White Paper:
Experiences Using Online Testing to Support Survey-Methods Research and Pre-Testing in the Federal Government

Originally published: June 2019

Contributing authors

Aleia Clark Fobia, U.S. Census Bureau

Jessica Graber, U.S. Census Bureau

Jessica Holzberg, U.S. Census Bureau

Robin Kaplan, U.S. Bureau of Labor Statistics

Brandon Kopp, U.S. Bureau of Labor Statistics

Kashka Kubzdela, National Center for Education Statistics

Bill Mockovak, U.S. Bureau of Labor Statistics

Rebecca Morrison, National Science Foundation

Paul Scanlon, National Center for Health Statistics

Erica Yu, U.S. Bureau of Labor Statistics

For correspondence, contact
Erica Yu, U.S. Bureau of Labor Statistics
Yu.Erica@bls.gov

Table of Contents

1. Introduction

Federal agencies are increasingly interested in supporting research activities such as questionnaire pre-testing and evaluation, cognitive interviews, usability testing, and methodological research with *online testing* methods. In this document, the term "online testing" refers to the automated collection of information from online participants who voluntarily participate in these research activities.

This document has two purposes. Firstly, it is intended to provide a foundational understanding of what online testing is and what how researchers working within the federal system can benefit from it. Secondly, the document explores experiences conducting online research within the federal system, as well as steps to navigate the procurement process directly or through contractors.

This document does not explain how to design or conduct surveys or use specific online data collection platforms. Rather, this document is an aggregation of the authors' experiences that may be useful for our peers as they expand their use of online platforms and services to support developmental research activities. This document does not discuss using online surveys for other purposes given concerns about representativeness. This document does not represent official federal policy or standards.

1.1 Who is this document for?

This document may be useful for federal employees and contractors using online methods to collect information to aid in the development of surveys. Although this scope is narrow, we note the experiences and insights discussed here may be useful also for researchers with other purposes or applications. Throughout this document, basic knowledge of surveys and research methods is assumed.

1.2 How should I use this document?

While the state of online testing at federal agencies is always changing, there are steps that researchers can take to be as prepared and informed as possible. The first three sections of this document, which should be read as a whole as early as possible in the process of beginning online testing, cover the groundwork for understanding online testing methods:

What are online panels? (Section 2)
Who are online participants? (Section 3)
How are online data collected? (Section 4)

Once readers have this foundation, the next five sections of the document outline concrete steps to prepare for and conduct online testing:

Deciding whether to conduct online testing (Section 5)
Choosing recruitment and data collection methods (Section 6)
Recruiting a sample (Section 7)
Fielding a study (Section 8)
Analyzing the data (Section 9)

And finally, this document ends with considerations specific to conducting online testing within the federal system:

Federal and agency-specific policies governing data collection (Section 10)

If you hire a contractor to conduct a study or some part of a study for you, ask the contractor to discuss these issues with you.

In the appendix, we provide case studies that describe real-life applications of online testing at federal agencies, including examples of research questions and study designs.

For information beyond the federal government focus of this discussion, other public documents, such as the American Association for Public Opinion Research (AAPOR) Report on Online Panels[1] and the ESOMAR/GRBN Guideline for Online Sample Quality[2] may be useful.

For tutorials on how to use specific online platforms, please refer to that vendor's website or other online resources. Throughout the document and addenda, the naming of a vendor in an example does not constitute or imply endorsement or recommendation, and we offer no technical support in resolving problems encountered using these platforms.

[1] AAPOR Report on Online Panels (2010). https://www.aapor.org/aapor_main/media/mainsitefiles/aaporonlinepanelstfreportfinalrevised1.pdf

[2] ESOMR/GRBN Guideline for Online Sample Quality (2015). https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-GRBN_Online-Sample-Quality-Guideline_February-2015.pdf

2. What are online panels?

Online panels are pools of individuals willing to participate in online studies. Unlike how researchers recruit for most in-person pre-testing projects at federal agencies (e.g., directly recruiting individual participants), researchers conducting online testing typically purchase access to participants from vendors who create and maintain online panels.[3]

The American Association for Public Opinion Research (AAPOR) has written two publications on online panels and non-probability samples (see AAPOR, 2010, 2013). We refer readers to these documents for a comprehensive review of this topic.

2.1 Types of online panels

The primary distinction between online panels is the way that the panelists join the panel. There are two kinds of online panels: probability and non-probability.

2.1.1 Non-probability panels

Non-probability panels, such as SurveyMonkey, Qualtrics, TryMyUI, and Amazon Mechanical Turk, rely on the basic idea that potential participants volunteer ("opt-in") to join without being recruited using traditional probability-based methods. In practice, this means that potential panelists either come across the panel's website or are directed to it by a social connection or advertising. These potential panelists then typically fill out an entrance survey, which collects information such as demographic characteristics. Following this, they may be given access to a website where they can select (or are assigned) surveys or other research tasks in which to participate. On some platforms, panelists may receive email notifications alerting them to new surveys or tasks.

Non-probability panels use both indirect and direct methods of bringing individuals into the panel. Indirect methods rely on potential panelists to come on their own to a vendor's recruiting portal. This could be because of word-of-mouth, general advertisements, or happenstance. Indirect methods do not focus on any particular demographic or subgroup, and can be thought of as metaphorically equivalent to "casting a wide net." In contrast, some panel providers also use direct methods to target specific population subgroups, typically to increase the representation of under-represented groups. Panel companies accomplish this by targeting people who visit particular websites and guiding them to the panel website to join the panel — a process that is sometimes called "river sampling." For instance, if a panel vendor was interested in increasing the number of people interested in hunting and fishing in their panel, they would place advertising and links to their panel site on outdoor recreation sites.

Because non-probability panels, by their nature, are less restrictive in their membership than probability panels, they are usually larger. This large panel size may be an attractive feature for some researchers when planning pre-tests, as they may be able to identify small and uncommon subgroups that are difficult to recruit otherwise. However, because non-probability panel providers do not know the exact dimensions of the universe from which they are pulling

[3] Alternatively, researchers may choose to find their own online participants, not from panels, through approaches such as word-of-mouth or paid advertising; considerations for this approach are discussed in Appendix: Other sources of online participants.

panelists, they cannot assign statistical likelihoods of selection into either the overall panel or any individual survey or task. Although researchers could use quota sampling to achieve a sample with characteristics that appear similar to their target population, the lack of information about probabilities of selection means that sampling weights cannot be assigned and the panel cannot be considered "statistically representative." Data collected from these panels should not be used for prevalence or point estimates.

2.1.2 Probability panels

As opposed to non-probability panels, probability panels, such as the Gallup Panel, KnowledgePanel, and AmeriSpeak, directly recruit all of their panelists from a known frame. Most use an address-based sample (ABS) frame (e.g., the US Postal Service's Delivery Sequence File), while some use dual telephone frames or mix address and phone frames. Because the composition of the panel is based on a statistical sampling design, each panelist in a recruited panel can be assigned a probability of selection, which allows the panel companies to assign weights to individuals and claim that their panels are "statistically representative."

Individuals in these panels participate in studies through invitation only; unlike panelists in some non-probability panels, panelists in probability panels cannot choose whether they are sampled for (i.e., able to participate in) an individual study. These managed processes may have the additional benefits of reducing fraud, such as by preventing individuals from signing up for multiple accounts, and fatigue, by limiting the number of study invitations each panelist receives typically to no more than a few invitations per month.

2.2 Sampling participants from online panels

Once a researcher chooses an online panel, there are two common methods for recruiting samples from online panels: managed recruitment, whereby the researcher specifies sample characteristics and the panel provider builds the participants sample, and marketplace "job" postings, whereby the researcher posts a study opportunity and waits for participants to volunteer and join the sample. A researcher's choice of panel type dictates how participants can be sampled. For example, if a researcher chooses to source their participants from a probability panel, then the panel provider will likely require the researcher also purchase managed recruitment services, In contrast, if a researcher chooses to source their participants from a non-probability panel, then the researcher may have the option to manage recruitment by him or herself.

2.2.1 Hiring a vendor to manage sampling

When researchers hire a vendor (e.g., SurveyMonkey, NORC, GfK, Qualtrics) to oversee the entire sampling process, they typically develop sample specifications and require the vendor to recruit a sample to meet those specifications. For an extra fee, you can purchase access to a sample of participants to match a desired set of characteristics. Most basic demographic information is collected at the time that the participant joined the panel but niche screening criteria will likely need to be asked at the time of data collection. In general, the more criteria you require and the harder it is to reach your target population, the more expensive this form of recruitment becomes.

After you specify your project requirements, the vendor contacts participants in the panel asking them to participate in a study. The vendor may first design the sample and then reach out to only those sampled panelists. The vendor may advertise a generic offering and then direct interested individuals to your study, up to your maximum number of desired participants. The vendor generally will continue to invite panel participants until your sample specifications are fulfilled. Once each eligibility criterion (e.g., demographic quota) is filled, the vendor will screen out ineligible panel members at no further cost to you.

2.2.2 Managing your own sample

When researchers instead manage the sampling process themselves, they typically post research opportunities ("jobs") on marketplaces where individuals browse or search for online tasks (e.g., Amazon Mechanical Turk, MicroWorkers, RapidWorkers). Potential participants browse or search through postings in the marketplace at their own convenience; researchers do not contact potential participants directly. Because you are posting to an online marketplace, there is no known frame or universe for understanding who might be seeing your posting. The characteristics of your sample may vary based on the day and time you post (e.g., some people may log in to the marketplace only on the weekends), the keywords you use to describe your post (some people may use filters to control what postings they see), and other factors, including ones that are out of the researcher's control, such as what other postings are available at the same time (those postings may flood the marketplace and push your posting out of view).

Overall, managing sampling on your own can be simple to implement and provide researchers with the most control over the fielding process. Because the researcher controls sampling, the researcher can conduct pilot studies, review data as it comes in, and revise the study as needed (within the bounds of your clearance from the Office of Management and Budget under the Paperwork Reduction Act[4] [PRA]). It is also typically the quickest way to sample participants; a study can be completed in a matter of hours. It is also the least expensive option and likely does not require a contract.

3. Who are online participants?

The composition of online samples is not necessarily representative of the U.S. population as a whole. Although online samples may not be representative of a particular population, studies using online non-probability panels have shown similar results to panels that were population-based (e.g., Mullinix et al., 2015). Understanding who participates in online panels will be useful for evaluating how online panels can be used in your research.

3.1 Characteristics of online participants

A number of research studies have investigated ways in which online survey participants may differ from the general population. People who participate in online research differ in demographics compared to the general population: online research participants tend to be younger, more highly educated, White, and more politically liberal than the general population (Fahimi et al., 2018; Ross et al., 2010). Online participants may also have fewer privacy

[4] 44 U.S.C. §§ 3501-3521.

concerns while using the Internet, a greater willingness to express opinions, more technological experience, participate in more activity politically and in their communities, and have more interest in the survey topic (AAPOR, 2010; Ross et al., 2010). Other studies show that participants in some panels differ in personality traits and are more introverted, higher in neuroticism, more open to new experience, less conscientious, and higher in autism spectrum traits and social anxiety (Chandler & Shapiro, 2016). Data quality may also vary between online panel providers (Pew Research Center, 2016), with some recent research showing more effortful and higher quality responses on particular panels over others (e.g., Anson, 2018).

3.2 "Professional respondents"

Another concern about online testing is so-called "professional respondents" or "professional survey takers." This typically refers to survey-takers who participate frequently in many different online panels to take advantage of the incentives offered (e.g., Hillygus et al., 2014). Researchers have been concerned that these professional respondents may decrease data quality; for instance, randomly clicking through surveys to complete them quickly and maximize payments. Research on the impact of professional respondents on sample composition and data quality is mixed, showing that sometimes no differences are found between participants who have taken many surveys and those who haven't (e.g., Matthijsse et al., 2015), whereas others show that professional survey takers speed through surveys and take shortcuts (Toepoel et al., 2008). While the data quality of professional participants is mixed, researchers can take measures to minimize their impact during the data collection and data cleaning phases, such as collecting more sample than needed to allow for possible exclusions, asking respondents to self-report the number of surveys they participate in each year, removing outliers, and removing those who show signs of satisficing.

3.3 Non-response bias

Researchers seeking representative samples must keep in mind that non-response is cumulative through the panel creation and study sampling process. For example, a panel provider may only get a 10% response rate to their original panel recruitment technique and so by the time panelists are sampled into any given online task or survey, the final response rate could be as low as 1% or 2%. If representativeness is important for a particular study, researchers working with probability panels should carefully evaluate the non-response bias.

4. How are online data collected?

Another component of online testing methodology is how to collect the data. Considerations in the data collection process begin with understanding what kind of data are being collected, which may dictate what platforms can be used, and include also managing data security and paying participants.

4.1 Types of data collection platforms

To field studies online and collect data from participants, researchers must use online software services, which we refer to as "platforms". Platforms available to researchers include those

designed primarily to capture survey responses, as well as platforms specializing in the administration of non-survey tasks and collection of multimedia responses or paradata (e.g., audio/video, mouse clicks). The data may range from numeric survey responses, to open-ended text, to audio and screen recordings. For example, surveys with open- or closed-ended questions designed to probe question comprehension may be administered to a representative sample of participants using a web survey or a convenience sample of online participants may be audio-recorded as they respond to questions using a think-aloud procedure. However, "online testing" does not refer to the use of web technology to conduct interviewer-administered studies (for example, the use of a web conferencing tool to conduct cognitive interviews with participants in multiple locations is excluded).

4.1.1 Survey platforms

Platforms primarily capturing survey responses (e.g., Qualtrics, SurveyMonkey) allow researchers to build an instrument with custom text and survey questions. These kinds of platforms are commonly used for online testing, as they are flexible and well-suited for a wide variety of tasks, such as A/B testing and web probing (i.e., online cognitive interviewing). Researchers should note that the features offered by survey platforms vary, and features that researchers are accustomed to including in production surveys may be difficult to implement in off-the-shelf solutions. Survey platforms are also typically fairly limited in terms of the paradata they can provide: some do not provide any at all, while others provide only limited paradata such as device, operating system, and time spent on screen.

4.1.2 Other platforms

Other platforms designed primarily to collect multimedia responses and paradata allow researchers to conduct more specialized types of research, such as usability or qualitative testing, card sorts, and learning tasks (e.g., Loop11, TryMyUI, GroupSolver, TreeJack). These platforms typically have additional capabilities that survey platforms do not. As with survey platforms, the options vary widely by individual vendor. Some may collect multimedia response data, such as participant audio/video, or advanced paradata, such as a complete history of mouse clicks. Some platforms collect participant reactions to a live website embedded within a research study, instead of needing to use screenshots or an external link. While well-suited for their intended purpose, they may be less adaptable for other types of research. These platforms may not have the necessary capabilities to host a survey or may only be capable of hosting a very simplistic survey. It is sometimes possible to identify a workaround by combining multiple platforms (e.g., asking TryMyUI participants to navigate their browser to a SurveyMonkey survey and collecting their reactions to it within the TryMyUI platform). These types of platforms may be more expensive than survey platforms.

4.2 Data security

The data collected through the online platforms discussed in this document typically are stored on third-party servers operated by the data collection platform vendor. While these servers may be protected by firewalls and other security systems, they are not generally considered by the federal government to be sufficiently secure for government collection or storage of personally-identifiable information (PII). Other kinds of data, such as non-PII survey and demographic

responses, behavioral paradata, and metadata, are typically allowed to be collected but policies vary by agency.

Alternatively, researchers may be able to use in-house data collection platforms, such as the platform used by their agency for production data collection, or develop protocols with vendors that comply with the Federal Risk and Authorization Management Program (FedRAMP).

4.3 Payments to participants

Rates of payments vary by platform and study. Within the federal government environment, researchers may aim to make payments to participants match as close as possible to industry standards so as to be effective in attracting participants. However, researchers should maintain an approach of treating incentive payments as tokens of appreciation for participation in the research, rather than as a wage.

5. Deciding whether to conduct online testing

Researchers may choose to address their research questions using any number of methods and techniques, depending on the nature of the question they are asking. For example, researchers seeking to establish which question about a behavior or attitude is best may wish to compare them using a quantitative, split-ballot test. Another researcher, seeking to understand what a respondent is thinking of when presented with a certain word or phrase, may wish to employ a qualitative strategy, such as cognitive interviews. Both qualitative and quantitative methods have value in survey pre-testing and methodological research.

Online testing is flexible enough that researchers can use it for both qualitative and quantitative projects. They may be used independently of—or as a complement to—traditional methods, such as cognitive interviews or in-person usability tests. When approaching the decision of whether to conduct online testing, researchers should understand that there are some research projects for which online testing methods are appropriate and there are other projects for which other modes of pre-testing such as in-person interviews would be a better fit. Researchers need to assess the nature of their research questions, the need for statistical validity, and the overall fitness-for-use of online testing methods for each study.

5.1 Logistical benefits

Traditional pre-testing methods employed by federal agencies typically require a lengthy timeframe for recruitment and data collection, which results in a relatively high cost. In a lab-based cognitive interview project, for example, staff must recruit the necessary number of individuals to participate by posting advertisements (e.g., in physical locations, on Craigslist, on Facebook, or a similar venue) or calling potential participants. In contrast, recruitment through online platforms typically requires only the time needed to specify any eligibility criteria or target demographics; participants are then recruited to these specifications without additional effort from the researcher.

In a lab-based cognitive interview project, researchers are involved in the preparation of the necessary materials for each of those interviews, whether they are paper printouts, decks of cards to be sorted, or interview protocols. The number of participants that can be scheduled may be limited due to the size of the lab or the number of interviewers available. Interviewers must make time to meet with each participant, including waiting for late-arriving participants and re-scheduling appointments. In contrast, an online study can be developed one time and then completed by many participants without any further work by the researcher.

If staffing or budget resources are limited, online testing can enable an agency to conduct research that might otherwise have been forgone. Online testing methods can be employed by offices with relatively small numbers of research staff who cannot logistically conduct and analyze large numbers of interviews. The online technology also means that data can be more easily collected from a diverse sample (e.g., geographic dispersion across the United States) or a targeted sample (e.g., hard-to-reach populations) without hefty travel costs.

The potential for quick turnaround times is another advantage: hundreds of participants can feasibly complete a study within a few hours. Many participants can participate in an online study at the same time, at any time of day, including evenings and weekends. Once the infrastructure for conducting online research is set up, the complete data collection process (after approvals and funding have already been secured) can take as little as a few hours. Keep in mind, however, that the full timeline that includes initiating a new contract with a vendor may add several months to the process, though this varies by agency.

5.2 Logistical limitations

For some online platforms, there are steep learning curves for beginners to develop instruments or questionnaires and analyze large datasets. Whereas traditional pre-testing methods may use no technology at all, online recruitment and data collection require learning new systems and software.

Online testing typically results in large datasets, for which statistical analysis is needed before researchers can draw conclusions, compared to smaller samples of in-person interviews where researchers can interpret results and form conclusions immediately.

5.3 Methodological benefits

Online testing methods enable researchers to answer different types of research questions than asked using in-person methods, from running experiments that require collecting data from large samples to collecting paradata about break-offs to collecting large datasets of text from open-ended questions. Small research questions that might not be pursued because they don’t require a full-length, in-person interview can be answered using only a few minutes of participant time.

Online studies can be used to make within-study comparisons between groups and to identify questionnaire or instrument problems during pre-testing. Importantly, the online aspect of the methodology also means that each survey or task is administered in a controlled way, without risk of interviewer bias, because the instrument looks the same from the first participant to the

thousandth participant. Online testing methods can also strengthen the conclusions of an in-person study by adding a different method (e.g., adding a quantitative method to a qualitative project) or participant sample to the project (e.g., reaching a more diverse population than is available to participate in an in-person study).

5.4 Methodological limitations

Mode differences mean that online methods may not be the best choice for testing interviewer-administered questions. Intrinsic factors of interviewer-administered questionnaires, such as interviewer and respondent behaviors, cannot easily be translated to self-administered online surveys. Participants may not realize that the online panel survey is sponsored by a federal agency or they may not treat the survey the same as they would a survey collecting data to directly produce official statistics. Longer studies that require significant participant time and attention may suffer high break-off rates or poor data quality.

The fixed nature of the method, whereby the questions and the instrument are administered the same way to all participants, also means that online methods may not be appropriate for exploratory studies for which the protocols and questions are still being refined and follow-up probes are not easily administered. Some researchers conduct preliminary in-person exploratory research to refine follow-up probes before conducting larger-scale online testing.

Given the remote nature of the participation, fraudulent responses are possible. Fraudulent responses may appear in such forms as an individual who randomly clicks through your study or a bot account (a computer program impersonating a person) that was created using false information and can quickly complete many surveys with random responses. For all online data collections, researchers must take extra steps to examine their datasets for such responses. Researchers may choose to exclude these responses from their final dataset, ultimately increasing the cost of data collection.

Another common concern is the representativeness of online samples. It is difficult to make blanket statements about the representativeness of online samples because sampling methods vary by platform and by study; however, online samples typically consist of younger participants with lower incomes, higher educations, and lower levels of health.[5] If using a non-probability panel, researchers may have no frame information. In general, online testing methods are not appropriate for studies targeting segments of the population who are not commonly found online (e.g., older adults) or when the online testing method itself is directly related to the concept being measured (e.g., understanding internet use behaviors). Despite these particular biases, we note that other testing methods, such as in-person interviews that rely on members of a local community willing to participate in research, arguably also suffer from representativeness concerns.

[5] AAPOR Report on Online Panels (2010).
https://www.aapor.org/aapor_main/media/mainsitefiles/aaporonlinepanelstfreportfinalrevised1.pdf

6. Choosing recruitment and data collection methods

Because the choices of recruitment and data collection platforms are so often linked together, we recommend taking both sets of needs into consideration simultaneously. Although the methodological needs of your research project may be the most important factors in data quality, your budget and time resources are necessary considerations early on in the project, too. Researchers may find that they must make compromises due to resource limitations. Typically, you only need to do this step of choosing platforms one time or once every few years. Unless your needs vary significantly between projects, you will likely use the same participant recruitment and data collection methods for all of your studies.

6.1 Recruitment needs

To plan your recruitment strategy, begin by considering your sample needs. For example:

- Does your project need a representative sample?
- Does your project target a specific subpopulation? Is it difficult to reach that group online?
- Will you be likely be able to find the data needed to identify your participant groups based on a panel's frame, or will you need to build a screening step into your survey?
- Will you manage sampling on your own or hire a contractor or the panel vendor to do the work?
- Does your project need to link participants across waves or surveys?

6.1.1 Probability or non-probability panels

Some research projects, such as making inferences to your target survey population about respondent behaviors, require a probability panel. For these projects, researchers should inquire with the candidate panel providers about how sampling weights are calculated when deciding which panel to use. Panel providers typically consider this proprietary information but this may be an important consideration in choosing between panels and affect how useful the data collected from the panel are. This issue is not applicable to non-probability panels.

For other research projects, such as pre-testing question wording changes, non-probability panels are sufficient. Even if you believe that your research questions do not require a representative population, we still recommend consulting stakeholders to find out whether research with non-representative samples will be accepted and trusted. Some non-probability panels may be extensively researched and you can find publications detailing population characteristics in academic journals. Vendors with probability panels, however, do not generally make this information publicly available but you may be able to find some information, such as how individuals are recruited into the panel and whether there are any selection biases, on their website.

When choosing a panel, researchers also need to consider whether the target population can be found in sufficient numbers on the panel. If your project requires participants who are trained to provide a specific type of data (e.g., think-alouds), then your search may be limited to only a few platforms where individuals have been trained to provide that type of data.

6.1.2 Sampling

Your choice of probability or non-probability panel may dictate your sampling management methodology. If choosing to sample participants from a probability panel, you must hire the panel vendor to manage recruitment. If choosing to sample participants from a non-probability panel, you can elect to manage sampling on your own or hire a contractor or the panel vendor to do the work.

In the most straightforward case of convenience sampling, you can make your study available to everyone on your chosen platform; otherwise, all data collection platforms will allow you to conduct screening of your sample as part of your questionnaire and these responses can be used to screen out ineligible participants. If the project's sampling needs are simple, such as a few characteristics for quota-sampling, then sampling needs do not need to be considered as a part of platform choice. If the project uses a complex sampling design, it may be impractical to manage the sampling on your own on some platforms and researchers should consider hiring a contractor to manage the process or paying the panel vendor to select the sample. Talk with the panel vendor to see what frame information is available that can be used for screening, such as demographics, number of previous studies completed, participation in previous studies that you have conducted, or evaluations from other researchers.

If targeting individuals under 18 years of age, check that the vendor is familiar with the Children's Online Privacy Protection Act (COPPA) and has methods of finding participants for you, such as through asking permission of panelists who are parents or guardians to survey their children.

6.2 Data collection needs

In the online environment, participant recruitment and study fielding happen concurrently: the study must be available from the moment you begin to recruit because the recruitment advertisement or email includes a link to the study. To plan your data collection strategy, begin by considering your data needs. For example:

- What kind of data are being collected?
- Are you or your staff able to use online software to develop an instrument and records all of the data, paradata, and metadata needed?
- Are there any legal requirements governing where data can be stored?

Identification of the most suitable platforms to use depends on the type of data you intend to collect, whether this requires particular software features or a specialized testing platform, and how detailed the data needs to be (e.g., basic paradata such as break-offs or detailed paradata such as where on the screen the participant clicks). If the primary goal of the research is to collect survey responses, a survey platform should be used. If the primary goal of the research is to collect non-survey responses, or if survey paradata is more important for the research than survey responses, explore other types of platforms to see if an existing alternative caters more specifically to your research needs. If no such platform exists or you are unable to use it, adapt your research to fit a survey platform or consider whether another methodology is more appropriate.

Some platforms offer more features and customization than others, which may immediately eliminate some platforms as candidates for running your study. Although many data collection platforms rely on "drag and drop" design interfaces that are simple to implement, advanced customization may require greater programming skill. Go to the platforms' websites to read their features lists and sign up for free trial accounts to explore using the software. If you determine that your agency does not have the appropriate staff to program your study, you can hire vendors to program the questionnaire as an additional service. Here are several examples of features, paradata, and metadata[6] that researchers may be interested in but are not offered on every platform:

- Custom question and response formatting, such as adding unit labels to the right of a text entry box
- Ability to loop through a series of questions repeatedly without programming each individual instance of a question
- Randomization of questions, question blocks, and response options
- Hard and soft edit checks
- Fills (e.g., fill a reference to a previously used name or pronoun)
- Questions or tables that can simultaneously capture multiple data types
- Time spent on a page
- Counts of how many times the participant clicks on a page
- Variable creation to calculate or generate data, such as random numbers
- Screen recordings of a participant's screen as they complete the study
- Voice recordings of participant think-alouds
- Overall number of ineligibles
- Locations of break-offs
- Type of device used
- Javascript for customizing the appearance of the survey or the data that are recorded
- Mobile optimization

In addition to these data needs, researchers also must consider the logistical aspects of data collection. Online testing platforms typically charge researchers a price per each individual who participates in a study (this charge is for the use of the online testing platform and is separate from any recruitment fees or participant compensation). However, there may be some variability in how participation is defined: for example, you may be charged for partial completes, screen-outs, and fraudulent responses[7], unless you specify otherwise in advance. For projects with strict screening criteria, this could raise costs substantially. Researchers should understand clearly what they are paying the vendor for.

[6] Some platforms may also provide personally-identifiable information (PII) such as IP address or latitude and longitude location, which researchers should not download or save.

[7] Some platforms allow researchers to rate participants or decline study completions for reasons such as fraudulent participant behavior or poor quality responses. Even though platforms offer this option, researchers should consider whether their agency's informed consent policies allow for behaviors such as declining to answer questions for any reason without penalty as well as the "optics" of their agency appearing to reject responses.

6.3 Resource and budget considerations

Every research project is limited by its budget and other resources. Although you may need to revisit these specifications as your recruitment and data collection strategies evolve, we recommend beginning with a set of basic specifications to direct the planning process. Most researchers find there is a trade-off between staff time and budget, whereby options that cost less money require more researcher time and effort.

Consider the following:

- How much funding can be committed to this project?
- Are there any timelines limiting when that money can be spent or what that money can be spent on?
- When are results needed?
- Do the project staff have the time and expertise to program the instrument?
- Do the project staff have the time and expertise to field the study?

Although fielding an online study requires substantially less time than in-person interviewing, there can still be some demands, such as piloting, responding to e-mail questions from participants, or releasing sample in batches or at timed intervals. Some platforms require a significant upfront investment of time to learn how to use the interface and what practices are considered normal (e.g., payment rate) within the context of a platform. Hiring a contractor or purchasing access to a managed panel typically includes assistance on these issues but still requires staff time to make the ultimate decisions. The alternative "do it yourself" platforms do not offer assistance beyond how-to guides, though other platform users may offer their own guidance. Colleagues at other agencies are also a great resource.

Hiring a vendor to handle recruitment or using a probability panel can drive up the cost of online research. On one end of the range, do-it-yourself options, where you manage the recruitment and data collection process yourself, can cost less than $5 per complete for a convenience sample, including payments to participants, recruitment commission fees, and data collection platform fees (not including researcher time). Hiring a vendor to manage recruitment from a non-probability panel may double or triple that cost. On the other end of the range, contracting with others to fully manage the process of recruiting from probability panels and collecting their data can cost closer to $100 per complete.

Online data collection is typically faster to complete than in-lab interviews, even with the larger sample sizes of most online studies. However, the timeline for initiating contracts and working with panel vendors is typically significantly longer than opting to manage the process yourself and may require more lead time (if not more staff time).

7. Recruiting a sample

For some research projects, drawing a convenience sample is sufficient; for others, such as projects that require matching to certain demographics or identifying a target subgroup, additional steps must be taken before a sample can be recruited.

7.1 Notifying potential participants

The recruitment platforms vary significantly in how potential participants learn about study opportunities: some platforms send targeted emails and conduct non-response follow-up while other platforms function more like public notice boards and only potential participants who happen to see your posting can participate. This difference may be of particular importance for researchers interested in response rates.

Despite their typically large sizes, online testing samples are often recruited quickly and through limited channels such as a single notification to panelists. This may translate to only a small range of panelists, such as those available on a Tuesday morning during workday hours, seeing your invitation before your data collection is complete. Depending on your research question and your sample size, it may be desirable to increase sample diversity. If hiring a vendor to manage your recruitment, request that they explain how they will do this as part of the procurement process. If doing your own recruitment, consider staggering sample releases to include weekdays, weekends, daytime and evening (Casey, et al., 2017). On some platforms, including a range of keywords in your study description may also help to ensure a wide range of potential participants learn about your study.

7.2 Finding eligible participants

Online platforms offer a range of ways to find participants, from simple fee-based options to more complex options to be done on your own. When considering recruitment methods, researchers should know that some approaches, such as screening questions, may need to be implemented through the survey testing platform rather than through the recruitment platform.

Regardless of whether you choose to implement no screening methods or one of the approaches below, be sure to collect demographic information, including age, sex, race, ethnicity, education, and any other information relevant to your research question that is useful for understanding your sample characteristics. These data may be available from the panel provider as frame information or you may collect it yourself as part of your survey.

7.2.1 Purchase screening criteria

Some recruitment platforms offer to screen participants based on frame data. You can request to target participant demographics such as location (such as a particular state of the United States or the United States in general), age or gender; or other characteristics (such as experienced hunters or recent college graduates). If you want your sample to match a selection of census demographics, you may be able to request a quota sample to target those participants. Researchers may also be able to select participants based on participant study history. Use of these data is typically based on a per-participant fee, with more niche specifications costing more money. Researchers should be cautious that some of these characteristics may not be stable over time (e.g., employment status, age of children in household) and a sample drawn based on these qualifications may not match your specifications if it relies on outdated information. Researchers should inquire as to how the platform acquired the data (including the age of the information), to ensure it is reliable and valid for your needs.

7.2.2 Write screening questions

As a supplement or alternative to purchasing screening criteria, researchers can use screener questions at the beginning of a study tailored to their specific research questions. Responses to the screeners can be used to qualify or disqualify individuals from participating in the study (e.g., using skip logic to end the study for any individual who reports not having health insurance in the last 12 months). Take care to write these questions so that it is not obvious which answer is the "desirable" option for gaining eligibility for your study. Most platforms automatically prevent individuals from participating in a study more than one time by using platform identifiers or IP addresses.

7.2.3 Quota sample

Researchers can specify groups of interest and desired sample sizes for each group. If your panel provider has these variables as part of their frame data, this step can be done by the panel provider in advance of your study; however, researchers not using managed recruitment methods will likely need to implement this approach at the time of the study's fielding through built-in features of the testing platform. After the quota for a group has been reached, any subsequent participants who fit that group will be routed immediately to the end of the study. Researchers should note that the latter option of quota sampling during the fielding period must be paired with slower-paced fielding (i.e., requesting only a fraction of the sample at a time) to avoid unintended oversampling when participants screen in before earlier participants complete the study and get counted toward the quota total.

8. Fielding the study

Researchers fielding online pre-testing studies face a different set of considerations than they do with in-person pre-testing. The differences in payment infrastructure, data collection, and mode of response need to be accommodated.

8.1 Payments to participants

Payments to online participants vary by platform and by study. For managed recruitment panels, the payment rate is typically fixed by the vendor. Researchers should consider a number of factors in determining the appropriate payment amount, including but not limited to:

- What is a typical payment for the platform for tasks that are similar to yours? Some platforms have fixed rates while others have no set prices but the population has expectations as to how much a study should offer. Because your study is just one of many opportunities being offered, researchers may need to align their payment amounts with platform norms.
- Is your target population hard to reach?
- How much time does the study take? On most platforms, payment is positively correlated with time; some participants even translate payment into effective hourly rates to evaluate whether to participate in a study.
- How much effort does the study take? Online participants may expect higher payments for tasks that involve open-ended writing or other higher levels of effort.

- Very low payment amounts may result in fewer participants willing to complete your study and may affect your (and your agency's) "reputation" on the platform. Participants learn through their own experience completing your studies or by reading information shared by other participants about their experiences.
- Studies have shown that high or low payment rates, within a range considered typical for the platform, do not affect data quality (e.g., Buhrmester, Kwang, & Gosling, 2011).

8.2 Informed consent regarding data security

Federal researchers must inform participants of the possible information security risks associated with online studies and they are not allowed to collect sensitive or personally-identifiable information (PII) without explicit permission from their agency. Some PII may be collected by the platform, such as contact information and details for payment, but the researcher may not be allowed to request or save those data. A confidentiality notice to participants should appear on the first page of your study and, if applicable, include a statement to the effect that continuing with the study is an acknowledgment of these concerns.

8.3 Designing for online self-administration

Online studies typically take 15 minutes or fewer and use simple instructions. If your questions were originally written for administration in a different mode, you may need to revise them. Given that online testing is self-administered and that data are collected quickly, researchers should ensure that instructions are clearly written and simple to follow and understand. If not, participants may not complete the study the way that you intended and you may not realize the problem until after many participants have already finished.

Relatedly, participants are increasingly using smartphones to complete online studies (Brosnan et al., 2017). If researchers try to guide participants to complete the survey on desktop by instructing participants not to complete the survey on mobile, there is often non-compliance. When designing your survey, check that your question formats are optimized for mobile as well as desktop screens.

Pilot testing your study with a small sample (e.g., 20 participants) may help you to reduce errors that may cost time and money if there are questionnaire typos, skip logic errors, or other instrument problems. Piloting also enables you to revise the study after preliminary analyses, within the bounds of your clearance from OMB under the PRA. Including an open-ended comment box at the end of your study for general feedback allows participants to inform you of any glitches they encountered or problems they had when answering your questions.

Some researchers add "attention checks", "instructional manipulation checks", or "trap questions" that serve as a flag for whether a participant is responding thoughtfully to your online study, but the literature is mixed on their effectiveness and some studies have shown that these harm data quality (e.g., Anduiza & Galais, 2017). Researchers should be cautious if considering including such questions.

Some platforms allow researchers to include captchas, or a type of challenge-response test that can be helpful in discriminating between human participants and bots. Given the low cost of including a captcha, researchers may consider including one at the start of the study as one of multiple tools to improve the quality of your data.

Some platforms allow researchers to rate the responses submitted by participants or even "reject" responses and not pay the participants. However, federal researchers typically do not reject responses due to ethical considerations and the "optics" of their agency appearing to reject responses. We recommend paying all participants who complete your task and subsequently considering excluding any low quality responses that you might otherwise have rejected from analysis.

Including benchmarking questions may be useful in weighting or otherwise evaluating the extent of bias in your sample. We recommend copying a few questions from your target survey and collecting data on characteristics of interest (e.g., demographics or household size) to serve as benchmarks, if appropriate for your study.

8.4 Instrument development

Some data collection platforms offer built-in features such as randomization of response options or question order, embedding of variables within the question displayed to participants, and times spent on pages. Researchers should check that any necessary information such as display order or embedded values are recorded in their final dataset. If this information is also necessary for partial completes, then check that this information is recorded early in the study rather than only at the end.

9. Analyzing the Data

Most data analysis considerations for online testing are the same as for other modes and types of data collection; however, given the size of most online datasets, the processes of cleaning and coding data may require more time. Online data collection platforms typically allow researchers to view data in real-time or at specified intervals during the fielding period. Data typically are downloadable in a spreadsheet, with labels taken directly from the instrument used to collect the data (e.g., question numbers or question text).

9.1 Data cleaning

Given the remote nature of online testing, it may not always be obvious to researchers whether a participant's data is of high quality. The following are examples of online participants that may warrant exclusion from your final dataset:

- Unambiguously inattentive participants, such as those who enter the wrong type of data into a field (e.g., a word instead of a number).
- Participants who enter gibberish or clearly irrelevant text into open-ended text fields.
- Participants with very long task durations.
- Participants with very short task durations.

- Participants with short page times on pages with instructions.
- Participants with seemingly duplicate or identical responses.
- Participants who leave multiple items unanswered or respond with "don't know" or "not applicable" responses.
- Participants who exhibit straight-lining or patterns of responses.
- Participants who explain or otherwise indicate that they misunderstood the instructions.

Although these factors may be data quality flags, research on data collected from online panelists has found that removing participants based on factors like speeding may not change findings such as distributions and correlations (Thomas, 2014); researchers should be conservative when excluding data.

9.2 Documentation

Alongside substantive analyses, researchers should include an explanation of the methods used to collect the data and, depending on the audience, a discussion of how your study's methods or sample deviate from the target survey design or population. This may include a description of the steps taken to find eligible participants, including any screener questions used, a comparison of your study's sample characteristics to the target survey's sample, and any steps taken to clean the data including reasons for any exclusions.

10. Federal and agency-specific policies governing data collection

Online data collections, similar to other pre-testing and research projects, are subject to both federal and agency-specific regulations.

10.1 Federal policies

OMB has issued regulations and guidance to promote agency compliance with the PRA (see https://www.whitehouse.gov/omb/information-regulatory-affairs/federal-collection-information/#PRAC). The OMB's memoranda "*Information Collection under the Paperwork Reduction Act"*[8] is a primer in the PRA and OMB's "*Facilitating Scientific Research by Streamlining the Paperwork Reduction Act Process"*[9] is particularly useful in navigating the PRA in a research context. OMB states that:

> [t]he PRA was designed, among other things, to "ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the Federal Government" and to "improve the quality and use of Federal information to strengthen decisionmaking, accountability, and

[8] Office of Management and Budget (2010). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/inforeg/PRAPrimer_04072010.pdf

[9] Office of Management and Budget (2011). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2011/m11-07.pdf

> openness in Government and society."[10] Federal agencies play a critical role in collecting and managing information in order to promote openness, increase program efficiency and effectiveness, reduce burdens on the public, and improve the integrity, quality, and utility of information to all users within and outside the government.[11]

Our experience is that OMB applies the same standards for online data collections as those for other offline pre-testing data collections conducted by your agency, with particular emphasis on its *Standards and Guidelines for Statistical Surveys*.[12]

10.2 Agency-specific policies

When you recruit participants for an online study, you are communicating with the public and representing your agency. Your agency may have strict guidelines about what platforms you can use and what content you can post. For example, your agency may require you to go through a lengthy process to be permitted to use a platform or to advertise your study on social media. Also, some agencies may require any project that involves vendors handling either data collected on behalf of the agency or draft survey materials that have not received final approvals to use a non-disclosure agreement.

10.2.1 Selecting a vendor

Researchers should have early discussions with their agency's administrative staff to understand what steps need to be taken to contract with or purchase online testing services from a vendor. Example topics to discuss include: whether a Request for Proposals is necessary; whether a sole source justification is necessary; whether there is an existing relationship with a contractor that the online testing work can be added to, such as an umbrella contract, so that you can avoid the process of starting a new contract; whether online testing is likely to be a one-time project or repeat expense.

If initial discussions are promising, researchers should involve the prospective vendors in this conversation as well, at an early stage. However, there are strict rules about when a researcher can discuss a project with a vendor; for example, you may only be allowed to speak with a vendor after they have responded to a formal Request for Information.

10.2.2 Paying a vendor

For most projects, all of your costs will be billed by the vendors providing access to their participants or software; you will not need to pay individual participants.[13] These may require special authorizations from your agency's budgeting and procurement offices; certain forms of participant payments, such as gift cards, may not be allowed at all. Some platforms offer the ability to pre-purchase credits that can be used at a later time. This flexibility allows researchers to work around budget restrictions and uncertainties; for example, using end of fiscal year funds

---

[10] 44 U.S.C. § 3501.
[11] 44 U.S.C. § 3506(b).
[12] Office of Management and Budget (2006). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/OMB/inforeg/statpolicy/standards_stat_surveys.pdf
[13] An exception is individual payments to anonymous online participants, for example in follow-up studies.

to pre-purchase credits to be used in the next fiscal year before budgets have been finalized. When exploring this option, be sure to check whether the credits come with an expiration date.

In some cases, paying vendors can be as straightforward as a credit card purchase. However, there are many situations where procurement becomes more complicated, and this will vary by agency. Researchers should have early discussions with their agency's administrative staff to understand what steps need to be taken to purchase online testing services. Example topics to discuss include whether the agency is able to accommodate invoicing or similar payment agreements from companies, which may differ from typical contract awards.

10.2.3 Data security and confidentiality

The federal government has a variety of statutes and policies that govern privacy, confidentiality, and security of data.[14] Researchers working in the federal government environment must also ensure that the platform is a vendor approved by the federal government to collect and store data. Your agency's authority to collect data may require strict data confidentiality standards that preclude use of an off-the-shelf third-party platform's online cloud storage for participant data (e.g., data collected under Title 13). These policies do not prevent the collection of data online but do mean that you must ensure that the platform you choose is approved for use at your agency to securely store participant data. Similarly, researchers should work with their agency to establish confidentiality policies regarding who will have access to the participant data.

Some agencies operating under the Confidential Information Protection and Statistical Efficiency Act (CIPSEA)[15] may allow researchers to collect information online as long as participants have been given an appropriate warning about risks as part of informed consent before agreeing to participate. Work with your agency to make sure that the language you use in the pledge you give to participants covers all possible concerns. For example, the Bureau of Labor Statistics uses the following language for OMB No. 1220-0141:

> *This voluntary study is being collected by [Agency] under OMB No. [XXXX] (Expiration Date: [Date]). Without this currently-approved number, we could not conduct this survey. We estimate that it will take on average 15 minutes to complete this survey. Your participation is voluntary, and you have the right to stop at any time. This survey is being administered by [Data Collection Vendor] and resides on a server outside of the [Agency] Domain. [Agency] cannot guarantee the protection of survey responses and advises against the inclusion of*

[14] Office of Management and Budget (accessed June 2019). https://www.whitehouse.gov/omb/information-regulatory-affairs/privacy/,
Office of Management and Budget (2017). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2017/M-18-02%20%28final%29.pdf

[15] Office of Management and Budget (2002). https://www.govinfo.gov/content/pkg/STATUTE-116/pdf/STATUTE-116-Pg2899.pdf
Office of Management and Budget (2007). https://www.govinfo.gov/content/pkg/FR-2007-06-15/pdf/E7-11542.pdf

> *sensitive personal information in any response. By proceeding with this study, you give your consent to participate in this study.*

Other researchers may choose to develop an in-house hybrid solution that uses the software from a third-party platform but stores data on the agency's own servers. Or, your agency may allow you to collect non-personally identifiable information on a third-party platform. Please check with your agency's Senior Agency Official for Privacy.[16]

References

AAPOR Report on Online Panels (2010). https://www.aapor.org/aapor_main/media/mainsitefiles/aaporonlinepanelstfreportfinalrevised1.pdf

AAPOR Report on Nonprobability Sampling (2013). https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Anduiza, E. & Galais, C. (2017). Answering Without Reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research, 29*(3), 497–519.

Anson, I.G. (2018). Taking the Time? Explaining effortful participation among low-cost online survey participants. *Research and Politics, 5*(3), 1-8.

Brosnan, K., Grün, B., & Dolnicar S. (2017). PC, Phone or Tablet?: Use, preference and completion rates for web surveys. *International Journal of Market Research, 59*(1), 35–55.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Chandler, J. & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology, 12*, 53-81.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal Differences Among MTurk Workers: Time-based sample variations and implications for online data collection. SAGE Open.

ESOMR/GRBN Guideline for Online Sample Quality (2015). https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-GRBN_Online-Sample-Quality-Guideline_February-2015.pdf

Fahimi, M., Barlas, F. M., & Thomas, R. K. (2018). A Practical Guide for Surveys Based on Nonprobability Samples. AAPOR Webinar.

Hillygus, D.S., Jackson, N., & Young, M. (2014). Professional Respondents in Non-Probability Online Panels. In Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 219-237). Chichester, UK: Wiley.

Office of Management and Budget (2010). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/inforeg/PRAPrimer_04072010.pdf

Pew Research Center (2016). Evaluating Online Nonprobability Surveys.

Office of Management and Budget (2010). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2011/m11-07.pdf

Office of Management and Budget (2006). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/OMB/inforeg/statpolicy/standards_stat_surveys.pdf

Office of Management and Budget (accessed June 2019). https://www.whitehouse.gov/omb/information-regulatory-affairs/privacy/,

Office of Management and Budget (2017). https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2017/M-18-02%20%28final%29.pdf

Office of Management and Budget (2002). https://www.govinfo.gov/content/pkg/STATUTE-116/pdf/STATUTE-116-Pg2899.pdf

Office of Management and Budget (2007). https://www.govinfo.gov/content/pkg/FR-2007-06-15/pdf/E7-11542.pdf

Matthijsse, S. M., de Leeuw, E. D., & Hox, J. (2015). Internet Panels, Professional Respondents, and Data Quality. *Methodology, 11*(3), 81-88.

[16] https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A108/omb_circular_a-108.pdf

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science, 2*(2), 109-138.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who Are the Crowdworkers?: Shifting demographics in Mechanical Turk. *Conference on Human Factors in Computing Systems – Proceedings*, 2863-2872.

Thomas, R. K., (2014). Fast and Furious… or Much Ado About Nothing? *Journal of Advertising Research, 54*(1), 17-31.

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of Design in Web Surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly, 72*(5), 985–1007.

Appendix: Case Studies

Case Study 1: Using the Amazon Mechanical Turk and Qualtrics online panels

Background

The Bureau of Labor Statistics (BLS) has been conducting research on why sleep duration in the American Time Use Survey, which collects data on sleep indirectly using an activity log of the past 24 hours, produces higher sleep duration estimates than other national surveys that use direct (or stylized) questions, such as "How many hours of sleep do you get?"

Previously, we had conducted cognitive interviews with participants in our lab, asking them about each stage of the response process in answering both types of sleep questions. In these interviews, we observed the same pattern as in the literature where participants reported getting more sleep in the diary vs. stylized measure. We also noticed an interesting finding where comprehension of the term 'sleep' affected estimates. Participants with broader definitions of sleep, including things like resting with their eyes closed, reading in bed, or trying to fall asleep, reported getting more sleep overall.

Description of Online Testing Research

BLS was interested in using online testing to confirm in-lab qualitative findings from cognitive interviews. In this case, we had conducted 29 cognitive interviews with participants in our lab. However, the participant pool was limited to those living in the Washington DC region, who were willing to participate in an interview during normal business hours for $40. Although there was a lot of value in conducting those cognitive interviews, they were limited to a very narrow pool of eligible participants. Thus, we wanted to use online testing to supplement those results and assess whether the same patterns of findings would hold up with a more geographically diverse sample. We also wanted a larger sample size to statistically test hypotheses generated from the cognitive interview results.

Research Questions

- Would we observe the same pattern of participants reporting more sleep with a diary vs. stylized measure in an online sample?
- Would randomly assigning participants to read a specific definition of sleep asking them to exclude non-sleep activities (vs. no definition) from their estimates affect reported sleep duration estimates?
- Would order of the sleep measures (diary first or stylized first) affect responses?

Selecting a Recruitment Platform

We required a large sample size because the study used a complex 3x2x2 design. Participants were randomly assigned to one of three framing conditions (i.e., being told the study was about health, employment, or time use), two definition conditions (definitions or no definitions provided), and two question order groups (stylized first or diary questions first).

In addition, we expected a relatively small effect size since study manipulations are subtle for online studies of this nature. For example, respondents may feel more anonymous while completing an online study than they would completing an interviewer-administered survey. Thus, manipulations such as the study framing may have a small effect since respondents tend to feel more comfortable providing information anonymously online without the presence of an

interviewer, and any sensitivity or social desirability may be minimized. A large sample size would also take into account break-offs, incomplete data, and participants who do not carefully read or follow the task instructions (which typically ranges between 4-10% of the sample in our experience).

Because this study was more interested in internal validity (the measurement of sleep duration across different types of questions) rather than the representativeness of any one population (e.g., obtaining a representative sample of the U.S. population), drawing on a general convenience sample was sufficient for this study.

Given these considerations, we decided to use Amazon Mechanical Turk (MTurk), a large non-probability panel that offers a "marketplace" where individuals can volunteer to participate in short tasks and be paid small amounts of money. This platform allowed for quick and relatively low-cost data collection from a large sample. Because we controlled the recruitment and data collection process on mTurk, we did not need to specify our final sample size in advance but could set our own criteria as needed during data collection.

And to evaluate our research questions using census quotas, we chose to also use the Qualtrics non-probability panel, which allowed us to select on three census demographics: gender, age, and education. The sample was designed to try to reflect census demographics in the U.S. Qualtrics used their frame information to send e-mail invitations to participants who met our target criteria. Once a quota was filled, recruitment stopped for that demographic variable (done through screener questions at the beginning of the study). If the quota was filled, the instrument closed and that respondent was not invited to participate in the study. If the quota had not yet been filled, respondents were routed to take the study. This was done until each quota had been filled. The process took a few weeks in total from the time we gave Qualtrics the finished instrument to receiving the data from them.

We specified the following conditions for Qualtrics, where our project's contract specified we would receive 900 responses (or "good completes") to the study:

- If someone drops out of the study (even if they finish all questions but do not make it to the end), they will not be counted in the good completes.
- If someone makes it through the entire study but skips some questions, they will still be counted as a good complete.
- Exclude those on mobile devices/tablets
- Exclude those outside of the U.S.
- Screen out those who took only a few minutes or over an hour to complete the study

Selecting a Data Collection Platform

Because of the complex skip patterns, random assignment to conditions, logical fills, the online time diary instrument that had to be constructed, and dynamic question wording needed for this research, we determined that the Qualtrics survey design platform (which can be used independently of the Qualtrics non-probability panel) was the best data collection platform. The questionnaire programming and instrument development needed for the study could be accomplished with this platform without advanced programming knowledge. Participants

recruited from both MTurk and Qualtrics panels were directed to complete the study on the Qualtrics data collection platform.

Pre-testing Phase

A primary difference between the in-lab and online testing protocols was the presence of an interviewer. In the cognitive interview lab study, an interviewer walked the participants through the time diary – a complex task where respondents must detail every activity they did from 4am the previous day until 4am of the day of the study. For the online study, the researchers had to adapt the time diary into one that could be self-administered online. To ensure the time diary task made sense to participants without an interviewer present to help them through the diary, we conducted a pretest of just the time diary portion of the online study to ensure it worked as expected. We recruited 6 participants (based on funding and the premise that most problems would likely be uncovered with that sample size) to complete the time diary while thinking aloud and had them respond to follow-up, scripted probes.

We selected the platform TryMyUI which has a panel of participants skilled in testing websites and answering questions while thinking aloud. TryMyUI provided 20 minute videos capturing participants' screens as they filled out the time diary and their voice as they thought aloud to complete the time diary. We used closed-ended probes to confirm whether the instructions on how to fill out the online time diary were effective, the task was clear, the questions were worded clearly, and the instrument worked as intended. For the most part, the instructions remained about the same with only minor tweaks and clarifications made, confirming that respondents would likely understand how to complete the self-administered version of the activity log.

Data Collection

Once the full online instrument was finalized, we checked it several times to ensure the skip patterns and functionality were working properly. This was done by previewing the instrument and ensuring selected answers took the study down the correct paths. Once we confirmed the study was working as intended, we conducted a brief pilot test with 10 participants recruited from MTurk. This was to ensure there were no major problems with data collection. We also downloaded the data file containing the recorded answers to ensure that the data output was as expected without any glitches, the random assignment to condition worked properly, and that everything was ready for the larger data collection. We also made sure the whole study from start to finish would take approximately 20 minutes to complete (this amount of time is recommended for online studies, or else participants begin to lose focus).

After we checked the pilot of 10 participants and were satisfied with those results, we posted the study on MTurk in sets of about 50 participants at a time, staggering sample release to include weekdays, weekends, daytime and evenings. This was done because, in the past, using MTurk has often led to very quick uptake of the study, and we wanted to maximize the chance of reaching different types of participants.

Samples

	Lab	**MTurk**	**Qualtrics Non-Probability Panel**
N	29 participants	1233 participants	939 participants
Gender	11 male; 18 female	54% female	52.2% female
Age (mean)	46 years (SD = 14)	36 years (SD = 11)	47 years
Age (range)	21 to 69 years	19 to 77 years	18 to 85 years
Other sample characteristics	None	None	Census quotas on gender, age, and education level
Data collection platform	N/A	Qualtrics	Qualtrics
Protocols	As many scripted and spontaneous probes of comprehension as needed	Few open-ended probes of comprehension	Few open-ended probes of comprehension

Data Cleaning

Once all of the data were collected, we went through the process of cleaning the data. For online studies, there are several ways of determining whether participants took the study as intended. The data collection platform we used allowed us to know the amount of time participants spent on the study and also on individual pages. For instance, if participants spent a very short time on the study, this could be an indicator of not taking the time to complete it as instructed. Participants who spent a very long time on the study may have been multi-tasking. We also used the time diary as a metric to help see which participants followed study instructions or not, omitting participants who entered too few activities in the activity log (e.g., three or fewer activities for the whole day), seemed to enter dates and times that indicated they did not read the instructions (i.e., not ending the final activity at 4am the day of the study as instructed), spent far too short of time completing the activity log, or reported very little sleep or far too much sleep on the activity log. We obtained most of these values by determining which participants could be considered "extreme outliers" on these metrics – either three standard deviations above or below the mean value. Some of these criteria were implemented by Qualtrics in that panel before we received the final data (i.e., those who took too short or too long a time to complete the study were removed in advance)– other calculations had to be made by the researchers upon receiving the data. Ultimately, we eliminated approximately 55 participants from both panels.

Results

We found that providing a specific definition of sleep brought the diary and stylized estimates slightly closer together, a small but significant effect. We also found a large order effect of the diary and stylized questions that either wasn't present, or was not detectable due to the small sample size, in the cognitive interview study. Participants who answered the diary first had a larger disparity between their reported sleep measures. Thus, the online study showed that comprehension of the word 'sleep' and the order in which the diary and stylized questions are asked may affect how people respond to questions about sleep.

Added Value of Online Testing

While it was not a complete replacement for the rich qualitative data obtained from the cognitive interview research, the online research was an excellent supplement to the laboratory research. The online study allowed for a larger sample size and the addition of experimental manipulations, allowing us to do analyses that wouldn't be possible with just lab studies. It also allowed us to use large samples to statistically test hypotheses generated from lab studies with quantitative data. The lab study uncovered information about the response process surrounding sleep questions and potential sources of measurement error. The online study helped confirm evidence that comprehension may play a role in how respondents report on sleep with quantitative data, and also enabled us to uncover a previously unknown order effect.

Case Study 2: Using TryMyUI

Background

The Bureau of Labor Statistics (BLS) collects data from establishments about job requirements through the Occupational Requirements Survey (ORS). As part of the ORS interview, interviewers (who are field economists trained to collect data from establishment respondents) collect information to determine specific vocational preparation (SVP), which is defined by the amount of lapsed time required by a typical worker to learn the techniques, acquire the information, and develop the competence needed for average performance. SVP levels can range from a short demonstration only to over 10 years.

To collect this information, interviewers are tasked with asking for four pieces of information: minimum education, professional certification requirements, prior work experience, and post-employment training. Preceding field tests had shown that the four items were not working as well as had been hoped, with reports of respondent misunderstanding and confusion. To assess reliability and accuracy of the field tests and to provide evidence for revising the target survey questions, a small focused study of telephone re-interviews using revised question wordings was planned. These calls were structured like cognitive interviews with spontaneous probing as needed and also served as ways to explore the causes of confusion.

Description of Online Testing Research

BLS wanted to use online cognitive interviews to supplement telephone re-interviews in a study to understand sources of possible respondent confusion due to question wording and design. Given that none of the four interviewers who would be conducting the telephone re-interviews were experienced or trained in cognitive interviewing methods, the research team wanted to use online testing methods to supplement their findings using controlled cognitive interview scripts and probes.

Research Questions

- Are the revised wordings of questions measuring Specific Vocational Preparation (SVP) clearly understood compared to the question wording used in a previous field test?
- Which parts of the question, if any, cause interpretation problems?
- Do results from online, unmoderated cognitive testing support or contradict results from the phone re-interview?

Selecting Recruitment and Data Collection Platforms

Given the focus on exploring sources of confusion in the question wording and design, we chose to use TryMyUI, an online unmoderated usability testing site similar to UserZoom, Loop11, Webnographer, and Usertesting.com. The participants recruited by TryMyUI receive training on thinking out loud and providing insights into their thought processes, which is critical for unmoderated cognitive interviews. Because TryMyUI participants are trained to read everything on the screen out loud, they typically do not skip instructions, which is also important for unmoderated cognitive interviews. Furthermore, because this study was interested in identifying sources of comprehension problems, a convenience sample was sufficient.

In order to use TryMyUI panelists, researchers must also use TryMyUI's data collection platform. Although TryMyUI was originally designed to conduct usability testing of websites and online tools, we were able to adapt the platform for our needs to conduct unmoderated cognitive interviews. However, the TryMyUI data collection platform only records the participant's audio and screen, and so the adaptation requires using another data collection platform to control the cognitive interviewing script, present the survey questions to the participants, and collect the survey responses. We chose to use SurveyMonkey in this case because our questionnaire needs were not complex and SurveyMonkey could be used to easily and quickly create our instrument.

Data Collection

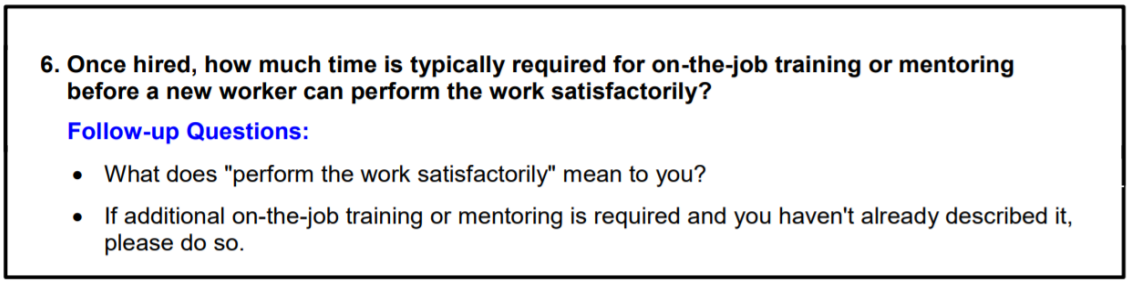
During the set-up of the project, we indicated the type of site to be evaluated as a "wireframe/prototype" so as to cue the participant that it is not a production system being tested. Since TryMyUI participants mostly participate in usability tests, it was important to explain that this is a different type of test, so the participants could re-orient themselves. In this case, the instructions to participants focused on directing them to carefully read the survey instructions. The TryMyUI instructions read:

> *This is not a typical usability test. Instead, the Bureau of Labor Statistics has developed questions that ask about the educational and training requirements for jobs, and we want your help testing them. To begin, please follow the instructions that appear on the web page to your right.*

And to cue participants to think out loud as if part of a cognitive interview, we used these instructions:

> *We would like your help evaluating questions that ask about the educational and training requirements for jobs.*
>
> *Please think out loud as you answer the questions.*
>
> *The questions should be clear and easily understood, so if anything causes confusion, please tell us.*
>
> *You may see the words Follow-up Questions on some pages, along with a question or two. These questions ask for additional information to help us better understand how you arrived at your answer.*

The participants were then asked to enter the title of his/her current job, along with a brief description of its main tasks. The survey questions were asked only about this one job. Since participants were "thinking out loud," they read the survey questions out loud, described their thought processes as they answered the question, and then read and answered the follow-up probes. An example survey question and its follow-up probes are shown below:



Before collecting our data, we conducted dry runs with a few participants. In addition to confirming whether the instructions and probe questions were interpreted by participants as intended, we also used the dry runs for testing whether the session would not last longer than our maximum session time (in our case, the limit was 20 minutes but it may be up to 30 minutes for other plans on the platform). As with any cognitive test, timing was somewhat tricky to interpret because some people will finish quickly (either have no problems with the questions or have little to say), whereas others will take the entire time. To ensure that participants did not feel rushed to finish the study, we included an instruction telling them that they wouldn't be penalized (rated poorly on the platform) if they ran out of time.

Samples

Eleven unmoderated online cognitive interviews were completed with participants from the TryMyUI panel. Despite this small number of cases, previous experience with think-aloud cognitive interviewing had demonstrated that small numbers of cases can still generate useful feedback and so we proceeded with the online study. Very general participant selection criteria were used: eligible participants must reside in the United States and be 25 to 75 years old. Because TryMyUI does not have a panel of establishments, we instructed participants to think about their own job or a previous job and respond as if they were answering a survey about that job. We specified our sample characteristic and size in advance to TryMyUI, which then used their frame information to recruit participants for our study. However, because TryMyUI operates on a pre-paid credit system, whereby you purchase credits that can later be used toward paying for participants, it is simple to add sample and modify eligibility criteria as needed.

In the parallel telephone re-interview study, 30 respondents from businesses that had participated in a preceding field test (which had ended four months before) were re-interviewed by telephone; establishments were selected to provide a wide range of industries and size classes. Respondents were asked the four educational requirement questions described above about two occupations.

Results

Data from both the telephone re-interviews and the unmoderated online cognitive interviews were coded by two researchers independently for possible sources of respondent confusion and misunderstanding. Because it was not possible to code these data blind to the mode of collection, we caution that there is potential for confirmation bias in the results.

The most problematic questions ("training/certification/licensing" and "time to adequate performance") were the most problematic in both testing modes, and similar misunderstandings were identified using both approaches. However, as previously mentioned, the phone

interviewers had been given the freedom to ask the questions in a different order. When doing so, they reported more success using an alternative question order, which seemed to address some of the respondent confusion. Given that the unmoderated online cognitive interview uses a fixed script, feedback about question order was only available through the re-interview. Had time and resources allowed, different question orders could have been tested in TryMyUI.

Interestingly, when asked directly in the respondent debriefing following the target survey questions, none of the 30 phone respondents reported that any of the target survey questions were confusing. A general consensus based on both the re-interviews and online cognitive interviews was that the set of revised questions used in the current study worked better than those used in previous field testing, but that some issues still remained. Therefore, a revised set of questions with a different order was generated for use in future tests. After the current test, the most effective question order was deemed to be: first ask questions about minimum education, then prior work experience, post-employment training, and required certification, licenses, or training. The full report can be found here: https://www.bls.gov/osmr/pdf/st150100.pdf

Added Value of Online Testing

Despite differences in testing methodologies, both the re-interview and the online test approaches provided useful, consistent feedback about the questions, which resulted in changes to the question wording. The online study allowed for the collection of more data and more diverse data than otherwise would have been possible and enabled the research team to control the cognitive interview questions and probes in a way that was not possible with the telephone re-interviews administered by the field economist interviewers. That the controlled format of the online testing supported the findings of the novice interviewers gave us confidence in the findings from the telephone re-interviews that we would not otherwise have had.

Appendix: Other sources of online participants

When researchers decide to not use online panels, they typically advertise a study on a website or mobile app (e.g., Facebook, Google AdWords, Reddit). Individuals who see these advertisements are not part of a panel and may not be looking for research studies when they encounter the advertisement. Researchers are effectively building a custom panel for their research project by themselves. The primary benefit to this approach is the possibility of reaching participants not otherwise available through online panels.

When paying to advertise on a platform, you may gain access to the platforms' built-in advertising tools to target demographics. These demographics are typically scraped from an individual's profile on the platform and range from basic characteristics such as country location to "premium" characteristics such as employment status, education, interests and hobbies, and behaviors on the platform. Some characteristics, age or gender for example, may be inferred based on the individual's behavior on the site if they are not listed explicitly in an individual's profile.

Some social media platforms may host populations so large that, while not representative of the general population, underrepresented groups are present in larger numbers than are available from online panels. However, the characteristics of your sample still depend on who sees your posting. This can vary based on the keywords you use to describe your posting (e.g., some people may use filters to control what postings they see), the people who follow or visit the site (postings to some platforms may only be displayed to individuals who explicitly allow content from identified individuals or organizations), and other factors, including ones that are out of the researcher's control, such as what other postings are shown at the same time (they may push your posting out of view).

As an alternative to social media postings or in conjunction with them, you may choose to recruit participants through snowball sampling or "word-of-mouth" methods. These methods include asking known individuals to ask others to complete the study or posting about the study on an organization's webpage. Relying on this method means that you as the researcher may lose control over factors such as sample characteristics and messaging regarding the purpose and context of the study. Additional screening through in-study embedded screener questions may help to target your desired population.

For all of these advertisement-based methods, screener questions at the beginning of your study can be tailored to your specific research question and can be used to qualify or disqualify individuals from participating in the study (e.g., using skip logic to end the study for any individual who reports not having health insurance in the last 12 months).