Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions

Jennifer Edgar¹, Joe Murphy², and Michael Keating³

Abstract
Cognitive interviewing is a common method used to evaluate survey questions. This study compares traditional cognitive interviewing methods with crowdsourcing, or “tapping into the collective intelligence of the public to complete a task.” Crowdsourcing may provide researchers with access to a diverse pool of potential participants in a very timely and cost-efficient way. Exploratory work found that crowdsourcing participants, with self-administered data collection, may be a viable alternative, or addition, to traditional pretesting methods. Using three crowdsourcing designs (TryMyUI, Amazon Mechanical Turk, and Facebook), we compared the participant characteristics, costs, and quantity and quality of data with traditional laboratory-based cognitive interviews. Results suggest that crowdsourcing and self-administered protocols may be a viable way to collect survey pretesting information, as participants were able to complete the tasks and provide useful information; however, complex tasks may require the skills of an interviewer to administer unscripted probes.

Keywords
cognitive interviewing, pretesting, crowdsourcing, survey methodology

Introduction
Cognitive interviews are a qualitative method often used to understand cognitive processes in the pretesting of survey questions (Willis, 1999). Interviews are typically conducted in a laboratory setting with members of a survey’s target population. Participants are recruited via posted flyers, word of mouth, newspaper ads, and, more recently, classified ads on online forums such as Craigslist (Murphy et al., 2007). Cognitive interview participants travel to the laboratory to complete face-to-face interviews and respond to prompts about their thought processes and reactions in answering questions. Although cognitive interviewing has been employed in different settings, such as with embedded probing in field interviews (Converse & Presser, 1986) and web probing has been used to capture open-ended items to assess item comprehension (Behr, Kaczmirek, et al., 2012), we characterize the traditional approach as involving the following combination of recruitment, setting, and administration: (a) recruitment via traditional media source (listed above), (b) travel by the participant to a cognitive interviewing laboratory or similar setting, and (c) interviewer-administered questions with scripted and unscripted probes.

Although results from cognitive interviews are not intended to be generalized, to best understand the appropriateness of survey questions, it is important to ensure that the sample for a cognitive interviewing study reflects, to the extent possible, the characteristics of the target population (Miller, Chepp, Wilson, & Padilla, 2014). However, cognitive interviews usually rely on a variety of convenience sampling methods (Baker et al., 2013). A researcher can only use cognitive interviews to better understand how questions may be interpreted, whether they are interpreted as intended, or how well they are suited for the intended target population.

Ideally, cognitive interview study samples represent the key subgroups of interest, that is, smokers or those who have made specific types of purchases (Miller et al., 2014), but commonly researchers must rely on participants who find the recruiting advertisements and are willing and able to participate in the given location at the given time. Almost always, cognitive interview studies are limited to one or two geographic areas, making the issue of geographic differences in language or how a given concept is experienced (e.g., participants who live in a moderate climate will think about “seasons” in a different way than someone who experiences four distinct seasons each year) of significant concern.

¹U.S. Bureau of Labor Statistics, Washington, DC, USA
²RTI International, Chicago, IL, USA
³RTI International, Research Triangle Park, NC, USA

Corresponding Author:
Joe Murphy, RTI International, 230 W Monroe, Suite 2100, Chicago, IL 60606, USA.
Email: jmurphy@rti.org

Creative Commons CC-BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
In addition to the limitations noted above, resources can often limit the number of cognitive interviews that can be conducted in advance of a survey. Although not often acknowledged in cognitive interviewing reports, sample sizes are often dictated by the amount of funding available, rather than the ideal number of participants. Working with a smaller than desired pool of participants can increase the chance that an important perspective is missed (Blair & Conrad, 2011). Time and resource limitations can also restrict the extent to which multiple rounds of testing can be conducted. Ideally, when pretesting new survey questions or forms, cognitive interviewing should be done using an iterative design, with each round building on the findings of the previous round (Willis, 2005).

Another limitation of cognitive interviewing concerns consistency of the approach used among those conducting the cognitive interviews. Cognitive interviewers can use both scripted and unscripted probes to delve into aspects of the response process to better understand issues with the survey questions. Especially with unscripted probing, interviewers can probe inconsistently across participants, and this has the potential to reduce the quantity of the data collected, if not all incomplete responses are probed for example. In addition to reducing the amount of information collected, thoughtful probing is required to ensure that high quality data are collected from all participants. If an interviewer probes inconsistently, or does not probe sufficiently to capture the true underlying response process, the quality of the data can be compromised. This impact on the quantity and quality of the data collected can be magnified in studies with more than one interviewer, when interviewers are probing inconsistently within and across their interviews.

Although cognitive interviewing has been a popular pre-testing method for decades, another research method termed “crowdsourcing” has gained popularity in recent years due to the speed and low cost with which information can be collected. Crowdsourcing has been described as the act of “tapping into the collective intelligence of the public to complete a task” (King, 2009) and involves recruiting a pool of willing individuals (the “crowd”) through convenient Internet platforms to complete discrete tasks, typically in their own homes, at a low cost, and in a short amount of time (Hill, Dean, & Murphy, 2013). The vast reach, quick deployment, and inexpensiveness of crowdsourcing address several of the limitations of traditional laboratory-based cognitive interviewing research. Using crowdsourcing, a researcher has low-cost access to a geographically and demographically diverse group of participants from whom information can be collected in a short time frame in a systematic way. Specifically, crowdsourcing may be a way to collect more information than what is traditionally collected in a cognitive laboratory setting, while overcoming some of the associated challenges of obtaining a diverse sample with limited resources.

In research, crowdsourcing has been used to generate ideas, collect data, and analyze large volumes of unstructured information (Keating, Rhodes, & Richards, 2013). In recent years, crowdsourcing has been applied to parts of the survey process, including brainstorming methods used during a survey design phase to drive a research agenda (Keating & Furberg, 2014), ad hoc field data collection for quality control (Kim, Lieberman, & Dench, 2013), supplemental data collection to add depth to survey data sets (Keating et al., 2013), and data analysis to predict study outcomes (Erdman & Bates, 2013).

Behrend, Sharek, Meade, and Weihe (2011) compared the quality of responses from a sample compiled from volunteers recruited on the crowdsourcing platform Amazon Mechanical Turk (MTurk) with those from a traditionally recruited sample of university students. They found few differences between the data collected from Mechanical Turk and students in terms of completion time or word count of responses. Quite importantly, they found Mechanical Turk to be much more efficient than the student sample, with hundreds of participants recruited in less than 48 hr and each requiring less than US$1 a piece in reimbursement for completion of a 30-min survey. Researchers (Behrend et al., 2011; Paolacci & Chandler, 2014) have found that Mechanical Turk offers a more diverse sampling pool than college campuses, suggesting that, in some situations, crowdsourced samples may be more representative than those obtained using traditional methods.

Given its low cost, speed, and geographic reach, we view crowdsourcing as a potentially viable option for recruiting participants and conducting survey pretesting activities. Here, we investigate the potential for crowdsourcing to recruit participants and conduct cognitive interviews. With crowdsourcing, there is typically no direct interaction between the person who designed the task (the researcher) and the person completing it (the participant). This means that all pretesting activities must be self-administered.

Exploratory research using an online panel and self-administered cognitive interviews found that the same types of information from a cognitive interview could be collected without the aid of an interviewer (Edgar, 2012, 2013). Responding to a series of questions on a web survey, participants were able to complete tasks commonly used in cognitive interviews: defining key terms and explaining their response process. Efficiencies were realized in the low cost and speed of data collection, and the demographic characteristics of the crowdsourced participants were generally comparable with the laboratory participants. These preliminary studies showed promise in the quantity and quality of the data collected online, but there were mixed results in the comparability of substantive findings when compared with traditional cognitive interviews for some questions tested (Edgar, 2012). A recent comparison of cognitive interview and online probing response quality found that although traditional cognitive interviews were higher in quality, the large
number of online responses compensated for their overall lower quality by uncovering a variety of error types and themes (Meitinger & Behr, 2016).

If a self-administered methodology can be used to collect information useful for the survey pretesting process, crowdsourcing participants could address some of the limitations found in traditional laboratory research. As noted above, crowdsourcing is a cost-effective way to collect information from a large, geographically diverse group of participants quickly. This would allow researchers to test in a more dispersed area and with more people than traditional methods would allow. In addition, some crowdsourcing platforms allow for sampling quotas to be used, letting researchers specify the key characteristics in advance of recruitment. Finally, using a self-administered methodology ensures that all participants are asked the same questions, and follow-up probes, in the same order. Potential interviewer effects are completely eliminated.

For this application, we define web probing–based crowdsourcing as involving the following combination of recruitment, setting, and administration: (a) recruitment via an automated Internet-based system, (b) no travel required or direct interaction between the participant and researcher, and (c) self-administered questions with scripted follow-up probes. Although crowdsourcing appears at face value to offer some advantages to traditional cognitive interviewing, there is some limitation in using it for research purposes. First, all participant demographic information is self-reported, so the accuracy of sample composition depends on the honesty of participants. To the extent that participants are motivated to misrepresent themselves to meet the requirements for a given project, and that the inaccurate characteristics relate to the information being collected, the conclusions may be unreliable.

Other crowdsourcing limitations overlap with those faced in traditional cognitive laboratory recruitment. Participants for both methods must volunteer and agree to participate to receive monetary compensation in the form of an incentive, and their motivation to expend effort when completing the tasks or interview may be low. Whereas in the laboratory, interviewers are able to gauge the level of effort a participant is extending for the interview and can choose to exclude the data if they believe there are issues with the data, there is no direct way to gauge the quality of crowdsourced information.

Building on past work, we consider three alternative crowdsourcing platforms for cognitive interview recruitment. We compare results from each platform with data collected using traditional laboratory methods in terms of participant characteristics, and cost. We also explore the effectiveness of collecting pretesting information using a self-administered form, evaluate the quantity and quality of data collected, and consider the effects on the substantive conclusions that would be drawn using each method.

Method
Recruitment

For this study, we recruited participants in four different ways. The first approach (traditional) recruited local participants from four different areas across the United States. The primary methods of recruitment were newspaper ads, flyers, word of mouth, and Craigslist ads. A total of 71 participants were recruited. The researcher defined demographic targets, or quotas, for recruitment. Researchers aimed to fill each quota, but when not enough participants were available to fill a quota, researchers generally accepted the next available participants. In addition to meeting the demographic requirements, participants had to be available to come to the laboratory or interview site during the times specified by the researcher—generally business hours on a weekday. Interviews were conducted in three states in the spring of 2006 (15 in New Mexico, 17 in Washington, and 20 in Wisconsin) and with 22 participants in Washington, D.C., in the summer of 2013. Participants answered 43 survey questions and follow-up probes, as well as additional unscripted probes based on their answers. Interviews ranged in duration, but averaged 45 min.

The second method of recruitment was an online, remote usability platform called TryMyUI. The service coordinates recruitment of and payment to participants, administration of tasks, collection and storage of audio recording of the respondent’s voice responses, and video recording of the respondent’s computer screen during the task. TryMyUI has a volunteer panel of participants with a range of known demographic characteristics and backgrounds, allowing for quotas to be set. TryMyUI typically asks panel members to evaluate websites and provide verbal feedback. Each participant receives a rating from the researcher after completing the task, and only panel members with high ratings are eligible to complete future tasks. Because the members typically evaluate websites and other online materials, they are used to reacting to stimuli and expressing their opinions out loud, experience which has the potential to help them provide in-depth responses to cognitive interviewing questions asking about their thought processes or response strategy. Researchers are able to specify quotas (e.g., five males with high school education), and as eligible participants come to the site, they may complete the task, which is offered as a “first come, first served” process until the quotas have been filled. Depending on the specifications of the quotas, TryMyUI may not be able to fill each quota (e.g., if there are not enough males aged 65 or older with a college degree), at which point the researcher is able to redefine the target quotas or increase the size of another group. After accepting the task, participants were directed to a 43-question web survey, designed to mirror the interview protocol used in the traditional cognitive interviews. TryMyUI limits tasks to 20 min, and most participants were able to complete the web survey in that time. We collected data from 44
self-administered interviews with participants recruited from TryMyUI in February 2013.

The third method used Amazon’s Mechanical Turk, one of the most popular current crowdsourcing platforms. Mechanical Turk allows for requesters to post human intelligence tasks (HITs) and for workers to complete these HITs for payment. At any given point, hundreds of thousands of HITs are available on the website, and there are currently more than 500,000 workers on Mechanical Turk. Most HITs are “micro-tasks,” taking only a few seconds or minutes to complete with a very small incentive, often less than 5 cents. Mechanical Turk is increasingly being used by researchers to recruit participants and to collect survey response data (Christenson & Glick, 2013; Chunara et al., 2012). Because Mechanical Turk HITs are typically very short, we included only one section of the full web survey in each HIT; each Mechanical Turk participant only answered 10 to 15 questions, estimated to take 5 min to complete. We collected data from 1,019 self-administered interviews with participants recruited from Mechanical Turk in September 2013.

The fourth method of recruitment used Facebook, a popular social networking site with more than a billion active users worldwide, including the majority of U.S. adults (Pew Research Center, 2015). We purchased advertising space targeted to specific demographic groups or those with particular interests or “likes.” For this study, we used targeted ads to let participants know they could receive a US$5 electronic gift card for completing a 10-min survey. Given that we were interested in a general population for this cognitive interview project, we presented the ads only to English-speaking Facebook users 18 years or older and living in the United States. Participants who clicked on the ad were directed to an abbreviated version1 of the web survey, estimated to take 10 min to complete. We collected data from 60 self-administered interviews with participants recruited from Facebook in September 2013.

Table 1 provides a summary of the study design.

Table 1. Recruitment Method and Interview Mode Summary.

| Interview mode | Traditional (classified ads, word of mouth, etc.) | TryMyUI | Mechanical Turk | Facebook |
|----------------|---------------------------------------------------|---------|------------------|----------|
| Face-to-face in cognitive laboratory | 71 | 44 | 1,019 | 60 |
| Average questions answered per respondent | 43a | 43 | 13 | 20 |
| Average length of interview (minutes) | 45 | 20 | 5 | 10 |

*aNot including unscripted probes.

Collecting the Data

A variety of questions were selected to test using the four different methods. Questions were selected based on their likely relevance to a variety of participants. They had all been used in prior pretesting studies, with relevant information about response processes and reactions able to be collected.

We followed traditional procedures for the laboratory cognitive interviews (Willis, 2005). A single interviewer conducted all interviews, using a scripted protocol. In general, participants first answered the survey question and then were asked follow-up probes aimed at understanding their response process. After the scripted probes, the interviewer asked unscripted probes as necessary to obtain as complete an understanding as possible. We used both open-ended probes (e.g., tell me what you thought about when you heard the question) and closed-ended probes (e.g., are the following items included in the category?). We audio recorded the interviews and took notes during each interview.

For the three crowdsourcing designs, we developed web survey versions of the cognitive protocol for self-administration. Follow-up probes, mirroring those used in traditional cognitive interviews, followed each survey question. For TryMyUI, a full-length web survey was administered. Two of the topics in this study (flu and smoking) were not included for TryMyUI, but additional topics from the laboratory interviews were. We administered abbreviated versions of the web survey to the Mechanical Turk and Facebook participants to align the estimated completion time with participants’ expectations for a short task. For Mechanical Turk, we developed three sub-surveys, each focusing on a single topic. For Facebook, we only included the topics discussed in this article in the web survey. Table 2 includes a list of the questions administered by data collection mode.

Within the protocol, we included typical cognitive interviewing tasks to evaluate survey questions of different substantive topics. There is variation in what a participant must do for each type of task, with each expected to have a different level of cognitive complexity. Reporting examples of a type of item is likely to be less complex than articulating the mental calculations used to arrive at an estimate. In Table 3, from top to bottom, we list the tasks in order of complexity, ranging from straightforward (e.g., provide examples of sportswear) to complex (e.g., explain response process in determining clothing expenditures).
Table 2. Questions Administered by Method.

| Module                  | Question                                                                 | Traditional laboratory | TryMyUI | Facebook | Mechanical Turk |
|-------------------------|--------------------------------------------------------------------------|------------------------|---------|----------|-----------------|
| Sportswear module       | Since the first of [reference month] have you or any member of your household purchased any swimsuits or warm-up or ski suits? | X                      | X       | X        | X               |
|                         | Yes                                                                      |                        |         |          |                 |
|                         | No                                                                       |                        |         |          |                 |
| Probe                   | What types of items did you think of when you saw that question about swimsuits, or warm-up or ski suits? | X                      | X       | X        | X               |
| Flu shot module         | Have you had a flu shot this season?                                     | X                      | X       | X        |                 |
|                         | Yes                                                                      |                        |         |          |                 |
|                         | No                                                                       |                        |         |          |                 |
| Probe                   | In your own words, how do you define the word “season” when you read the question above? | X                      | X       | X        |                 |
| Smoking module          | Have you smoked at least 100 cigarettes in your entire life? [SELECT ONE] | X                      | X       | X        |                 |
|                         | Yes                                                                      |                        |         |          |                 |
|                         | No                                                                       |                        |         |          |                 |
| Probe                   | How did you mentally calculate the answer to the last question?          | X                      | X       | X        |                 |
| Clothing module         | How much have you or any member of your household spent on clothing since the first of May? | X                      | X       | X        | X               |
|                          | Probe 1 What did you think about when deciding how much you’ve spent on clothing? | X                      | X       | X        | X               |
|                          | Probe 2 What clothing items have you or any member of your household bought since the first of October and how much did each cost? Please fill in up to 6 items (Example: 1 shirt, US$20). | X                      | X       | X        |                 |
| Demographics module     | Q1 What is your gender?                                                  | X                      | X       | X        | X               |
|                         | Male                                                                     |                        |         |          |                 |
|                         | Female                                                                   |                        |         |          |                 |
|                         | Q2 What is your age?                                                     | X                      | X       | X        | X               |
|                         | 18-24                                                                    |                        |         |          |                 |
|                         | 25-34                                                                    |                        |         |          |                 |
|                         | 35-44                                                                    |                        |         |          |                 |
|                         | 45-54                                                                    |                        |         |          |                 |
|                         | 55-62                                                                    |                        |         |          |                 |
|                         | 63 or older                                                              |                        |         |          |                 |
|                         | Q3 What is the highest level of education you have completed?             | X                      | X       | X        | X               |
|                         | Less than high school                                                   |                        |         |          |                 |
|                         | High school degree                                                       |                        |         |          |                 |
|                         | Some college                                                             |                        |         |          |                 |
|                         | College degree                                                           |                        |         |          |                 |
|                         | Some graduate school                                                     |                        |         |          |                 |
|                         | Completed graduate school                                                |                        |         |          |                 |
|                         | Q4 What is your approximate average annual household income?              | X                      | X       | X        | X               |
|                         | US$0-US$14,999                                                           |                        |         |          |                 |
|                         | US$15,000-US$29,999                                                      |                        |         |          |                 |
|                         | US$30,000-US$49,999                                                      |                        |         |          |                 |
|                         | US$50,000-US$74,999                                                      |                        |         |          |                 |
|                         | US$75,000 and above                                                      |                        |         |          |                 |
TryMyUI participants answered the follow-up probes (e.g., what did you think of . . . ?) as verbal think-alouds recorded by the computer and then transcribed. Mechanical Turk and Facebook participants provided typed answers to all questions via a web form.

In analyzing the data, we calculated the number of words provided in each response by participants and the number of examples provided in their answers. We also subjectively rated the relevance (yes or no) and usefulness of the response (on a 4-point scale, from completely unusable to complete). The subjective metrics were blind coded by two researchers, and inter-rater reliability was calculated using Cohen’s kappa. We made one round of revisions to the coding instructions based on preliminary reliability results, and researchers revised their codes as applicable. Each final kappa exceeded $.70$ for both subjective measures, suggesting a good level of agreement. To illustrate the coding process, we present sample responses and how we coded them in Table 4.

### Results

#### Recruitment

To evaluate the extent to which the different methods resulted in samples mirroring the general population (the target population for the tested survey questions), we compared participant demographics with various benchmark data sources. Although the interviews were conducted over a wide time span (laboratory participants as far back as 2006), a review of the selected benchmarks does not suggest substantial changes over the course of our study period. Demographic

---

**Table 3. Tasks, Survey Questions, and Follow-Up Probes.**

| Task | Survey question | Follow-up probe |
|------|-----------------|-----------------|
| Provide examples (sportswear) | Since the first of [reference month], have you or any member of your household purchased any swimsuits or warm-up or ski suits? | What types of items did you think of when you saw that question about swimsuits, or warm-up or ski suits? Please list up to six items. |
| Define concept (flu season) | Have you had a flu shot this season? | In your own words, how do you define the word “season” when you read the last question? |
| Explain answer (smoking) | Have you smoked at least 100 cigarettes in your entire life? | How did you mentally calculate the answer to the last question? |
| Explain response process (clothing) | How much have you or any member of your household spent on clothing since the first of [reference month]? | What did you think about when deciding how much you’ve spent on clothing? |

**Table 4. Sample Responses to Follow-Up Probes.**

| Task | Follow-up probe | Sample response | Number of words | Relevant response? | Completeness rating |
|------|-----------------|-----------------|-----------------|--------------------|---------------------|
| Provide examples (sportswear) | What types of items did you think of when you saw that question about swimsuits, or warm-up or ski suits? | “One piece, bikinis, sweatpants” “Comfort” | 4 | Yes | Complete |
| | | | 1 | No | Completely unusable |
| Define concept (flu season) | In your own words, how do you define the word “season” when you read the last question? | “The period between October and March” “It is fall, and this is the time of year when we discuss such things” | 6 | Yes | Complete |
| | | | 15 | No | Some usable information |
| Explain answer (smoking) | How did you mentally calculate the answer to the last question? | “I’ve never smoked in my life, so it was easy” “Wouldn’t need to” | 10 | Yes | Completely unusable |
| | | | 3 | No | Completely unusable |
| Explain response process (clothing) | What did you think about when deciding how much you’ve spent on clothing? | “I thought about what I bought, how many of them, and the quality of them to estimate the cost” “How much clothing was absolutely needed” | 19 | Yes | Mostly complete |
| | | | 6 | No | Completely unusable |
quotas were requested from both TryMyUI and the lab methods, but given the nature of the convenience sampling (e.g., sampling from the online panel), they were not always met. For Mechanical Turk and Facebook recruitment, it was not possible to screen participants based on demographic characteristics and so we did not use quotas.

In Table 5, we present participant demographics and answers to substantive questions by method. We did not conduct tests for statistical significance given the nonprobability nature of the recruitment. By gender, all methods were comparable with the benchmark, with the exception of the TryMyUI sample, which was 75% female. TryMyUI overrepresented the 25- to 44-year-old age group compared with the benchmark, and underrepresented the 63+ group. Facebook participants were overrepresented by 18- to 24-year-olds and underrepresented 25- to 44-year-olds relative to the benchmark. The laboratory and TryMyUI cases were more concentrated in the Western United States compared with the benchmark and included fewer states than Mechanical Turk and Facebook. All methods provided samples that were more highly educated than the benchmark—those with a high school degree or less education were underrepresented compared with the benchmark. Participants by method generally matched the benchmark income distribution.

We also compared participants with external benchmarks to ensure that we were not recruiting a biased sample for the cognitive interviews and that participants did not differ from national averages on the substantive questions being studied (Table 5). For instance, one survey question asked whether the participant had smoked at least 100 cigarettes in his or her life. Ideally, the percentage of participants who report yes would be consistent across method, and similar to a national benchmark. The laboratory and TryMyUI cases

| Table 5. Participant Demographic Characteristics by Method (and Ratio of Method to Benchmark). |
|---------------------------------------------------------------|
| Benchmark | Traditional Laboratory | TryMyUI | Mechanical Turk | Facebook |
|-----------|--------------------------|---------|-----------------|----------|
| n         | 71                       | 44      | 1,019           | 60       |
| Gender    |                          |         |                 |          |
| Male      | 49%                      | 48% (0.98) | 25% (0.51) | 53% (1.08) | 40% (0.82) |
| Female    | 51%                      | 52% (1.02) | 75% (1.47) | 47% (0.92) | 60% (1.18) |
| Age       |                          |         |                 |          |
| 18-24     | 13%                      | 10% (0.77) | 11% (0.85) | 16% (1.23) | 27% (2.08) |
| 25-44     | 34%                      | 27% (0.79) | 55% (1.62) | 38% (1.12) | 17% (0.50) |
| 45-62     | 31%                      | 37% (1.19) | 15% (0.48) | 23% (0.74) | 35% (1.13) |
| 63+       | 21%                      | 26% (1.24) | 5% (0.24)  | 2% (0.10)  | 22% (1.05) |
| Region    |                          |         |                 |          |
| Northeast | 18%                      | 0% (0.00)  | 16% (0.89) | 20% (1.11) | 25% (1.39) |
| South     | 37%                      | 31% (0.84) | 32% (0.85) | 41% (1.11) | 22% (0.59) |
| Midwest   | 21%                      | 27% (1.29) | 16% (0.76) | 27% (1.29) | 27% (1.29) |
| West      | 23%                      | 42% (1.83) | 36% (1.57) | 17% (0.74) | 25% (1.09) |
| States represented | 51     | 6 (0.12)  | 24 (0.47)  | 45 (0.88)  | 30 (0.59)  |
| Education |                          |         |                 |          |
| High school degree or less | 45%     | 24% (0.53) | 23% (0.51) | 9% (0.20)  | 23% (0.51) |
| Some college/college degree | 45%     | 57% (1.27) | 60% (1.33) | 80% (1.78) | 62% (1.38) |
| Graduate degree | 10%     | 19% (1.90) | 17% (1.70) | 11% (1.10) | 15% (0.80) |
| Annual household income |         |         |                 |          |
| US$0-US$30K | 23%     | 33% (1.43) | 25% (1.09) | 29% (1.26) | 33% (1.43) |
| US$30K-US$75K | 36%     | 36% (1.00) | 43% (1.19) | 53% (1.47) | 33% (0.92) |
| US$75K+    | 41%                      | 30% (0.73) | 32% (0.78) | 18% (0.44) | 33% (0.80) |
| Percentage who have smoked at least 100 cigarettes in entire life | 39%     | 18% (0.46) | Not collected | 40% (1.03) | 37% (0.95) |
| Average amount spent on clothing in last 3 months | US$360  | US$450 (1.25) | US$260 (0.72) | US$295 (0.82) | US$252 (0.70) |

*Benchmark: 2013 Estimates of the Resident Population by Single Year of Age (18+), U.S. Census Bureau, Population Division.
*Benchmark: Adult civilian persons, 2013 Current Population Survey, Annual Social and Economic Supplement.
*Including Washington, D.C.
*Collected from 21 participants.
*Collected from 30 participants.
*Benchmark: 2012 National Health Interview Survey.
*Collected from 255 participants.
*Benchmark: 2013 Consumer Expenditure Survey.
have not smoked 100 cigarettes compared with the benchmark. Mechanical Turk and Facebook respondents were comparable with the benchmark on this item. For clothing expenditures, laboratory responses resulted in a higher amount than the benchmark, and the crowdsourcing methods resulted in lower mean expenditures than the benchmark. Although differences may be a result of the small sample size, they may also indicate that research samples are somehow substantively different than the survey population of interest.

In terms of cost, the primary and most directly comparable cost is participant incentives. Across all methods, participant incentives were contingent on a completed interview or survey. For TryMyUI and Mechanical Turk, the incentives were normed to payment expectations on the respective platforms. We provided laboratory participants with a US$40 incentive for a 45-min average interview, TryMyUI participants received US$10 for a 20-min task, Mechanical Turk participants received US$0.75 for a 5-min task, and Facebook participants received US$5 for a 10-min task. Additional laboratory costs included interviewer and recruiter labor and check processing fees for the laboratory. It is estimated that the recruiter spent about 1 hr per interview to screen, schedule, and confirm appointments. Each laboratory interview also required about 1 hr of interviewer labor. The time spent developing the web survey and setting up the study was relatively consistent across the three crowdsourcing modes. Additional costs for the crowdsourcing designs included a US$20 per-participant fee in addition to incentives for TryMyUI, a 10% fee on HITs in Mechanical Turk, and US$300 for Facebook advertising, which was based on the number of clicks on each ad (i.e., pay-per-click). Only about one person in 10 who clicked on a Facebook ad completed an interview. Researcher time spent compiling and analyzing the web survey data was consistent for each question regardless of mode, with the exception of TryMyUI, which required individual participant audio files to be transcribed before they were analyzed. Table 6 provides a summary of incentives and costs by design.

### Quantity of Data

Cognitive interviews aim to elicit information from participants about their response processes to evaluate the questions being tested. Generally, researchers find more reliance in the conclusions as the amount of information collected increases. In the laboratory, interviewers are able to use various methods, such as follow-up probes or encouraging facial expressions, to elicit additional information from participants. When using self-administered modes, however, we rely on the participant to provide enough information to be useful without the benefit of interviewer intervention. One way to evaluate the successfulness of self-administered modes in eliciting information is to consider the number of words that participants used in their responses.

Examining the number of words in participants’ responses, we see in Table 7 that the number of words used by laboratory participants increases with the complexity of the task, from a median of four for providing examples of sportswear to 32 for explaining their response process in reporting clothing expenditures. For the simpler tasks (providing examples of sportswear and defining flu season), word counts for the crowdsourcing designs were comparable or even higher than in the laboratory, suggesting that the presence of the interviewer and ability to probe may not be as strong a benefit for the simpler tasks as it is for the more complex. Because interviewers have the ability to follow up with additional probes in the laboratory, the success in obtaining a higher median

### Table 6. Participant Incentives and Other Costs by Design.

|                      | Traditional (classified ads, word of mouth, etc.) | TryMyUI | Mechanical Turk | Facebook |
|----------------------|--------------------------------------------------|---------|-----------------|----------|
| Incentive per interview | US$40                                            | US$10   | US$0.75         | US$5     |
| Incentives per hour of participant time | US$53                                            | US$30   | US$9            | US$30    |
| Additional costs | Interviewer labor, US$3.25 processing fee per incentive | Additional US$20 per-participant service fee | Additional 10% service fee | US$300 in advertisements |

### Table 7. Median Number of Words in Response to Follow-Up Probe by Design and Task.

| Task                                | Laboratory | TryMyUI | Mechanical Turk | Facebook |
|-------------------------------------|------------|---------|-----------------|----------|
| Provide examples (sportswear)       | 4          | 9       | 8               | 7        |
| Define concept (flu season)         | 6          | Not collected | 10              | 4        |
| Explain answer (smoking)            | 12         | Not collected | 14              | 7        |
| Explain response process (clothing) | 32\*       | 44      | 15              | 7        |

\*Data captured after all probes.
count of words for the more complex tasks is not surprising and illustrates the potential downside of self-administered cognitive interviews. In general, additional, unscripted probes were only used for the response process explanation (clothing) tasks, as most laboratory participants completed the other tasks without the need for additional probing, again supporting the idea that interviewers may not be needed to collect information on straightforward tasks. It should be noted that our count included all words, not only relevant words. A review of the data however suggests that the proportion of relevant words does not vary by total length of the responses. Furthermore, we examine quality of the data in the next section.

Quality of Data

Beyond the quantity of data obtained by the different methods, we sought to compare the quality of responses obtained. The first dimension of quality we examined was relevance—that is, whether the response was appropriate given the topic of the follow-up probe. As shown in Table 8, there was not a consistent pattern of relevant responses by task or design. For the most straightforward task, providing examples of sportswear, a majority of participants provided relevant responses, from a low of 68% for TryMyUI to a high of 98% for Mechanical Turk. Although it may be surprising that 32% of TryMyUI participants provided irrelevant responses, it could be that the task was in fact too straightforward, especially compared with the website usability tasks they commonly complete. The irrelevant responses indicated a misunderstanding of the task—it may be that participants assumed that because the question itself included examples (swimsuits, warm-up or ski suits), the follow-up probe was asking for something else. Irrelevant responses to the sportswear item tended to focus on situations when people used sportswear, or times when they had bought sportswear, rather than example items.

The task of explaining “flu season” appears to have been confusing for some participants. In the laboratory, 68% provided a relevant response. Mechanical Turk’s rate of relevant responses was 86% and Facebook’s was 80%. Irrelevant responses to this item may have resulted from the fact that people are generally not used to explaining what commonly used terms mean. The fact that laboratory participants provided the lowest percentage of relevant responses may suggest an interviewer effect; it is possible that participants did not expect the interviewer to violate conversational norms in asking something so straightforward. When asked on a self-administered survey, however, participants appeared to take the task at face value and a higher percentage were able to provide a relevant response.

As opposed to the flu question, most participants provided relevant responses for the smoking question—93% for the laboratory and Mechanical Turk and 97% for Facebook. This task asked participants a probe that likely made sense, especially in the context of having just answered the survey question. The task of explaining the response process in responding to the clothing item was most complicated, though nearly all participants in the laboratory and TryMyUI provided relevant responses (100% and 96%, respectively). Even when self-administered, a majority of participants were able to provide a relevant response—81% for Mechanical Turk and 80% for Facebook. In all, there was no clear evidence that the crowdsourcing designs or self-administration resulted in substantially lower levels of relevant answers collected than the laboratory; the quality of the information collected, in terms of relevance, was generally consistent across modes.

From Tables 7 and 8, we see that quantity in terms of word count and quality in terms of response relevance were generally comparable across designs, especially for the more straightforward tasks included in the study. Differences begin to emerge, however, as participants are asked to complete more complicated tasks such as explaining their response process, with quantity and quality being higher for the laboratory. It may be that these more complicated tasks are where the presence of an interviewer, who can probe for more information and provide clarification of the tasks, is critical.

To delve further into the quality differences for the most complex task in this study, we considered the quality of responses in terms of completeness. That is, beyond simply providing a relevant response, was the response complete enough to be able to assess a participant’s response strategy? We coded the responses from the explaining response process task in terms of completeness using four categories, as described in Table 9.
Using the completeness ratings, we assigned each response a “grade,” to allow for comparison of the results across modes. An A indicates a high quality, complete answer, and an F indicate a poor incomplete answer as described in Table 9. Table 10 provides the distribution of the usefulness ratings to the clothing expenditure follow-up probe. We focus on this single item as an example, because it was included in all treatments with a very similar presentation. We present the laboratory distribution both with just the initial probe (“how did you arrive at your answer?”) as well as the total response after any unscripted probing the interviewer added to clarify the response (average of 2.4 unscripted probes per participant). For each design, we also calculated an average “grade” indicating the overall quality of the responses in terms of completeness. As with the other quality ratings, this measure is subjective, but gives a sense of the overall success of each design in obtaining usable information for this complex measure. We calculated “grades” as follows: complete—no additional information needed = 100/A, mostly complete = 85/B, some usable information = 75/C, completely unusable = 60/F.

We find that based on just a single, scripted, follow-up probe, none of the methods was able to obtain useful information from a majority of participants. Examining the usable responses, we see that many participants did not understand the task at hand even if their responses were relevant or on topic. Many discussed the reasons they buy clothing, personal situations, or the types of clothing they like to buy rather than their thought process in determining their answer. Such misinterpretations of the task probably could have been easily redirected by an interviewer with some feedback and a follow-up probe. In fact, when the additional, unscripted follow-up probes are factored in, the laboratory interviewers are able to gain useful feedback from almost all respondents; with all probes, the laboratory design receives an A grade. With the single scripted probe, the laboratory design receives a C grade.

Among crowdsourcing designs, we see TryMyUI as the most effective (with a grade of C), followed by Mechanical Turk (D), and then Facebook (F). TryMyUI participants provided the “best” answers to this probe among the crowdsourcing methods, perhaps aided by verbal rather than typed response or their experience thinking out loud and explaining their reactions in past usability tasks. In essence, they are already trained for this type of task. TryMyUI and Mechanical Turk participants are motivated to successfully complete the task because

### Table 9. Assessed Usefulness for Clothing Expenditures Probe Responses.

| Assessed usefulness     | Description                                                                 | Sample response                                                                 |
|-------------------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Complete (A)            | Enough information to be able to code response strategy without probing      | “How much have I spent on footwear? I bought two pairs of shoes, and they were $50 a pair, so I came up with $100” |
| Mostly complete (B)     | Only a little probing would be required to be able to code response strategy | “I remember we went clothes shopping because we needed some shorts. There were some good sales so we bought a few more things. We are very happy with our purchases, it was out of need, so that is why we made the purchases.” |
| Some usable information (C) | Some information to identify a response strategy, but considerable probing would be needed to code a response strategy | “The amount of individual clothing items I bought” |
| Completely unusable (F) | No information that could be used to code a response strategy               | “Quality, if I need it” “Our income”                                           |

### Table 10. Usefulness of Clothing Expenditures Probe Responses by Design.

| Level of usefulness                  | Usefulness “grade” | Lab probes | Scripted probe only | TryMyUI | Mechanical Turk | Facebook |
|--------------------------------------|--------------------|------------|--------------------|---------|-----------------|----------|
| Complete—No additional information needed | A                  | 89         | 5                  | 7       | 2               | 0        |
| Mostly complete                      | B                  | 6          | 25                 | 50      | 17              | 7        |
| Some usable information              | C                  | 4          | 50                 | 39      | 62              | 37       |
| Completely unusable                  | F                  | 2          | 20                 | 4       | 19              | 53       |
| Average “grade”                      |                    |            |                    |         |                 |          |
they have volunteered for the panel and know they will be evaluated by the researcher. Facebook participants, by comparison, likely had the least motivation to provide useful information in the interviews, as they were not actively seeking “work” when they were approached to complete the task and have nothing to gain personally beyond the incentive for participation.

Regardless of the method, a majority of crowdsourced participants were not able to complete this complex task without the aid of an interviewer, whereas in the laboratory, the interviewer was able to use probes to collect usable information from almost 90% of participants. This suggests that crowdsourced, self-administered web surveys are better suited for the simpler comprehension tasks but that the more complex cognitive interviewing tasks require an interviewer to appropriately guide the participant through the process.

Finally, we were interested to learn whether crowdsourcing would result in the same conclusions as the laboratory-based cognitive interviews. That is, would each design lead us to the same understanding of respondent reactions or need for item revision? For the four tasks included in this study, we assessed how conclusions from the crowdsourcing designs compared with the traditional design. For providing examples of sportswear and defining flu season, we compared the percentage accurately comprehending the question based on their response to the probe. For the explanation of the answer to the smoking question probe and the type of response process for expenditures on clothing, we compared the percentage using a reasonable response strategy. Next, we assessed whether the results suggest each crowdsourcing design supplied the same conclusion as was gained from the laboratory cognitive interviews. We considered the conclusions the same if the crowdsourcing metric evaluated was 10 or fewer percentage points from the traditional metric. Although this threshold is somewhat arbitrary, it is a reasonable approach to compare the conclusions by design.

Table 11 provides the results of these comparisons by task and design. If the laboratory design, with its known limitations, is used as the basis for comparison, all three crowdsourced designs lead to the same conclusions for the simplest of the tasks, providing examples (sportswear). The same conclusion was also reached across methods for the explanation of answer (smoking). For concept definition (flu season), Mechanical Turk led to a different conclusion than the laboratory or Facebook results. As the most complicated task, the conclusions reached for response process explanation (clothing) differed across all three modes. This inconsistency is critical because it suggests that researchers would reach a different conclusion if they used only one of these designs, and likely take a different action in terms of revising the question, depending on the designs.

**Discussion**

In this study, we compared traditional laboratory-based cognitive interviewing with web-based self-administered interviews with respondents recruited from three crowdsourcing platforms. These results suggest that crowdsourcing offers the potential to recruit a larger number of participants who are more diverse geographically and at a lower cost than traditional methods. Even for the crowdsourcing platforms that did not allow the setting of quotas, similar representation among demographic subgroups was obtained, and all crowdsourced samples were more geographically and demographically diverse. The ability to collect information from a wider range of participants may help researchers achieve more representative samples, and strengthen the results and recommendations. The inability to set—and achieve—quotas in some crowdsourcing platforms is a limitation when the researcher needs to control the distribution of participants along one or more dimensions. In our case, strict quotas were not sought and we were able to achieve a fairly balanced sample without the use of quotas.

The cost and time required to collect information required to evaluate questions is a common limitation for traditional cognitive interviewing. Crowdsourcing provides a promising alternative to overcome this obstacle by providing researchers a low-cost approach for quickly collecting data from a large number of participants. Eliminating the need for a recruiter to screen, schedule, and remind participants saves considerable time, and in general participant incentives are generally lower when crowdsourcing than when using traditional methods. There may be additional time on the back-end however, with self-administered interviews requiring a review of the recording and/or open-ended answers to analyze the data because there are no interviewer notes or summaries.

In terms of data quantity and quality, for more straightforward cognitive tasks, such as asking participants to provide examples of sportswear, the crowdsourcing alternatives provided similar quality results. However, for more complicated tasks, such as understanding a participant’s response strategy, the quality of the information collected using the self-administered crowdsourcing alternatives was inferior to the traditional methods. When a purpose of the interview is to capture a nuanced cognitive process such as response strategy, the presence of an interviewer administering unscripted probes can be crucial. For instance, when using a single, scripted probe “how did you arrive at your answer?” to follow up a question on clothing expenditures, most participants were unable to provide obtain useful information, regardless of the method of recruitment or mode of administration, but additional unscripted interviewer probes were successful. Although more specific scripted probes may be more effective than general ones, it may also be some tasks (e.g., straightforward comprehension tasks) are better suited for crowdsourced, self-administered testing; more complex cognitive interviewing tasks require an interviewer to appropriately guide the participant through the process.
Our research suggests that crowdsourcing can be a viable “fit for purpose” supplement to traditional cognitive interviewing, particularly when there may be regional differences or input from a large sample is needed quickly. It should be noted that the Mechanical Turk platform, in particular, does not allow for the selection of specific sub-populations for recruitment, which means it is most efficient for general population studies. However, some services now offer, at an additional cost, Mechanical Turk samples pre-screened for a variety of demographic characteristics (Fowler, Willis, Moser, Ferrer, & Berrigan, 2015).

Comparative analyses of the recruitment costs and quality with specific sub-populations is a topic well suited for future study.

From our findings, web self-administration appears to be a viable cognitive interviewing mode for simple tasks. Incorporating crowdsourcing during the preliminary stages of cognitive interview studies may leverage the advantages of the approach, allowing researchers to gain insight into concept comprehension by collecting information from large numbers of participants. By collecting information from larger number of participants, we are able to collect information that might be missed with traditional sample sizes. Although we found that lower quality data were collected from the crowdsourced platforms, it might be possible to compensate for the quality by increasing the quantity or supplementing with traditional, in-depth, exploratory cognitive interviews to provide deeper insight into the more complicated aspects of the question evaluation. Researchers should carefully consider demographic characteristics and experiences of participants and use traditional and crowdsourcing methods in concert to optimize data quality and quantity.

Although self-administered cognitive interviewing methodology has not, thus far, incorporated unscripted probes, web survey technology makes adding a full spectrum of tailored probes based on a response possible. For example, researchers could develop a web instrument that displays a probe when the respondent answers a question too quickly or uses only a few words. Past research suggests that, although the wording and format of probes must be carefully developed, asking such follow-up questions on web surveys can be an effective method to collect in-depth information about the respondent’s response process (Behr, Bandilla, Kaczmirek, & Braun, 2012; Behr, Kaczmirek, Bandilla, & Braun, 2012). The use of automated probes for web self-administered cognitive interview is another topic worthy of future investigation.

Advantages of crowdsourcing are that it can be fast, inexpensive, and allow for greater geographic dispersion; can sometimes result in a more experienced audience (e.g., TryMyUI); and can be used to target-specific groups (e.g., Facebook). Disadvantages include the lack of unscripted or follow-up probes and absence of an interviewer. Another disadvantage is potential panel bias. If panel members are different from the population in a systematic way, this could lead to bias. In particular, researchers have raised concerns with “professional respondents” in Mechanical Turk who may differ from the general population because of the large number of studies in which they participate (Marder, 2015). There may be similar concerns with laboratory participants, with some participating in many studies. Quantifying and determining the effect of “professional respondents” on the outcomes of cognitive interview research is another area ripe for future investigation.

Considering these advantages and disadvantages, crowdsourcing seems to be a promising addition to traditional pretesting methodologies. Figure 1 highlights three potential project examples. For projects that have the time and budget for traditional cognitive interviews, crowdsourcing can be used as a first step in a pretesting process. Researchers can integrate a variety of comprehension, paraphrasing, confidence, recall, specific, and general probes in a web survey instrument (Willis,
The breadth of respondents in a crowdsourcing process, numbering in the hundreds or even thousands, allows researchers to identify potentially glaring issues with their questionnaire before they move into a cognitive interview that focuses on depth. If the project has additional time and budget, then a second round of crowdsourcing could be implemented after the cognitive interviews, to serve as a final pretest. For projects that do not have the time or the budget for traditional cognitive interviews, crowdsourcing may serve as a quick pretesting method, allowing researchers to identify and resolve some of the issues surrounding their questionnaire that otherwise might not be found.

Moving forward, further evaluation of the use of crowdsourcing as a recruitment method and the effectiveness of self-administered surveys to collect cognitive interviewing type information is needed. In addition, more research is needed to investigate other crowdsourcing alternatives (e.g., Google Consumer Surveys, Promoted Tweets, online survey panels) and to determine more precisely where and when the need for an interviewer is necessary.

Declaration of Conflicting Interests
The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Opinions expressed in this paper are those of the authors and do not reflect official policy of the U.S. Bureau of Labor Statistics.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Cognitive laboratory and TryMyUI interviews were funded by the U.S. Bureau of Labor Statistics. Mechanical Turk and Facebook interviews were funded by RTI International. The authors received no additional financial support for the research and/or authorship of this article.

Notes
1. Laboratory and TryMyUI participants completed additional substantive topics not discussed in this study. Mechanical Turk and Facebook participants only completed the topics described here.
2. For each of these questions, researchers identified the response strategy as most likely leading to an accurate response. Other strategies were likely to lead to an inconsistent response if it were used by all respondents.

References
Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., . . . Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, M. (2012). Cognitive probes in Web surveys: On the effect of different text box size and probing exposure on response quality. *Social Science Computer Review*, 32, 524-533.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: Which factors have an impact on the quality of responses? *Social Science Computer Review*, 32, 487-498.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Weihe, E. N. (2011). The viability of Crowdsourcing for survey research. *Behavior Research Methods*, 43, 800-813.

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly*, 75, 636-658.

Christenson, D. P., & Glick, D. M. (2013). Crowdsourcing panel studies and real-time experiments in MTurk. *The Political Methodologist*, 20, 27-33.

Chunara, R., Chhaya, V., Bane, S., Mekaru, S. R., Chan, E. H., Freifeld, C. C., & Brownstein, J. S. (2012). Online reporting for malaria surveillance using micro-monetary incentives, in Urban India 2010-2011. *Malaria Journal*, 11, 43.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized survey questionnaire*. Newbury Park, CA: Sage.
Edgar, J. (2012, May). *Cognitive interviews without the cognitive interviewer?* Paper read at 67th Annual Conference of the American Association for Public Opinion Research, Orlando, FL.

Edgar, J. (2013, May). *Self-administered cognitive interviewing.* Paper read at 68th Annual Conference of the American Association for Public Opinion Research, Boston, MA.

Erdman, C., & Bates, N. A. (2013, November). *The U.S. Census Bureau Mail Return Rate Challenge: Crowdsourcing to develop a hard-to-count score.* Paper read at Federal Committee on Statistical Methodology Research Conference, Washington, DC.

Fowler, S., Willis, G., Moser, R., Ferrer, R., & Berrigan, D. (2015). *Use of Amazon MTurk online marketplace for questionnaire testing and experimental analysis of survey features.* Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference, Washington, DC. Retrieved from http://sites.usa.gov/fcsm/files/2016/03/C2_Fowler_2015FCSM.pdf

Hill, C. A., Dean, E. F., & Murphy, J. J. (2013). *Social media, sociality, and survey research.* Hoboken, NJ: John Wiley.

Keating, M. D., & Furbeg, R. D. (2014). *A methodological framework for Crowdsourcing in research.* Proceedings of the 2013 Federal Committee on Statistical Methodology Research Conference, Washington, DC. Retrieved from https://fcsm.sites.usa.gov/files/2014/05/H1_Keating_2013FCSM.pdf

Keating, M. D., Rhodes, B. B., & Richards, A. K. (2013, November). *Applying Crowdsourcing methods in social science research.* Paper read at Federal CASIC Workshops, Washington, DC.

Kim, A. E., Lieberman, A. J., & Dench, D. L. (2013). *Case study comparing data collected via Crowdsourcing vs. Trained Data Collectors.* Proceedings of the 2013 Federal Committee on Statistical Methodology Research Conference, Washington, DC. Retrieved from https://fcsm.sites.usa.gov/files/2014/05/H1_Kim_2013FCSM.pdf

King, S. (2009). *Using social media and crowd-sourcing for quick and simple market research.* Retrieved from http://money.usnews.com/money/blogs/outside-voices-small-business/2009/01/27/using-social-media-and-crowd-sourcing-for-quick-and-simple-market-research

Marder, J. (2015). *The Internet’s hidden science factory.* Retrieved from http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory

Meitinger, K., & Behr, D. (2016). *Comparing cognitive interviewing and online probing: Do they find similar results?* *Field Methods.* Published online before print January 31, 2016, doi:10.1177/1525822X15625866

Miller, K., Chepp, V., Wilson, S., & Padilla, J. L. (2014). *Cognitive interviewing methodology.* Hoboken, NJ: John Wiley.

Murphy, J. J., Sha, M., Flanigan, T. S., Dean, E. F., Morton, J. E., Snodgrass, J. A., & Ruppenkamp, J. W. (2007, November). *Using craigslist to recruit cognitive interview respondents.* Paper read at Midwest Association for Public Opinion Research, Chicago, IL.

Paolacci, G., & Chandler, J. (2014). *Inside the Turk: Understanding Mechanical Turk as a participant pool.* *Current Directions in Psychological Science, 23,* 184-188.

Pew Research Center. (2015). *Demographics of key social networking platforms.* Retrieved from http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

Willis, G. B. (1999, August). *Cognitive interviewing: A “how to” guide.* Paper read at Annual Meeting of the American Statistical Association, at Research Triangle Park, Research Triangle Institute, NC.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* London, England: Sage.

**Author Biographies**

Jennifer Edgar is director of the Behavioral Science Research Center in the Office of Survey Methods Research at the U.S. Bureau of Labor Statistics. Her research focuses on context effects, qualitative research methods, cognitive sources of measurement error, questionnaire design, and interviewer training and evaluation.

Joe Murphy is a senior survey methodologist at RTI International. His research focuses on new technologies to improve data quality, timeliness, and cost efficiency.

Michael Keating is a survey research scientist at RTI International. His research focuses on new technologies for data collection and innovation management.