***

While I see that these kinds of technologies are going to change the way we use some of our resources, I do not think it is anywhere close to the level of a simple drop-in solution that a lot of people maybe assume it is. I think there is another piece here, and ideas like the silicon samples are a good example of it. Many of those approaches are just not validated in a very strong way. Something like bias in training data would be dramatically reinforced by assuming that you understand people based on a silicon sample, because you simply do not even know what the biases are. That is the whole point of biases. If you knew how to measure them, you would be able to account for them on some level.

In many studies that we run, we look at the output of human behavior, and we realize that we must be missing something, that there must be a dimension that we are not aware of, explaining changes in the data that are not explained by any other component. This is an example for which I cannot imagine a simple approach like a silicon sample working very well. I am worried that people might overly trust them and assume that they are getting representative data about humans, but that is not true.

GESIS: Are there other things that you see people constantly getting wrong when thinking and talking about crowdsourcing. Are there things that people would expect to get from it, but that crowdsourcing just cannot offer?

Mark Whiting: Often people will try crowdsourcing and they will get bad results, and they immediately blame the workers. There have been a lot of papers that say you should not do this. There is a lot of good evidence to say that that is not just cruel, but also misinformed. Almost always these bad results are not the fault of the worker, but the fault of the requester who was designing a task in a way that it could easily be misunderstood, maybe on purpose.

> ***The labor of a good requester is much more complicated than I think most people expect, especially with this ethical piece to it.***

There is a wide range of challenges in designing a good task. Some of them are simple things, like explaining what you really mean by a question very clearly, which will directly make people read and answer it more accurately. But some of them are nuanced things, like people who want to study some aspect of human behavior that ends up being ethically complicated when done in a crowdsourcing context. Some economic studies fall into this situation, where some aspect of economic decision-making is being studied, and in order to make it real, people are actually paid more or less based on their behavior and answers. Sometimes that feels like we are treating these people as guinea pigs with punitive consequences to their goodwill participation, which we need to be very careful about.

The labor of a good requester is much more complicated than I think most people expect, especially with this ethical piece to it. People think you can just throw a survey on Mechanical Turk and you will get the right results out. That is just not the case. You have to put in a lot of work to set things up properly to get the level of results that you want.

> ***It is very easy to think about crowdsourcing as anonymous or as an API call to humans. But there are humans at the other end [...]***

Another thing is the pay, which is much more complicated than almost anyone thinks. For example, there has been a lot of discussion about whether or not Amazon Mechanical Turk should institute a minimum wage. There is actually a lot of disagreement and heterogeneous perspectives within the worker community about how pay should work and what is ethical on their platform. It is very easy to think about crowdsourcing as anonymous or as an API call to humans. But there are humans at the other end, and you really need to think about how to motivate them and how to treat them like humans, which people sometimes seem to overlook.

GESIS: You worked on this challenge of paying participants fairly. Could you tell us about these efforts? And related, what are other best practices and things to keep in mind when doing a crowdsourcing study?

> ***I would say almost every study run from my lab uses fair work now.***

Mark Whiting: I am just going to talk a little bit about the fair work platform that we developed with Hugh Grant and Michael Bernstein [6], trying to resolve the ongoing challenge of figuring out how much to pay people for their participation. Usually, what requesters are doing is trying to work out how long it might take people to complete the HIT (human intelligence task). Then they are making some judgment about whether that is a reasonable amount of time, whether people are procrastinating for some reason or speeding through it for some other reason, so that in the end the time spent would not be as high as they thought it was going to be. Fair work tries to minimize that problem by promising a wage of $15 an hour. The particular number is not so critical here, but promising a particular wage and ensuring that this wage will be established based on how long it actually takes the median worker to complete the task. We use this technique very widely, I would say almost every study run from my lab uses fair work now.

I think many other labs have adopted similar policies. Overall, I think payment remains complicated, but it dramatically simplifies the problem if there is this very simple rule-based system to decide how much you should pay people before you try to run a study.

Another dimension to always think about is the relationship between the requester and the workers. Paying fairly actually is a very good way to build up a strong relationship with workers that can even lead to getting more high-quality workers because of reputational points on various external platforms. It can also lead to more insights about what is going on with our studies. If we have strong relationships with a lot of workers and a study is not working well, they will often send us an email that turns out to be very helpful to understand how we can fix what we are doing.

> ***I have noticed that the biggest difference for getting the best results is to have an open mindset for truly collaborating with the workers.***

Really collaborating with these participants, the workers on Mechanical Turk or on other platforms like Prolific [7], makes a dramatic difference in the quality of results you can expect. Part of that relationship is the payment, another part is treating them as real humans who you can have conversations with, and finally it is about being open to having processes that are open to that kind of evolution and development.

I have done a lot of crowdsourcing, and I have noticed that the biggest difference for getting the best results is to have an open mindset for truly collaborating with the workers. They know a lot about what they want and need to be successful in doing a task. Many of these people are doing crowd work as a way to earn a living, not as a way to spend all day cheating other people, and they are not necessarily doing it as a way to just consume time. They legitimately want to do it for a wide range of reasons that have been well studied. But we need to respect that and support their interests just as much as we support our own.

GESIS: In your talk at this conference (IC2S2-23) about the population level common sense, one of the things that you were talking about is how difficult it is to identify this common sense and that there is a very narrow set of things that could pass for common sense. Do you think this has implications for wisdom of the crowd approaches and data coding tasks?

> ***[…] do your own HITs repeatedly […] do your own online activities and experiments repeatedly […]***

Mark Whiting: The wisdom of the crowd space has really nice work on that question, speaking to the fact that certain kinds of questions are going to be worse if you aggregate individuals to answer them. Other kinds of questions might be very accurate. As long as you are aware of those properties, when you do things like data coding, you can be OK. The challenge is, however, that it is often hard to know the properties of the question that you are asking before you have seen quite a lot of answers to it, or maybe before you yourself have tried to answer it. This speaks to another thing that I highly recommend, which is to do your own HITs repeatedly, to do your own online activities and experiments repeatedly, so that you really learn the ins and outs of what makes them good and what makes them complicated and tricky.

I think what this common sense study (which is now out [8]) is really saying is that people do not necessarily disagree on big things, but they might disagree on many little things.

Many details might just not quite align. If you are aware of this, it does have implications to the way we crowdsource. Many people try to deal with this by, for example, averaging a series of different responses. That probably works a good deal of the time. This idea that the level of agreement on small detailed things is much less than you expect could explain variability in data that we had assumed was just weirdly non-specific and heterogeneous. It could be that individual level properties are much more productive there. This is important, but not very easy to consider. Something that we and many other labs do when it comes to a hard task is that we will give people a lot of training so that we can know if they are going to be good at it. We will then continue to test them repeatedly to see if they are as good as we thought they were. In that setting you can deal with this sort of heterogeneous response a little bit better.

At the end of the day, it really comes down to the specific properties of the questions you are asking. Answers to some questions are just always going to be inconsistent. I think it is safe to assume that it is at least in part due to the differences between individuals.

GESIS: You already talked about the importance of building good relationships with crowd workers and training them properly for data coding tasks. But how does this kind of training look and how does it differ from the more traditional approach of training students or research assistants for such a task?

Mark Whiting: Many crowd working platforms provide some ways to deal with training. For example, on Mechanical Turk, you can set up tests for people to take and they can get qualifications based on the tests they do. The test can involve things like reading a text, thinking about it, and then thoughtfully answering a few questions. This way, you can really check if your workers are good at a specific activity.

In addition to the training aspect, one of the huge benefits of having research assistants is the two-way communication, meaning that you are getting value out of them which you might easily overlook if it was just one-way communication. We try to bring this to crowdsourcing, engage the participants or workers and actively talking to them about what they think about the tasks you are asking them to do.

One way that we sometimes do this is by making versions of a survey for data coding that, instead of having only the questions that we want them to answer, have a secondary question that just asks about how hard it was to answer or if they can help us understand any confusion about that question. This way we can systematically and continuously improve all the materials together. I think it is a very successful technique for bridging the gap between what you get from a research assistant and what you get from a crowd worker.

There are other things, too, that play a role, like education level or alignment on background. Especially for relatively straightforward coding tasks a majority of the difference comes from that, meaning that in fact, if you do a good job, you can account for all those differences effectively.

> *In a few cases that we have worked on, we spent hundreds of hours refining survey instruments for coding certain kinds of data because the questions were so nuanced and it was so challenging to get right.*

There are also times where you want to do very challenging coding tasks, which is also possible to do with crowd workers. But you have to spend a lot of time getting the materials right. In a few cases that we have worked on, we spent hundreds of hours refining survey instruments for coding certain kinds of data because the questions were so nuanced and it was so challenging to get right. Eventually we were satisfied with the results that we were getting out of our MTurk participants, but it required setting up specialized qualification tests, meaning that some group of people had answered a lot of these and knew exactly what they had to look for. It also meant refining our own questions a lot and being very careful to ask them in a very specific way.

GESIS: For people that would want to get started with crowdsourcing their studies, are there any resources that you could recommend? You already mentioned some of the papers and people working in the space, but maybe there are some additional hands-on tutorials and best practice guidelines?

Mark Whiting: There are often tutorials on combining crowdsourcing and experiments at conferences. There are also workshops about Empirica at many conferences now. I think materials like that are probably very useful for people who are thinking about doing experiments on one of these platforms.

Data coding is a little bit more complicated, because the properties are very specific to the particular problem and kind of data coding you are looking at. Mechanical Turk and Prolific both now provide pretty good resources for getting started on their platforms. There are also pretty good communities around each of those platforms where you can go to see the kinds of dynamics that workers are expressing and discussing. You can also see the kinds of tasks that they are enjoying and not enjoying. And of course, in many of those communities, you can just go and talk to them. However, this might not always be the best option. If you are thinking of doing studies that might annoy people, you need to be really careful. We have had very heated interactions with workers who thought we were cheating them. Even though we managed to resolve these cases eventually, that part can be a little bit stressful. And so, again, I would recommend to always keep in mind that you are dealing with other people's time.

There are also many resources around orchestrating hits, managing compensation, like fair work, or other wrappers for Mechanical Turk that can be worth playing with. Getting familiar with the API tools can dramatically change a person's outlook on building systems with these kinds of platforms.

One other thing that is very much worth it and that I ask – if they can – all my students to do is to create an account on Mechanical Turk and use it to earn a few dollars. You will learn a lot of interesting things about the workers' experience on the platform, like the kinds of tasks that they are getting, the work that they have to do to maintain their reputation and the interactions that they have to engage in to satisfy requesters.

It makes it much easier to be a better requester if you have tried being a worker.

It makes it much easier to be a better requester if you have tried being a worker. I say I ask them to try this if they can because there are a lot of limitations on who can set up accounts on these platforms. If you are in the United States, you have to be American if you are trying to attach it to an American bank account, or you have to have a green card. You have to be a permanent resident of some sort. In other countries, they probably have a similar style of law, because they file taxes for you. There have been a lot of researchers talking about doing this, especially in the HCI community, they have tried to be very supportive of the worker outlook.

With Michael Bernstein and other colleagues, I have done some work on trying to understand how to build better communities for workers [9], trying to understand even how to build better novel platforms [10]. But there are also many other researchers who have written qualitative reports on their own time trying to earn money on one of these platforms. In trying so on your own, or even in reading these reports, you realize all sorts of interesting aspects of that worker experience that would be totally hard to guess if you did not try.

GESIS: If you could make a wish to the universe for some type of research artifact that would help facilitate your research in this area. It could be a package, an app or a whole research agenda -- what would that wish be?

Mark Whiting: Lots of things. However, there is one thing that I am particularly interested in. Internally, we build a lot of infrastructure. We have a panel of many thousands of MTurk workers whose data we integrate across studies, and we have the appropriate IRBs (Internal Review Board) to do that. This gives us much richer information and lets us do higher-level experiments. We are also building similar kinds of datasets of papers, for which we are then doing this data coding I was talking about earlier. This gives us large bodies of literature that we want to study consistent properties of.

What to me would be an exciting future research infrastructure are systems to aggregate the individual level behaviors, you could call this 'large-scale digital trace aggregation', in such a way that can do reasonable inference across this data without exposing any PII (personal identifiable information), but also to do it at a scale that would give you something far beyond crowdsourcing. There are a few platforms that are trying to do this. Mozilla Rally [11, 12] for example, tried a way for people to opt in to collect lots of data about their web usage, as an integrated digital trace system, and then provide that data to researchers in a secure or anonymous fashion.

Extending this vision to much more data so that we can do amazingly rich causal inference on things that are happening in the world would really change the way we think about the social sciences and crowdsourcing in particular.

GESIS: Thank you very much for the interview, Mark, thanks a lot for these insights!

References

1 https://www.mturk.com/

2 https://empirica.ly/

3 Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158-2171. https://doi.org/10.3758/s13428-020-01535-9

4 Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences, 47*, e33. https://doi.org/10.1017/S0140525X22002874

5 Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

6 Whiting, M. E., Hugh, G., & Bernstein, M. S. (2019). Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, pp. 197-206). https://doi.org/10.1609/hcomp.v7i1.5283

7 https://www.prolific.com/

8 Whiting, M. E., & Watts, D. J. (2024). A framework for quantifying individual and collective common sense. *Proceedings of the National Academy of Sciences*, *121*(4), e2309535121

9 Whiting, Hugh, & Bernstein, 2019, see ref. 6.

10 Gaikwad, S. N. S., Whiting, M. E., Gamage, D., Mullings, C. A., Majeti, D., Goyal, S., ... & Bernstein, M. S. (2017). The daemo crowdsourcing marketplace. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1-4). https://doi.org/10.1145/3022198.3023270

11 https://blog.mozilla.org/data/2021/05/05/announcing-mozilla-rally/

12 https://github.com/mozilla-rally

All links in the text and in the references were retrieved on July 15, 2024.

Suggested citation

Whiting, M. (2024). *Expert Insights into Crowdsourcing and Studying Human Coordination. An Interview with Mark E. Whiting* (GESIS Guides to Digital Behavioral Data, 17). Cologne: GESIS – Leibniz Institute for the Social Sciences.

Series editors

Danica Radovanović, Maria Zens, Katrin Weller, Claudia Wagner

Publisher

GESIS Leibniz-Institute for the Social Sciences, Cologne