

Crowdsourcing in the Cognitive Interviewing Process

Joe Murphy^a, Michael Keating^b, Jennifer Edgar^c

^aRTI International, Chicago, IL, jmurphy@rti.org

^bRTI International, Research Triangle Park, NC, mkeating@rti.org

^cU.S. Bureau of Labor Statistics¹, Washington, DC, edgar.jennifer@bls.gov

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

Crowdsourcing holds promise as an alternative or supplement to traditional cognitive interviewing methods. Faced with the challenge of assessing the extent to which a question or concept is understood before finalizing a survey instrument, researchers typically recruit subjects for an in-person administration of draft items using think-aloud methods to capture feedback. This can provide deep understanding of individuals' reactions to a question or set of questions, but is limited by the number of interviews that can be completed in a short amount of time and at a given cost. Cognitive interview participants are typically not selected based on probability methods and, beyond demographics, there is usually not an understanding of how they may differ from those in the population of interest for the study. Crowdsourcing allows for the option of recruiting many participants quickly and cheaply. Participants can be provided a question or set of questions to consider and to which they can react. Using an existing crowdsourcing platform, these interviews can be self-administered over the web to provide reactions to questions as quickly as in one afternoon. We discuss results of a pilot test of this method of testing survey items using crowdsourcing including advantages and disadvantages as compared to traditional cognitive interviewing methods.

Introduction

Crowdsourcing has been defined as the act of “tapping into the collective intelligence of the public to complete a task” (King 2009). This simple definition could apply to any number of research activities and one may argue it even applies to survey research. By completing a survey or conducting a qualitative assessment such as cognitive interviewing, we are asking individuals to supply a set of information to be used for some analytic purpose. Crowdsourcing is generally seen as a new phenomenon, brought about by the networked nature of the internet and web platforms set up to accomplish discrete tasks by a pool of willing individuals at a low cost and in a short amount of time (Murphy, Hill & Dean, 2013). In fact, many researchers have begun using crowdsourcing platforms for data collection (Keating, Rhodes, & Richards, 2013). Behrend et al. (2011) compared the quality of responses given from a sample compiled from volunteers on Mechanical Turk to a more common and traditional sample of university students. They found better internal consistency with Mechanical Turk, but more of a social desirability effect with responses. They found few differences between the Mechanical Turk and university samples with regards to completion time or word count and more diversity among Mechanical Turk respondents. Importantly, they found Mechanical Turk to be much more efficient than the university sample, with hundreds of participants recruited in less than 48 hours requiring less than one dollar a piece in reimbursement for completion of a 30-minute survey.

Crowdsourcing has four distinctive features: broad reach, a motivated crowd, participants well suited to complete the task, and infrastructure to facilitate the task completion. In probability-based surveys, each member of the sample frame has a known and non-zero chance of selection into the study sample. This makes general population sampling via crowdsourcing infeasible at this time because there is no crowdsourcing platform that covers the general population well and in a way to facilitate probability based sampling. However, for pretesting activities

¹ Disclaimer: Opinions expressed in this paper are those of the authors and do not reflect official policy of the U.S. Bureau of Labor Statistics.

such as cognitive interviewing where representative point estimates, for example, are not the ultimate goal of research, crowdsourcing may offer an attractive alternative to traditional methods for several reasons.

Cognitive interviews are used in the survey development process to evaluate sources of response error in survey questionnaires. The focus is on the cognitive processes that respondents use to answer questions. (Willis, 1999). Traditionally, cognitive interviews have been conducted in survey laboratory settings with members of the target population for a survey who have been recruited through a variety of methods including posted flyers, word-of-mouth, newspaper ads, and more recently ads on electronic forums such as Craigslist (Murphy et al., 2007). Recruited participants come into the lab to complete a face-to-face interview using think-aloud and similar methods to obtain reactions to draft survey questions and explore potential sources of measurement error in the instrument. These methods offer many benefits but have a fair number of drawbacks as well. For instance, those who answer cognitive interviewing ads and take the effort to travel to a facility to complete the interview are highly motivated, and may differ substantially from the target population. Also, the activity is typically limited to a local area and to those with the time to travel in and complete the task. The extent to which these individuals represent the target population may be quite limited on dimensions like geography, socio-economic status, and so on. Once in the lab, interviewers are able to delve into detail on individual aspects of the survey items to learn what sources of confusion or cognitive difficulty may be present, and can tailor the questions to the participant's responses. Finally, securing participants and providing incentive for their participation can require a level of resources that may limit the number of total participants, which increases the chance that a rare but important perspective is missed. In other words, traditional methods allow a researcher to go "deep" with a small set of participants and understand the nuance of some survey items, but limit the ability to go "wide" to gain a large number of viewpoints.

These limitations, in conjunction with the resources required to recruit and interview participants in the lab setting lead to exploratory research using an online panel to conduct self-administered interviews (Edgar, 2012, 2013). This work found that the same type of information from a cognitive interview could be collected without the aid of an interviewer. Responding to a series of open and closed questions, participants were able to explain their response process, their interpretation of key terms and complete tasks that demonstrated their comprehension of the question. Audio recording of think-alouds and text boxes captured the data which then was analyzed using the same qualitative methods used to analyze cognitive interview data. Efficiencies in resources required to recruit and collect the data were realized, and the demographic characteristics of the web panelists were generally comparable to the lab 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 to traditional cognitive interviews for some questions tested (Edgar, 2012).

In this paper, we consider a few newer options for recruiting participants and conducting cognitive interviews using systems that benefit from the "power of the crowd." The first alternative is a web-based usability panel called TryMyUI that was designed to provide website developers with a base of reviewers who could comment on different aspects of design and functionality for improvement. Participants in that panel complete a short self-administered protocol by speaking their responses to cognitive interview prompts, which are recorded and later reviewed by the researcher. The second is recruitment using a popular crowdsourcing platform, Amazon Mechanical Turk. Workers sign up with Mechanical Turk to complete very short tasks and the researcher can purchase access to available workers to complete a study in just hours. The third is recruitment using the social media platform Facebook. Ads can be purchased and targeted based on user demographics and "likes" and those who agree to participate complete a web survey. A more complete description of each crowdsourcing platform evaluated follows.

TryMyUI is an online, remote usability testing service. The service coordinates recruitment and payment of participants, administration of tasks, and collection and storage of audio recording of the participant's voice and video recording of the participant's computer screen. TryMyUI has a volunteer participant panel with a range of demographic characteristics and backgrounds. One strength of this service is the experience of the panel members, each of whom is rated by the researcher requesting the task after the study is complete. Only panel members receiving high ratings are eligible to complete future tasks. Since the members typically evaluate websites and other online material, they are used to reacting to stimuli and expressing their opinions out loud, experience which has the potential to help them provide excellent cognitive-interviewing type data.

Amazon Mechanical Turk is one of the World's the most popular crowdsourcing platforms. Mechanical Turk allows for requesters to post "Human Intelligence Tasks" (HITs) and for workers to complete these HITs for

payment. At any point in time, hundreds of thousands of HITS are available on the website. In general, the site focuses on HITS that are “micro-tasks” and may take a person only a few seconds or minutes to complete. Mechanical Turk is increasingly being used by researchers to recruit respondents and to collect survey response data (Chunara et al., 2012; Christenson & Glick, 2013). The speed of the platform is very fast and the cost of data collection is very low, making the platform very appealing for researchers with lean budgets. Initial research has shown that it is possible to recruit hundreds of respondents within a few days. Behrend et al. paid their respondents \$0.80 to complete a 30 minute survey (Behrend et al., 2011). When comparing this to the cost of a traditional face-to-face cognitive interview that paid \$25, this is a 96.8% cost savings. There are over 500,000 workers on Mechanical Turk and most reside in the United States. In a 2010 study, Ipeirotis showed that the demographic makeup of Mechanical Turk was a bit different from the general United States population. Specifically, Mechanical Turk workers were younger, better educated, but had lower income levels (Ipeirotis, 2010). Considering that cognitive interview samples are not typically expected to represent the general population, and that data can be collected at rapid speeds and low costs on the platform, Mechanical Turk offers a very appealing and intriguing value proposition to collect feedback on a questionnaire.

Facebook is a popular social networking site with over 1.1 billion active users worldwide, including the majority of U.S. adults (Pew, 2013). One feature of the site is ability to purchase advertising space targeted to specific demographic groups or those with particular interests or “likes.” In our case, we invited participants using these ads to complete a short survey in exchange for a small incentive. While the typical Facebook user is not likely looking for the opportunity to complete surveys or related activities when visiting the site, the vast user base and relative low cost of advertisement makes it a promising and potentially powerful platform for this application of crowdsourcing.

In this paper, we consider crowdsourcing methods as alternatives to traditional cognitive interview recruitment and conduct. We consider how these alternative methods compare to traditional in terms of the cost of recruitment, the speed with which one can recruit participants, the ability to target and recruit from the survey’s population of interest, and the quality of information obtained from those recruited and conducted via these alternative methods. Ultimately, we are interested in learning what conclusions are drawn from the different methods and how they compare.

Methods

Our experimental design included administering a common subset of cognitive interview items in each of the modes described previously, however not all modes included all questions. For instance, the “micro-task” setup of Mechanical Turk necessitated limiting the interviews to five minutes and administering questions in modules that were each presented to only a subset of participants. The design involved five groups of participants, detailed in **Table 1**.

Table 1. Recruitment and Interviewing Methods

Group	Recruitment	Interview mode	Number of participants	Average length of interviews (minutes)
1	traditional	face-to-face	71	25
2	traditional	web-based self-interview with typed responses	18	20
3	crowdsourcing: TryMyUI	web-based self-interview with typed and audio responses	44	15
4	crowdsourcing: Mechanical Turk	web-based self-interview with typed responses	1,020	5
5	crowdsourcing: Facebook	web-based self-interview with typed responses	60	10

An overview of the recruitment and interview methods for each group follows.

Groups 1 and 2: Traditional recruitment, lab setting

We conducted traditional cognitive interviews in a laboratory setting; participants came to the researcher at a prescheduled time. Potential participants were originally identified through a variety of sources, including ads in newspapers, online bulletin boards, fliers in community locations and word of mouth and then entered into a database for future selection. This was the source of participants for this study; we identified 71 participants from the database, screened based on demographic characteristics to aim for representativeness across key characteristics, and invited them to the lab. Participants had to be able to come to the testing site during the times specified by the researcher. Prior to the interview, we called participants to confirm their appointment. Although exact time spent recruiting was not captured, it is estimated that the recruiter spent about an hour identifying, screening, scheduling and reminding each participant.

For the cognitive interviews, we followed traditional procedures. We used a scripted protocol to conduct the interview. In general, participants first answered the survey question and were then asked to explain how they arrived at their answer. The interviewer asked follow-up questions as necessary to obtain as complete an understanding as possible of their response process. We used both open ended (e.g. please give examples) and close ended (e.g. are the following items included in the category) questions. We audio recorded the interviews and took notes, which were entered into a spreadsheet following the interview. Most interviews lasted approximately 20 to 30 minutes.

In addition, a small subset of participants recruited using traditional methods were asked to complete a web-based self-administered version of the instrument. The purpose was to discern whether any differences in feedback or outcomes could be detected using a self-administered mode vs. face-to-face cognitive interviewing.

Group 3: TryMyUI

Recruiting from the TryMyUI panel is based on groups defined by the researcher (e.g. 5 males with high school education). As eligible participants come to the site, they may accept the task, which is offered as a “first come, first serve” process until the quotas have been filled. For this study, we requested 45 participants to complete a web survey which was hosted on SurveyMonkey. Follow-up questions, mirroring those used in the traditional cognitive interviews, followed each survey question. In TryMyUI, the open-ended follow up questions (e.g. what did you think of?) were completed as a think-aloud, which was captured via audio recording. Survey Monkey captured the close-ended follow-up questions. TryMyUI limits tasks to 20 minutes; most participants completed the task in less than 20 minutes.

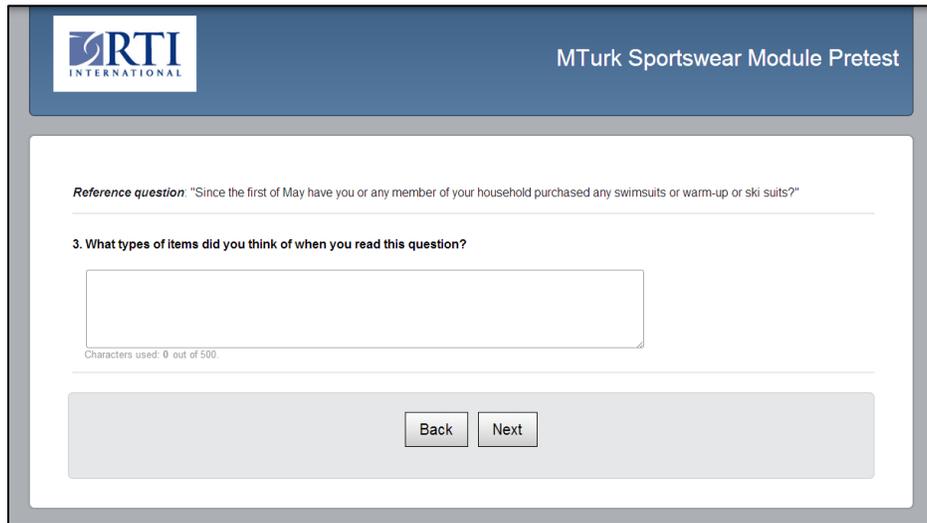
Group 4: Mechanical Turk

We posted HITs on Mechanical Turk that paid people \$0.75 to complete the survey. The HIT linked workers to a web survey that contained the survey questions, the cognitive interviewing protocol, and a set of demographic questions. As shown in **Figure 1**, we inserted the reference question above cognitive interview probes to aid respondents and prevent recall challenges. At the conclusion of the survey, workers were given a randomly generated number PIN code. Workers entered this PIN code in their HIT template and submitted it for payment. We collected 288 responses over a period of two days.

Group 5: Facebook

We recruited participants on Facebook using targeted ads. These ads can be purchased and typically appear on the right hand side of a Facebook users' screen. When purchasing an ad, the buyer has the option to target by demographic information such as location, age, gender, language, and education (assuming Facebook users enter these details in their profiles) and interests (a.k.a. “likes”). The ad placed stated that one would receive a \$5 electronic gift card for completing a ten minute survey. Given that we were interested in a general population for this cognitive interview project, we began by limiting the Facebook target only to those living in the United States, 18 years of age or older, and who spoke English. We experimented with several ad types and found the best success with those offering a \$5 donation to the Red Cross rather than a \$5 gift card (Murphy, 2013).

Figure 1. Screenshot of the Desktop Version of the Web Survey with a Reference Question

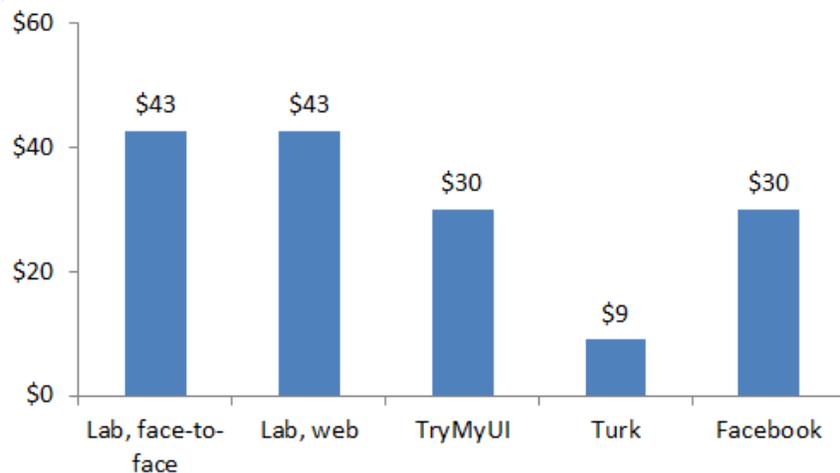


Results

We examined several results by mode, including the costs of recruitment and conducting the interviews, the geographic and demographic distributions obtained, the amount of information shared by mode, the amount of information that was relevant, an assessment of accurate comprehension, and insight into the thought process.

For cost, the one component allowing an equivalent comparison by mode was incentive costs, which made up the majority of total costs. Incentives required were highest for Groups 1 and 2—the in-person treatments—and lower for Groups 3-5—the crowdsourcing platforms. **Figure 2** provides a comparison of incentive costs per respondent hour by group.

Figure 2. Incentive Cost per Respondent Hour

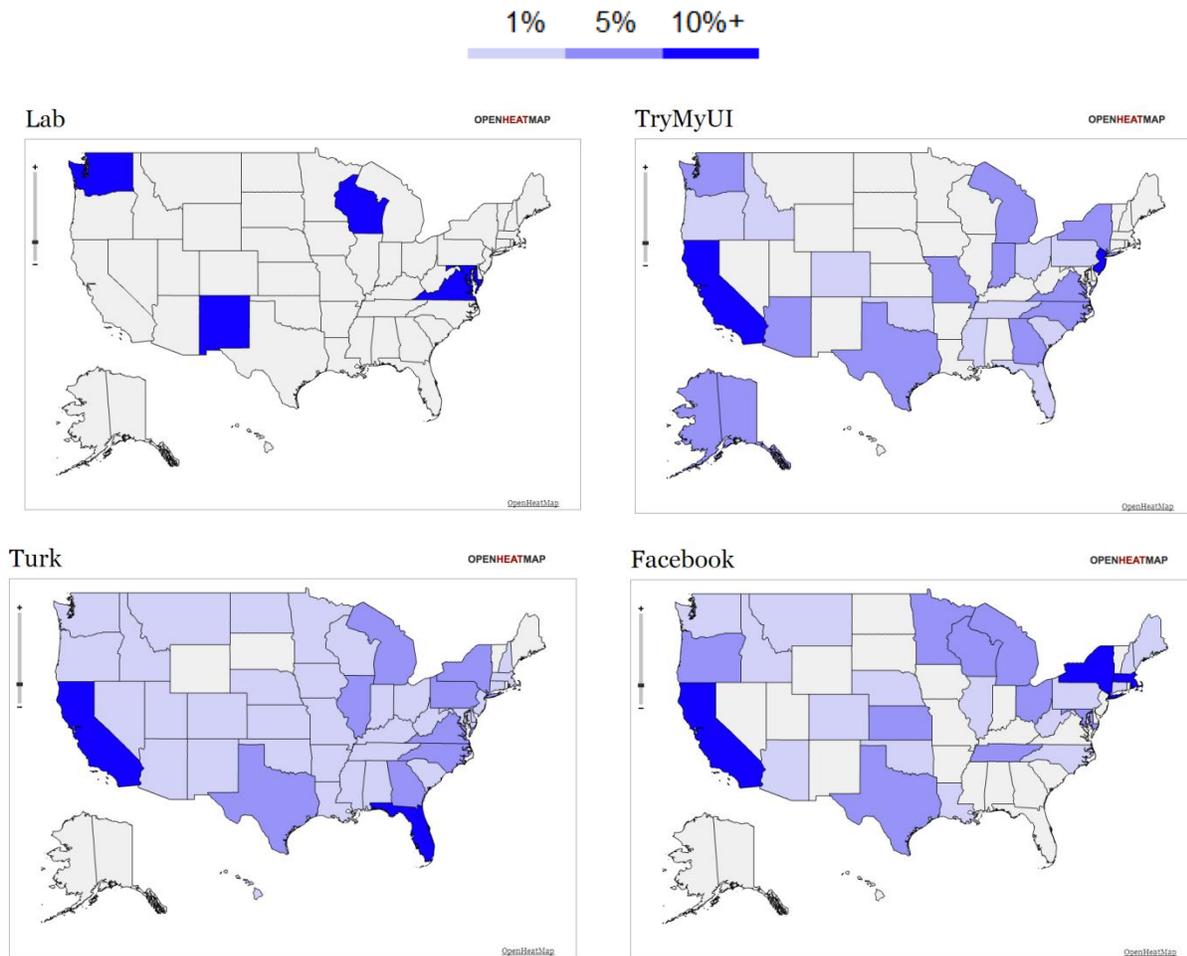


TryMyUI and Facebook had about the same incentive totals at \$30 per respondent hour. Mechanical Turk was much less expensive at \$9 an hour in respondent incentives. Other costs were more difficult to compare by mode, including the screening and interviewing staff time for the lab cases, a 10% fee charged by Amazon for the Mechanical Turk service, and the cost of the Facebook ads. Those ads added about \$300 in cost to Group 5. The

ads used a pay-per-click campaign where Facebook was paid each time someone clicked on our ad, regardless of whether they completed the interview. Only about 1 in 10 who clicked on the ad ended up completing the interview.

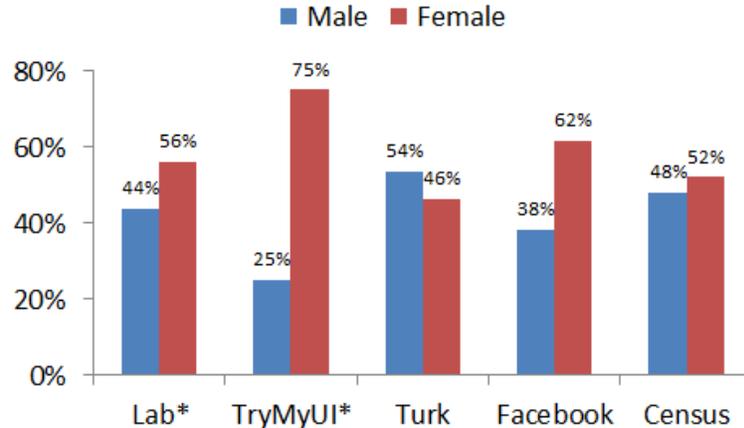
Regarding geography, the lab cases all came from 4 local regions, by design. TryMyUI had a much more even demographic spread across the U.S., though there appeared to be some concentration in the western U.S. Facebook also had a good representation of different states and regions. Mechanical Turk, with over 1,000 completes, provided a wide geographic coverage with respondents in almost every state. The lab interviews in this study were unique from most in that it was conducted in four locations; typically lab interviews are conducted at the single location of the researcher. **Figure 3** present geographic reach by state and mode.

Figure 3. Percent of Respondents in Each State by Mode



Looking at demographics, we compared our results to U.S. Census benchmarks to see how well these samples matched the general population age 18 and older. The lab and TryMyUI modes employed quota sampling to explicitly try to match the benchmark as closely as possible. On gender, the demographics of the lab groups came close to the Census benchmark, but TryMyUI completes were more female than male by a rate of 3 to 1. The Facebook recruitment also resulted in more females and Turk included slightly more males than females. **Figure 4** presents the gender distribution by mode with Census figures for comparison.

Figure 4. Gender Distribution by Mode



*Lab and TryMyUI recruitment used quota sampling

By age group, the lab participants match the Census benchmark rather closely. Facebook participants, not surprisingly, included more 18-24 year olds. Mechanical Turk included more 25-34 year olds and TryMyUI included more 35-44 year olds. It was interesting to observe that the different crowdsourcing platforms resulted in overrepresentation of different groups, suggesting one mode might be more efficient than others based on the population of interest. **Figure 5** presents age distributions by mode.

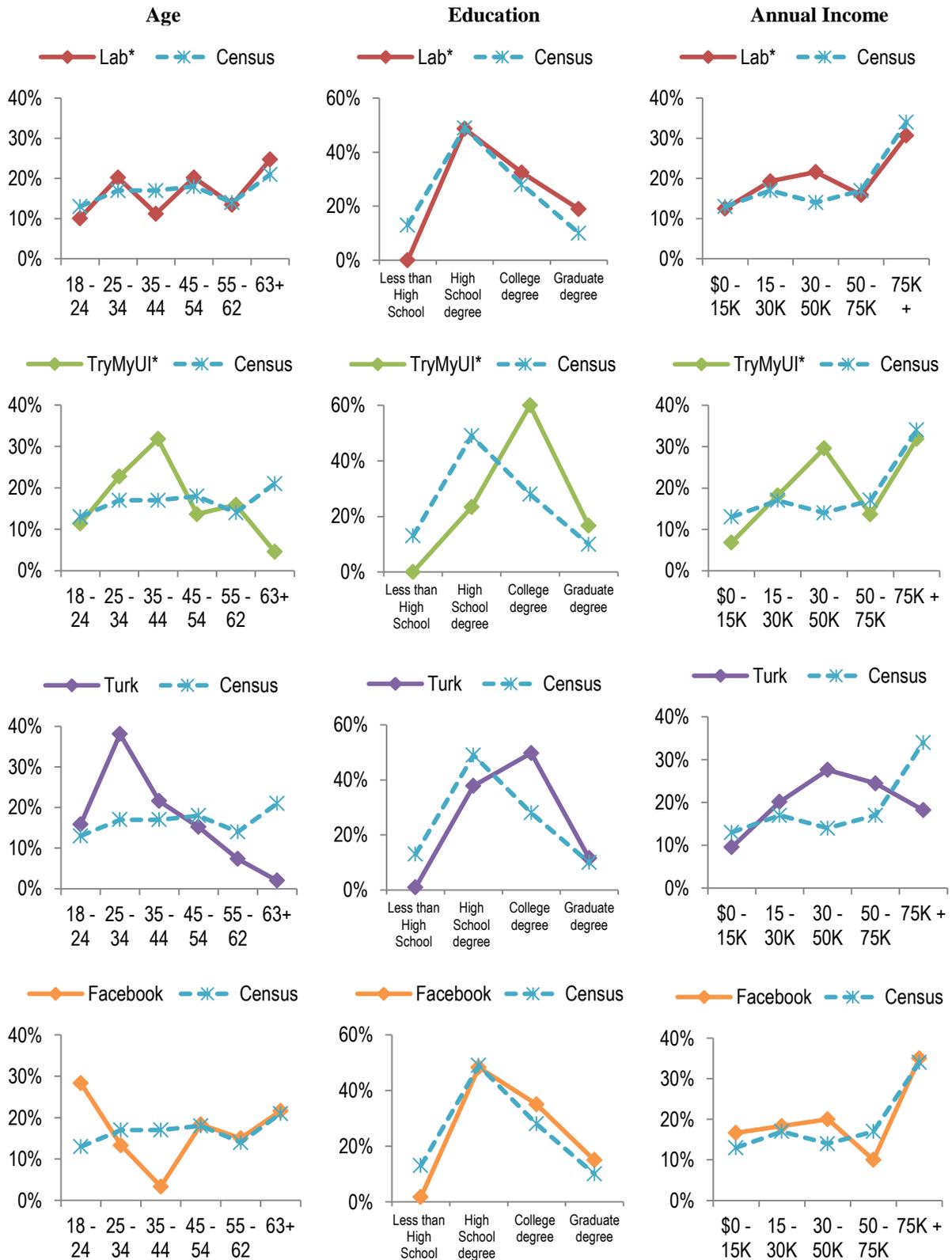
Concerning level of education, none of the modes evaluated generated many, if any, participants with less than a high school education. TryMyUI and Turk skewed toward the more educated, where the lab and Facebook samples matched the Census benchmark rather closely. **Figure 5** presents the education distribution of participants by mode.

By annual household income level, the lab and Facebook again matched the Census benchmark pretty well overall. All modes included proportionately more participants in the 30-50 thousand dollar income range than the benchmark. **Figure 5** presents the income distribution of participants by mode.

Regarding the actual feedback we obtained from participants about the survey items, we first looked at the amount of information shared for the item “Since the first of May have you or any member of your household purchased any swimsuits or warm-up or ski suits?” We asked what types of items people thought of when they answered the question and counted the words in their responses to obtain a comparison in the amount of information obtained from participants in each mode. We saw the highest amount of words elicited from the TryMyUI participants and fewest from the Facebook participants. While the modes are difficult to compare here given the mix of administration types, a pattern starts to emerge with the panel-based crowdsourcing platforms appearing to have a more engaged set of participants than in Facebook, where we were engaging with people who may have never completed this type of task before and, in that way, may be more like the population we would ultimately be targeting in a general population survey. **Table 2** presents the median number of words provided by participants in a follow-up probe about reporting expenses on sportswear.

We were also interested how much of the information provided was actually relevant and helpful in revealing the participants’ thoughts on the subject and providing actionable feedback on the survey. We coded responses using the NVivo software package and created a variable indicating relevant vs. irrelevant responses. Relevant examples included statements like “one piece swimwear, snow pants, or tennis shoes” and irrelevant examples include statements like “how it would look on stage, I didn’t buy any, and none” which were not informative for our purposes. As shown in **Table 2**, Turk and TryMyUI gave us very relevant feedback overall. The lab rate of relevant information was lower, but for the face-to-face lab cases, there was the opportunity to probe respondents to provide more information, so this would theoretically rise to 100%. Facebook had the lowest level of relevant information provided, but still the great majority of responses were on task.

Figure 5. Demographic Distributions by Mode Compared to Census Benchmark



*Lab and TryMyUI recruitment used quota sampling

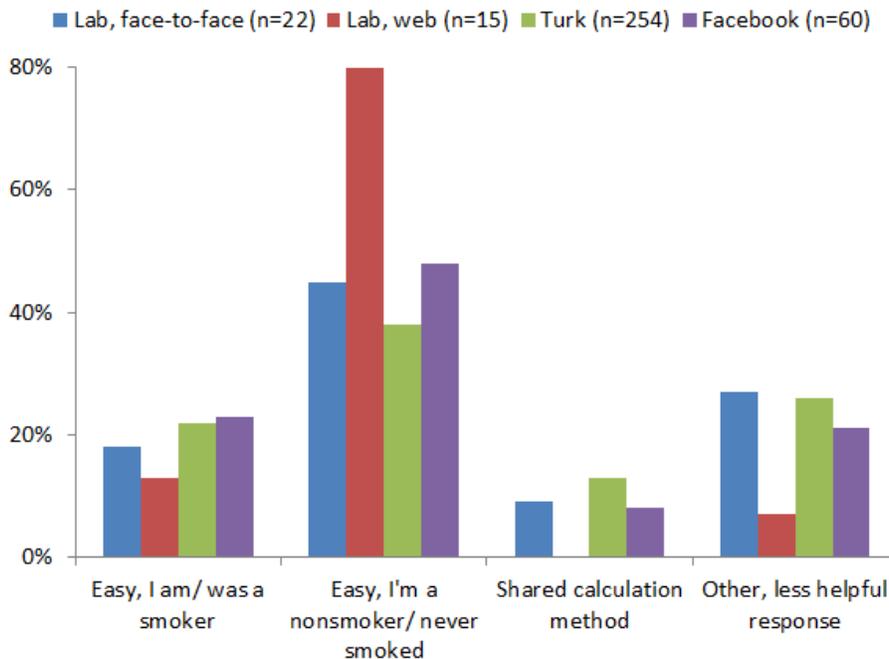
Another piece of information we evaluated regarded comprehension of the task and the items in the cognitive interview. We included a check in to make sure participants understood the items that would be included under the sportswear question. These were designed to be easy and just to make sure they were paying attention and giving good quality responses. We coded accuracy to these items as the percent coded correctly as include vs. exclude. As shown in **Table 2**, we see a similar pattern here as with the previous measures. Turk and TryMyUI had the highest rates of comprehension on these items and Facebook, while not horribly low, did have the lowest comprehension.

Table 2. Follow-ups to Swimsuits or Warm-up or Ski Suits by Mode

	Lab, face-to- face	Lab, web	TryMyUI	Turk	Facebook
Median number of words in response	7	6	9	8	4
Percent of information that was relevant	87	89	85	98	86
Percent accurately comprehending question	80	78	81	84	70

We were also interested to compare what insights into the thought process we would obtain in the different modes. We included a standard lifetime smoking item “Have you smoked at least 100 cigarettes in your entire life?” in all modes except TryMyUI. We followed this up with the probe “If you were asked this survey question, how would you mentally calculate the answer?” and coded responses into whether participants indicated this was easy to answer because they are or were a smoker or nonsmoker, whether it required calculation and the respondent gave insight into the calculation process, or whether they provided another, less helpful response. The sample size for the lab group completing on the web was very small, so we believe this may explain their extent of mismatch with the other modes. As shown in **Figure 6**, overall, we see generally equivalent distributions to this item by mode. Comparing the bars under “shared calculation method” and “other” it appears there are few respondents by mode that actually provided good insight their calculation process, regardless of mode.

Figure 6. Responses to “How Would You Mentally Calculate?” Whether Ever Smoked 100 Cigarettes



Discussion

This study provided many insights into some of the relative advantages and disadvantages of crowdsourcing methods for cognitive interviewing and some evidence for specific platforms. The major advantage of the lab and face-to-face administration is that we can use spontaneous probing on specific responses. However, recruitment is restricted to the local area and it is a time intensive and expensive process. The lab is probably best for items or constructs requiring in-depth exploration of the response process.

The advantages of TryMyUI were the rapid data collection, geographic dispersion, and motivated and knowledgeable participants. However, we could not use spontaneous probing, did not get many lower education participants, and the gender balance was skewed towards females. TryMyUI respondents were very experienced at giving technical feedback and thinking aloud. This platform may be best for getting straightforward verbal think-alouds quickly from more participants than one could in the lab.

Mechanical Turk was very quick for collecting data. Each of the HITs resulting in over 250 respondents were completed in just a couple days and provided good geographic coverage. It was relatively very inexpensive, with an incentive rate of about \$9 an hour. However, like with the other crowdsourcing methods, we could not probe spontaneously and HITs are typically very short and may limit us to just a few items per participant. Respondents appeared experienced in giving this type of feedback but perhaps not to the extent of the TryMyUI group. Turk may be best when large samples, rapid turnaround, and limited resources are drivers. Turk participants seemed to fit a niche somewhere between novice respondents and an expert review group, so it may be good resource in situations where that level of feedback is needed.

Finally, Facebook offered fast data collection and good geographic coverage among a base that was not necessarily looking to participate in cognitive interviews, and perhaps that is a benefit in some cases when there is most interest in how the real population will react to a survey. The recruitment strategy can make a big difference, so, like with a survey, that strategy needs to be thought through and be informed by what has worked well in the past. Like with the other self-administered modes, we did not have the option of spontaneous probing with Facebook, which was a disadvantage.

Moving forward, we hope to further evaluate each mode and crowdsourcing alternative for optimal design. There may be a way to address the limitation of spontaneous probes with the crowdsourcing methods by making an interviewer available on call by telephone or Skype, or more advanced automated methods for programming in probes based on natural language processing. We would like to look into other crowdsourcing alternatives like Google Consumer Surveys, promoted Tweets, and other panels that are like TryMyUI and replicate a single set of questions with as few differences by mode as possible. We attempted to do that here but since this study was done in stages, there were some limitations to accomplishing that. Finally, figuring out just where and when the presence of the interviewer pays off is important. If self-administered cognitive interviews can provide the same information as interviewer-administered in certain scenarios, we should determine what scenarios those are and then use that information to tailor our designs.

References

- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43, 800-813.
- Chunara et al. (2012). Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010-2011. *Malaria Journal* 2012 11:43.
- Christenson, D.P. & Glick, D. (2013). Crowdsourcing Panel Studies and Real-Time Experiments in MTurk, *The Political Methodologist*. Vol 20, No. 2.
- Edgar, J. (2013). Self-Administered Cognitive Interviewing. Presented at the 68th Annual Conference of the American Association for Public Opinion Research, Boston, MA.

Edgar, J. (2012) Cognitive Interviews without the Cognitive Interviewer? Presented at the 67th Annual Conference of the American Association for Public Opinion Research, Orlando, FL.

Ipeirotis, P. (2010). Demographics of Mechanical Turk. *Center for Digital Economy*, Research Working Papers, 10.

Keating, M., Rhodes, B., & Richards, A. (2013). Crowdsourcing: A Flexible Method for Innovation, Data Collection, and Analysis in Social Science Research. In *Social Media, Sociality, and Survey Research*, (eds. Hill, C.A., Dean, E., & Murphy, J.). New York: Wiley.

King, S. (2009). Using Social Media and Crowd-Sourcing for Quick and Simple Market Research. <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>

Murphy, J. (2013). Altruism: Alive and Well on Facebook? *SurveyPost*. <http://bit.ly/Hjyrwl>

Murphy, J., Hill, C.A., & Dean, E. (2013). Social Media, Sociality, and Survey Research. In *Social Media, Sociality, and Survey Research*, (eds. Hill, C.A., Dean, E., & Murphy, J.). New York: Wiley.

Murphy, J., Sha, M., Flanigan, T. S., Dean, E. F., Morton, J. E., Snodgrass, J. A., & Ruppenkamp, J. W. (2007). Using Craigslist to recruit cognitive interview respondents. Presented at the Annual Meeting of the Midwest Association for Public Opinion Research, Chicago, IL.

Pew Research Center (2013). Social Networking Use. <http://www.pewresearch.org/data-trend/media-and-technology/social-networking-use/>

Willis, G. B. (1999). Cognitive interviewing: A "How To" guide. Annual Meeting of the American Statistical Association. Research Triangle Park, NC: Research Triangle Institute.