Chapter 14. Considerations for Internet Participant Selection: Algorithms, Power Users, Overload, Conventionalization, and Participant Protection

John R. Gallagher
University of Illinois

> Purposeful [participant] selection, though, is more than a technique to access data; our selection choices frame who and what matters as data (Freeman, 2000). These choices interface the other methods in a study to ultimately become the stories that are told. Consider, for example, the intersection of participant selection and interview analysis. The participant's story is embedded in a matrix of researcher choices: research questions, selection criteria, interview style, analysis technique, and countless other choices. Thus, purposeful selection is a mechanism for making meaning, not just uncovering it. From this perspective, purposeful selection is epistemological; researchers construct versions of reality grounded in their selection choices. (700)
>
> *– Earl Reybold, Jill Lammert, and Stacia Stribling*

Theorizing participant selection needs to account for a small number of users who are responsible for a large amount of internet activity. Influencers and celebrities dominate Twitter, thereby creating misperceptions about how often the "average" Twitter user participates. Users who participate frequently and have influence on other users often have different perceptions and habits than those who are lurkers or users who participate less often. In terms of methodology, these "power users" may skew dataset averages because data collected about them are not representative. Recruiting power users is but one key consideration when selecting participants.

This chapter addresses five challenges for participant selection with respect to internet research. These challenges are 1) algorithms, 2) power users, 3) overload of possible participants, 4) conventionalization of experiences, and 5) participant protection from online toxic communities. In doing so, I advocate for active reflection on the ways participant selection processes shape an empirical internet-based study. Reflecting on participant research helps to "question our own assumptions" to "actively" and "progressively" change our own habits (Agboka 299). In turn, this questioning can begin to address cultural hegemony in academic research (Agboka 299).

First Challenge: Algorithms May Invisibly Shape Participant Selection

Algorithms can shape how participants are selected. Algorithms spotlight, often invisibly, atypical users, such as power users. Algorithms are mathematical expressions used to scale up human-based decisions. If we critique algorithmic bias, then we critique human bias that has been transformed into *automated routines*. One aspect that separates algorithms from their programmers and designers, however, is their ability to make decisions without human input. From this perspective, algorithms possess a sense of agency that is akin to human agency but is still dependent on human agency. Algorithms have been around since ancient times. For example, Babylonians used algorithms for factorization. For the past several decades, grocery store companies have used algorithms to determine the items customers are likely to buy based on their shopping history. Web crawlers, such as Google, index the internet via algorithms that use keywords and other variables to systematically make the world wide web searchable. With the rise of vast, real-time social media networks and the scale of digital infrastructure, algorithms display, collate, and filter users that researchers can see.

Considering how algorithms shape participant selection is a key challenge for internet researchers. We need to account for the degree to which our access to participants is determined by algorithms or other models that sort users automatically. What we encounter as researchers on our own web pages or social media feeds is not what all users encounter. This issue is not a post-modern or post-structuralist concern either: web pages are literally different depending on the algorithms used to tailor web pages based on user activity, account history, and tracking data.

Algorithms typically use some form of reach, or what we might colloquially call popularity, to sort users into various "bins." Some bins are easily viewed whereas others fall into obscurity. These differences lead to websites literally appearing differently to different users. A concrete example of this algorithmic differentiation is algorithmic price discrimination. Synthesizing multiple studies, law professor Oren Bar-Gill writes the following:

> Uber, Amazon, Staples, and the online video game store Steam were found to vary price by geographic location and, in Uber's case, also by the time of day. B&Q, a British multinational company, tested in its brick-and-mortar stores digital price tags that interfaced with customers' phones and adjusted the displayed price based on the customer's loyalty cards data and spending habits. Grocery stores are experimenting with digitized and personalized pricing using e-coupons. Allstate was criticized for optimizing prices based on its calculated like-

> lihood that individual users would comparison-shop before purchasing insurance. (218)

Price discrimination presents methodological concerns for internet researchers: web pages are not only interactive ("Web 2.0"), but they are also customized. When researchers recruit internet users to interview, survey, and observe, those participants may likely experience different web pages. To be clear, this algorithmic price discrimination and other algorithmic determinations are not simply a matter of different interfaces due to accessing websites via mobile or desktop technologies; this issue can be identified by asking participants how they access the internet. Rather, algorithmic determinations are less identifiable because they function in the background of web pages, often taking innumerable variables into account. Consider that social media networks such as Facebook, Instagram, LinkedIn, TikTok, Twitter, and YouTube use thousands of variables to determine social networks newsfeeds, i.e., who sees what posts.

Accounting for what participants view on their web pages clarifies selection criteria by making those criteria explicit rather than assumed. It leads to transparent research procedures and consistency for comparison across participants. I have found three ways to account for algorithms when recruiting participants. First, I ask participants about the advertisements they see on their respective digital accounts. Second, I request some screenshots from participants. These images can help researchers determine how participants' view of web pages differ from our own. Screenshots can also help researchers more clearly determine differences between power users and non-power users by documenting how notifications and other interfaces shape participants' viewpoints and perspectives. For example, power users are likely to get many notifications. Third, before formally recruiting participants, I ask potential participants about their perspectives about algorithms and the degree to which algorithms play an active role in their activity. I usually do this over email. This third consideration helps me to understand participants' own metacognition about how they are sorted by algorithms.

Second Challenge: Power Users

The second challenge to participant selection, as I've alluded to, is to identify "power users" to understand how selection will shape research findings. The term "power" does not imply that power users are good or bad. I also want to avoid a conflation of the term with Foucauldian notions of power. Rather, "power" users possess an ability to hold influence over other users and generate a larger amount of activity when compared to typical users. I use power to invoke the concept of power laws, or that idea that the functional relationship of one variable changes in proportion to the other. Long tail functions are one example of power laws. An anecdotal example of power laws is the "1 percent rule of Web 2.0 culture" wherein one percent of users are responsible for the content and the rest of users do not produce content.

Internet researchers have found the phenomenon of power users in a variety of contexts and digital spaces. I found it in my research of a Facebook forum (Gallagher *Interactive*; Gallagher "Five") and in the distribution of commenter activity on articles in *The New York Times* (Gallagher et al.). I found that out of ~450,000 comments from *The New York Times*, the top 0.49 percent of commenters account for approximately 17.2 percent of all comments (163). Technical communication scholars Liza Potts, Rebekah Small, and Michael Trice found a similar pattern in that nine users (out of tens of thousands) were responsible for 20 percent of posts on reddit (359). Communications scholars Todd Graham and Scott Wright found an even more radical distribution of what they call superparticipants, wherein .4 percent of users were responsible for 47 percent of a forum's activity (631). Brian Weeks, Alberto Ardevol-Abreu, and Homero Gil de Zúñiga performed a survey of opinion leaders, determining that "power users" (my term) can leverage their active participation into real influence on other users and for online forums themselves.

Attending to powers users is important because I worry that my own work is guilty of recruiting power users and not more representative of typical users on their specific platform. This worry is informed by my personal life: I am married to an engineer who performs cutting-edge research in the area of cavitation rheology as well as cutting of soft materials (think here of cutting polymers). This engineer consistently challenges my research, calling into question many unstated assumptions. Also called into question are issues of generalization: how can I make decisions and conclusions based on just a few case studies? I need to clarify and reflect on my choices not only in my scholarship but also in my daily life whenever I discuss my research in a domestic setting.

Understanding how a power user can impact data was, for me, a useful lesson on how to learn of the importance of power users' participation (and whether to recruit them or not). In my dissertation, I conducted observations of a private Facebook group because I was interested in how writers develop strategies for incubating participation in a closed group. The group discussed politics, often in heated fashion. I obtained IRB access and consent from the administrator, Tracy Monroe (pseudonym), and the members themselves. I had a hunch that Monroe was not only administering the group (deciding who could join and who needed to be removed) but also actively managing the group. Monroe set rules and told me that she tried to model behavior for the group by posing questions and initiating posts.

My hunch about what was happening with Monroe's case, however, was not enough to indicate the *degree* to which Monroe was managing the group. I had many *assertions* about Monroe's activity, but my engineering spouse asked me to defend my claims with data-driven methods. For this reason, I web-scraped the entire forum (August 1, 2012 - August 1, 2013), which resulted in 5622 total posts and comments. Monroe was responsible for ~28 percent of all posts and ~26 percent of all comments. The average character length of her posts was 377 charac-

ters whereas the average character length of the rest of the entire group was 250. Given that the forum had 129 members at that time, these numbers demonstrated that Monroe was a power user. She was the most frequent group member, and her posts were the longest. She exerted a great amount of influence on the group.

This large dataset helped me to contextualize and describe Monroe's activity as a power user and to describe other group members as more or less typical. For instance, when Monroe wrote short posts or comments, I could consequently describe this behavior as atypical for Monroe's own activity, but these shorter posts reflected the activities of the typical member of the Facebook group under study. Without the general trends of the group, I would have had no broader context to fall back upon when describing forum members' individual posts or Monroe's activity.

Monroe's case study helped prepare me for being careful about participant selection and performing a lot of contextual work *before* contacting possible participants. In this sense, it was a crucial experience for what I call "selection context" or what Reybold, Lammert, and Stribling call "subjective focus" (701). For them, selection is more than a rote set of choices. They write:

> Selection as method requires researchers to be aware that choosing sites and participants for our research is more than a technical process. As Peshkin (2001) reminded us, these choices are the 'selection and choice of what to perceive' (p. 251). How we perceive the research issue impacts who we perceive to be at the core of that issue and thereby what we hope to learn from those whom we have identified. (703)

With this perspective, Reybold, Lammert, and Stribling advocate for selection as an extension of researchers' "…theoretical and conceptual framework" (702). From Monroe's case, I learned I needed to call my own assumptions as well as my participants' assumptions about their digital activity into question because the scale of internet research makes it difficult to determine how different users view such networks. For example, as Kristin Arola ("Design") argues, digital technologies that use standardized interfaces ascribe behaviors to users, thereby engendering normative, often colonizing, behaviors. These behaviors can simultaneously lead users to perceive that other users are having similar experiences to their own. But as Arola ("Land-Based") has demonstrated, there are a variety of digital designs possible to users, which can produce a litany of possible experiences—possibly overloading researchers with too many participants to address.

Third Challenge: Overload

For internet research, the sheer volume and variety of participants presents practical considerations. In many ways, internet research inverts typical participant selection: the problem isn't a lack of willing participants but too many possible

participants. Recruitment thus becomes less about finding participants and more about developing detailed selection criteria for participant recruitment and winnowing down possible pools of participants for sampling.

Selection procedures are the most important step for researchers (internet or not) to develop because they help researchers identify who should or could be included in a research design. We need to identify and interrogate our 1) own research inquiries, 2) the tools we have at our disposal, and 3) our methodological and epistemological commitments. I avoid the phrase "research questions" in the previous sentence because we often cannot formulate our inquiries into formalized questions until we have identified tools, conducted background research, and thoroughly consulted relevant literature. From this standpoint, selection procedures (not yet principles of selection) are an iterative, non-linear, and recursive process. In this way, selection procedures generate principles of selection, or the categories and elements that help me to determine who I should ask to participate in my research projects.

When I am beginning a new project, I initially sketch out these three elements on a blank piece of paper. I usually make three columns. I prefer unlined paper because I can draw arrows to generate connections between my inquiries, tools, and commitments. I try to avoid using the screen because I have ADHD and screens tend to overstimulate my thought process and my eyes. However, developing a personalized process is part of the research process. I have found this personally rewarding too, as it helps me to think iteratively through my own perspectives while I formulate questions. Upon developing these procedures, I consider the *principles of selection* that determine the types of people I aim to recruit.

Principles of selection allow internet researchers to grapple with overload in coherent ways. With the development of circulation studies (Gries; Edwards "Circulation"; Eyman) and spreadable media (Jenkins), data, discourse, and messages are on multiple platforms simultaneously, often with contradictory audience reception and varying amplifications. By amplification, I mean different messages can be increased or decreased depending on discourse producer, audience reception, platform, and interface. Due to this overwhelming amount of information, we need reflexive, detailed principles of selection that use some form of real-time analysis or note-taking.

Sara Riddick has offered one such approach through what she calls "digital drifting" or where researchers take notes on the affective nature of real-time events that are streamed through social media platforms. Riddick's approach calls for researchers to observe the live reactions used on Facebook or YouTube video to gauge how audiences receive a particular message, which in Riddick's case are political speeches. Riddick's approach can be leveraged effectively as a tool for selecting participants because researchers can find audiences who are reacting to discourses and attempt to recruit those users.

More broadly, principles of selection encourage us to inculcate higher awareness of sampling techniques and approaches. All researchers should aim to limit

under/over sampling and eliminate sample bias.[1] Under and over sampling in-volves taking too many or too few datapoints, respectively, from a dataset. For participant selection, that dataset is potential participants. To be clear, principles of selection may focus on a particular group, leading researchers to focus on one class or group of participants (this focus is not sample bias but researcher se-lection). Oversampling, in qualitative human subjects research, occurs less fre-quently due to the labor involved. Under sampling, conversely, occurs frequently due to the labor-intensive process of in-depth qualitative research, such as eth-nographies that require time intensive participant observations and interviews.

I have found the following principles of selection to be especially useful: ac-cessibility, time frame(s), and participant's knowledge about research inquiries. For internet research specifically, however, other principles might need to be con-sidered. Platform usage, audience reception (for example, comments), and likes/retweets/views and other qualitative affordances (Tarsa) are three such principles of selection. One principle of selection I found useful in my work was to reori-ent principles of selection to include users who had large amounts of audience reception, in other words, lots of comments. These users, who often were power users themselves, could thus speak about considering audiences after their texts were published, thereby enabling me to determine the activities writers engage after the publication of their work. Another principle of selection could include whether (and to what degree) digital writers respond to their audiences. These two principles could be applied to non-internet research, but they are both espe-cially important for internet researchers who aim to account for digital writers and content producers who circulate their work on platforms such as TikTok, Twitter, Snapchat, Reddit, and Facebook.

Fourth Challenge: Conventionalization of Experiences

Conventionalization is an expectation of regularized and routinized patterns of behaviors. Conventionalization is a key issue with respect to participant selec-tion. Internet researchers need to be aware how and to what extent the partici-pant responses they receive are manufactured not by experiences but by techno-logical templates, cultural norms, and individual memories. Derek Edwards and Neil Mercer, in a study of classroom conventionalization, describe the concept as a "cultural basis of thinking and remembering, especially with the process of 'conventionalization,' through which cultural symbols, signs, and texts, and the mental schemata that used them, took on their recognized properties" (92). Digital networks because they use prefabricated, standardized templates (Arola "Design"), encourage users to conventionalize their experiences. These standard-

---

1.   Sample bias is different than researcher bias. Researcher bias is inescapable, large-ly due the methodological and epistemological commitments researchers bring to their projects.

ized interfaces enable possible participants to extrapolate their experiences to the general user or the culture of the internet platform or digital space.

For example, I am likely to encounter dramatically different conventionalization if I recruit participants on Reddit who identify as women versus those who identify as men. Alternatively, if I recruited power users, these participants are likely to conventionalize and routinize replying as a norm of the forum or space whereas typical users would report replying less frequently. In my own research, I try to limit making cultural extrapolations inferences from a single participant or even group of participants unless I have a representative sample that properly samples the population under study.

In my experience, conventionalization is difficult to identify until a large pool of participants has been recruited. Once a sample has been identified and recruited, I tend to rethink my recruitment procedures because my current participants may be warping (in good or bad ways) the data I have been collecting and analyzing. I attempt to account for conventionalization via this reconceptualization process. But, as with all naturalized routines and habits, conventionalization is simply an important element that researchers must be aware of when they select participants.

Fifth Challenge: Participant Protection from Toxic Communities

The final challenge I address in this chapter is considering how to protect participants' identities as *part of the recruitment process*. Internet-based participants face a greater threat from their participation in academic research because the scale of possible threats and harm is greater than if they were an offline participant. It's also important to remember this harm is also possible for researchers who identify as women, something that researchers Derek Sparby, Adrienne Massanari, and Whitney Phillips have addressed in their scholarship. Identifying possible publication venues and what happens to scholarship after it is published is vital to protecting participants. For example, considering where articles are stored, such as public venues, should be considered when participants are recruited. Related, testing participants' activity through online search and determining if participants are at-risk for coming up in easy-to-access searches could be an element in participant selection.

With respect to this latter element, my personal preference is to search online for participants' identities and activities before recruiting them as a participant. I tend to collect numerous texts from a potential participant, usually via an automated process called web-scraping. After I collect those texts, I plug in different sentences, phrases, and "turn of phrases" from the participant to determine if those texts come up in a Google search. This prevents future participants from being targeted by toxic online communities if that community accesses my research.

Conclusion: Granularity as a Response to Scale

All researchers engaged in qualitative human subject's research need to develop metacognition about participant selection. It's an important step but can be overlooked as an uncomplicated one. It is even more critical for internet researchers due to the scale of digital networks. While this chapter has addressed five challenges for participant selection, there are many that remain unaddressed. All these issues grapple in some way with the idea of scale—because the internet is, after all, a massive network.

While computer scientists and engineers try to model massive network behaviors and address scale in their research, *granularity* is an alternative answer to the question of the internet's massive scale for qualitative researchers. And it's one that writing studies and other qualitatively oriented fields are equipped to address. The stories of participants, told in detail, help to make the internet more than a set of websites driven by corporate profit and user data. How, why, what, and when user-participants communicate, write, inscribe all points to the granular detail needed for internet research. When selecting participants, then, I advocate for finding participants who can narrate their digital experiences in detail and who have extensive records of their digital lives. More importantly, internet researchers need to dwell in the spaces of their participants, likely even before recruiting them. I believe, then, that determining why each participant is selected makes for good practice. Being considerate about each participant's narrative could push a research project forward helps to be deliberate about how and why participants are selected.

Works Cited

Agboka, Godwin Y. "Decolonial Methodologies: Social Justice Perspectives in Intercultural Technical Communication Research." *Journal of Technical Writing and Communication*, vol. 44, no. 3, 2014, pp. 297–327, https://doi.org/10.2190/TW.44.3.e.

Arola, Kristin L. "A Land-Based Digital Design Rhetoric." *Routledge Companion to Digital Writing and Rhetoric*, edited by Jonathan Alexander and Jacqueline Rhodes, Routledge, 2018, pp. 199-213.

---. "The Design of Web 2.0: The Rise of the Template, The Fall of Design." *Computers and Composition*, vol. 27, no. 1, Mar. 2010, pp. 4-14.

Bar-Gill, Oren. "Algorithmic Price Discrimination When Demand Is a Function of Both Preferences and (Mis)Perceptions." *The University of Chicago Law Review*, vol. 86, no. 2, 2019, pp. 217–54.

Edwards, Derek, and Neil Mercer. "Reconstructing Context: The Conventionalization of Classroom Knowledge." *Discourse Processes*, vol. 12, no. 1, 1989, pp. 91–104, https://doi.org/10.1080/01638538909544720.

Edwards, Dustin W. "Circulation Gatekeepers: Unbundling the Platform Politics of YouTube's Content ID." *Computers and Composition*, vol. 47, 2018, pp. 61–74, https://doi.org/10.1016/j.compcom.2017.12.001.

Eyman, Douglas. *Digital Rhetoric: Theory, Method, Practice*. U of Michigan P, 2015, https://doi.org/10.3998/dh.13030181.0001.001.

Gallagher, John R. "Five Strategies Internet Writers Use to 'Continue the Conversation.'" *Written Communication*, vol. 32, no. 4, 2015, pp. 396–425, https://doi.org/10.1177/0741088315601006.

---. "Interactive Audience and the Internet." 2014, U of Massachusetts-Amherst, dissertation, https://scholarworks.umass.edu/dissertations_2/190/.

Gallagher, John R., Yinyin Chen, Kyle Wagner, Xuan Wang, Jingyi Zeng, and Alyssa Lingyi Kong. "Peering at the Internet Abyss: Using Big Data Audience Analysis to Understand Online Comments." *Technical Communication Quarterly*, vol. 29, no. 2, 2020, pp. 155–73, https://doi.org/10.1080/10572252.2019.1634766.

Graham, Todd, and Scott Wright. "Discursive Equality and Everyday Talk Online: The Impact of 'Superparticipants.'" *Journal of Computer-Mediated Communication*, vol. 19, no. 3, 2014, pp. 625–42, https://doi.org/10.1111/jcc4.12016.

Gries, Laurie. *Still Life with Rhetoric: A New Materialist Approach for Visual Rhetorics*. Utah State UP, 2015.

Jenkins, Henry. *Spreadable Media: Creating Value and Meaning in a Networked Culture*. New York UP, 2013.

Massanari, Adrienne. "# Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures." *New Media & Society,* vol. 19, no. 3, 2017, pp. 329-46.

Phillips, Whitney. *This is Why We Wan't Have Nice Things: Mapping the Relationship Between Online Trolling and mainstream culture*. MIT Press, 2015.

Potts, Liza, et al. "Boycotting the Knowledge Makers: How Reddit Demonstrates the Rise of Media Blacklists and Source Rejection in Online Communities." *IEEE Transactions on Professional Communication*, vol. 62, no. 4, 2019, pp. 351–63, https://doi.org/10.1109/TPC.2019.2946942.

Reybold, L. Earle, Lammert, Jill and Stribling, Stacia. "Participant Selection as a Conscious Research Method: Thinking Forward and the Deliberation of 'Emergent' Findings." *Qualitative Research*, vol. 13, no. 6, 2013, pp. 699-716, https://doi.org/10.1177/1468794112465634.

Riddick, Sarah A. "Deliberative Drifting: A Rhetorical Field Method for Audience Studies on Social Media." *Computers and Composition*, vol. 54, 2019, pp. 1-27.

Sparby, Erika M. "Digital social media and aggression: Memetic rhetoric in 4chan's collective identity." *Computers and Composition*, vol. 45, 2017, pp. 85-97.

Tarsa, Rebecca. "Upvoting the Exordium: Literacy Practices of the Digital Interface." *College English*, vol. 78, no. 1, 2015, pp. 12–34.

Weeks, Brian E., Alberto Ardèvol-Abreu, and Homero Gil de Zúñiga. "Online Influence? Social Media Use, Opinion Leadership, and Political Persuasion." *International Journal of Public Opinion Research*, vol. 29, no. 2, 2017, pp. 214–39, https://doi.org/10.1093/ijpor/edv050.