Man versus Machine: Evaluating IVR versus a Live Operator for Phone Surveys in India

Dipanjan Chakraborty[1]*, Indrani Medhi[2], Edward Cutrell[2], and William Thies[2]

[1]IIT Delhi
dipanjan@cse.iitd.ac.in

[2]Microsoft Research India
{indranim,cutrell,thies}@microsoft.com

ABSTRACT

Many organizations in the developing world need to conduct phone surveys to collect data from low-income respondents. Such organizations generally have two options: employ a live operator, or utilize interactive voice response (IVR). Despite the relevance of this question, we are unaware of any work that rigorously compares the accuracy, speed, and cost of an IVR survey relative to a live operator.

In this paper, we address these questions by giving two identical interviews – one using IVR, and one using a live operator – to 31 low-income job seekers in India. The IVR interview included a brief introduction by a live operator, to provide context for the call. Out of the 20 people who completed both surveys, we found that IVR incurs a 4.0% error rate (95% C.I. 2.5% – 6.1%) and requires 2.5 times longer for users. We summarize our experience as a set of recommendations for practitioners in this space.

1. INTRODUCTION

The rapid increase in teledensity in many developing countries has made it feasible to conduct large-scale surveys over the phone rather than relying on face-to-face interviews [32]. This has the potential for organizations to increase their reach while decreasing costs. At the same time, phone surveys come with their own set of challenges. They need to retain high data quality, even in the midst of unreliable network connections and varying attention levels on the part of respondents. Also, they often require a dedicated staff of call center operators to track and administer the calls. Employment of such operators can quickly become a financial bottleneck that prevents a survey from scaling up to reach the full target population.

Interactive voice response (IVR) systems represent a potential alternative to the expense of a live operator. By using a computer to administer the survey, employees can spend time on less repetitive and more important tasks. IVR also offers flexibility in scheduling calls: an automated system can make many parallel calls during peak hours (say, 6-8pm) while live operators are typically employed throughout the day. However, IVR has drawbacks as well. It requires all users to follow a fixed script, which can be slow, tedious, and impersonal relative to a live conversation. It requires considerable technical setup, which is beyond the reach of many organizations. Moreover, IVR systems are notoriously frustrating and difficult to use. For untrained users, how does the accuracy and speed of an IVR survey compare to that of a live phone survey? And on the whole, can an IVR survey decrease the costs incurred by the organization?

As pertinent as these questions are, we are unaware of any prior comparison between IVR and a live operator for data collection in the developing world. Perhaps the most relevant work is Lerer et al. [16], who conduct an IVR survey for untrained rural teachers in Uganda. While they evaluate several mechanisms to help users complete the survey, the accuracy of entered data remains unknown. Likewise, Medhi et al. have compared IVR systems to text and graphical interfaces, but focus on task completion instead of data accuracy [17]. Studies in rich countries have concluded that IVR can cut operational costs by a factor of ten [3], but it is not clear if this result extends to developing countries, where live operators are cheaper relative to technology and low-educated users might have more difficulty using IVR.

In this paper, we do a formal evaluation of an IVR system versus a live operator, in the context of a real-world phone survey in India. Working in collaboration with Babajob, an online jobs portal in Bangalore, we interview low-income job seekers and gather profile information that would be of interest to potential employers. To compare IVR with a live operator, we interview each user twice, with identical questions but alternate interfaces, and compare the responses to judge the data entry accuracy. A unique aspect of our IVR interview is that it includes a brief introduction by a live operator, to provide context for the call. We conduct tests in an ecologically-valid scenario encompassing 31 drivers, many of whom are actively seeking jobs. Our interview consists of 25 questions that were gathered from real employers, and refined via iterative prototyping with drivers.

Our results indicate that IVR is a viable option for conducting large-scale phone surveys in India. Out of the 20 participants who completed both surveys, we observe a 4.0% error rate (95% C.I. 2.5% – 6.1%). That is, out of 500 total questions administered, 20 were shown to have been entered incorrectly on the IVR interface. We also find that certain elements of our design were effective at reducing errors; for example, the presence of confirmation prompts ("press 1 to

The first author conducted this work during an internship at Microsoft Research India.

DEV'13, January 11–12, 2013, Bangalore, India.

keep your answer, press 2 to change it") on multi-digit questions reduced the overall error rate by a factor of 1.6. With regards to speed, we observe that the IVR interview requires about 2.5 times longer than the live interview. Given current costs in India, this implies that an IVR interview could offer modest reductions (about 1.5x) in the cost of operators and airtime. However, we do not estimate the cost of configuring or maintaining the IVR technology, which could outweigh the costs of operators and airtime, especially for small organizations.

We close with a set of hypotheses on how to further improve the accuracy of IVR data collection, as well as a discussion addressed to organizations such as Babajob that are facing a real-world decision between IVR and a live operator.

2. RELATED WORK

Interactive voice response systems have been around for decades, and there is a large body of literature surrounding their design and application; see [5] for a sound historical perspective. Our work sits at the intersection of three related threads of research: design of usable IVR systems, measurement of survey bias across various modalities, and applications of IVR in the developing world. Despite the maturity of each of these research sub-fields, we are unaware of any prior study that measures the impact of usability challenges on the accuracy of data collected via IVR. Herein lies the novelty of our work.

Designing usable IVR systems: Many authors have advocated guidelines for effective design of usable IVR systems, typically with the goal of enabling users to successfully complete a given task over the phone. For example, Gardner-Bonneau et al. provide 10 recommendations for IVR designers [8] while Oberle offers 11 tips for adapting an IVR to its target audience [19].

Most relevant to our work is that of Lerer et al., who utilize IVR to collect data from teachers in rural Uganda [16]. This work shares our goal of conducting automated surveys for untrained users. However, its principal metric remains that of task completion: what fraction of the users complete the questionnaire without hanging up. No effort is made to measure whether the data entered to the system is an accurate reflection of the users' intent.

Other authors have provided specific guidelines for improving the usability of IVR for novice and low-literate users in the developing world. Medhi et al. compare an IVR spoken dialogue system with a rich multimedia device for mobile banking tasks [17]. They find that while users have no hesitancy speaking to the IVR system, challenges with accent, vocabulary, and the general notion of "speaking to a computer" can inhibit task completion. Patnaik et al. investigate various interfaces for mobile data collection, spanning SMS, electronic forms, and a live operator [24]. They find that a live operator is up to 10 times more accurate than SMS or electronic forms, though they do not compare to any IVR solution.

Several authors have investigated the tradeoff between spoken and typed inputs for IVR applications in the developing world. Patel et al. find that farmers in rural India demonstrated higher rates of task completion using typed (DTMF touch-tone) inputs, as opposed to speech [21]. Grover et al. came to a similar conclusion in Botswana, where users of a health information service preferred touch-tone to speech even though there was not a significant difference in performance [10]. Conversely, Sherwani et al. found that well-designed speech interfaces enabled higher task completion than touch-tone for literate and low-literate health workers in Pakistan [31]. We do not aim to advance the speech vs. touch-tone debate in this paper; we rely on touch-tone inputs whenever a choice is required in the IVR.

To summarize, while the HCI community has developed several recommendations to improve task completion over IVR, as of yet there has not been any measurement of data entry accuracy via IVR. Completing tasks and entering data are related but distinct problems. For example, data entry can often tolerate errors in certain fields, whereas any error in navigating a task hierarchy often leads to a common failure condition. Also, data entry typically has more instances of complex data types, such as multi-digit numbers, as opposed to the multiple-choice answers typical of most IVR systems. At the same time, data entry is a very structured application, where users could perform well with simple guidance.

Measuring bias in IVR surveys: In the health, psychology, and social science communities, IVR systems have been examined from an alternate standpoint, which is their ability (or inability) to collect unbiased reports of sensitive information. For example, it has been demonstrated that respondents are more likely to express extreme positive responses in aural interviews (via telephone and IVR) as opposed to visual interviews (via mail and Web) [7]. A study inquiring about university records found that students were more likely to give accurate replies to sensitive questions in a Web interview than an IVR interview [14]. A survey of sexual behavior compared responses from self-administered questionnaires, face-to-face interviews, and daily IVR diaries, all of which were given to the same respondents [28]. Results showed under-reporting of sensitive information in the face-to-face interviews and over-reporting in the self-administered questionnaires, relative to the IVR system.

While these studies have some overlap with our goal of measuring the accuracy of IVR data collection, they have one fundamental difference: the variations observed in other studies were deliberate on the part of respondents, due to their preference to withhold sensitive information. In contrast, our goal is to measure whether users can successfully utilize an IVR interface to submit the data that they intend to submit. In other words, we measure errors stemming from usability problems as opposed to privacy concerns.

Applying IVR in the developing world: Recently there have been numerous projects that seek to design and deploy IVR systems for the benefit of low-income communities in the developing world. Examples span diverse domains including agriculture [22, 25], citizen engagement [18, 29], community radio [12, 13], health [6, 27, 30], entertainment [2, 26], mapping [15], and education [9].

Our research extends and complements these projects by developing a real-world IVR interview for low-income job seekers. As many of the projects above include some elements of data collection, we anticipate that our lessons learned will also be relevant to improving and extending their reach.

There are also several platforms available for building IVR systems, including IVR Junction [33], Freedom Fone [4], ODK Voice [11], Awaaz.De [23], Tangaza [20], and Gram Vaani [1]. We utilize a variant of IVR Junction for our work.

3. IVR DESIGN

Our goal is to design and evaluate an interactive voice response system that performs an automatic "job interview" with low-income job seekers. The results of the interview could be browsed by potential employers, thereby making it easier for them to connect with qualified applicants. In this work, we limit our focus to the domain of drivers: individuals seeking full-time employment driving a vehicle for a family or a company.

To develop a questionnaire that is both meaningful for employers as well as comprehensible to drivers, we employed an iterative design process. We obtained an initial list of questions by surveying online advertisements by those looking for drivers on Babajob.com (a mobile and Internet jobs portal in India). We refined this list via in-person conversations with several people who have hired drivers in the past, including the transportation manager at a large corporate office. Our resulting prototype questionnaire encompassed numeric questions (e.g., what is your expected salary?), yes/no questions (e.g., are you married?), multiple-choice questions (e.g., what is your level of education?), as well as free-response questions (e.g., how would you navigate to an unknown location?). Numeric, yes/no, and multiple-choice responses were encoded as DTMF key presses, while free responses were recorded in the user's own voice.

To evaluate our prototype, we performed three Wizard-of-Oz trials with drivers at a corporate office, as well as six automated IVR trials that were administered to registered drivers in an online jobs portal. Based on the experiences of these initial users, we improved the questionnaire in several ways. The most important learnings were as follows:

- **Anticipate that users may be in noisy or distracting environments.** Some of our participants had difficulty hearing the prompts because they received our call in a noisy environment, or were unable to give it their full attention. We responded by offering such users the option of rescheduling the call to a time that would be quieter or less busy. Also, we amended the IVR to advise users that each prompt will be repeated if they do not answer; thus, in the event of distractions or noise, they can wait for the repeat instead of trying to answer a question they do not understand.

- **Illustrate multi-digit entry using multiple examples.** Users had difficulty understanding multi-digit entry (e.g., salary field), even when an example was provided by the IVR. Sometimes users did not respond at all, and sometimes they entered the exact digits that were used in the example (even though it did not correspond to their salary). To alleviate this problem, we provided multiple examples of multi-digit entry in the instructions.

- **Explain that '0' is a valid answer to certain numeric questions.** For example, on the question "how many times have you received a traffic ticket?", participants did not know what they should enter if they have never received a ticket. We clarified the prompt to instruct them to press zero if they have never received a ticket. (In retrospect, an even better solution could have been to first ask whether or not they had received any tickets, and if so, to ask for the exact number of tickets.) As another example, for a question asking how many passengers are usually in the driver's car, a participant did not know how to answer because he drove commercial goods instead of people. We eventually decided to remove this question.

- **Enable users to skip sensitive questions.** For sensitive questions ("do you smoke?" and "do you drink?"), some users were disconnecting the call instead of answering. Thus, we added an option to skip three questions. Potential employers could obtain this information separately during a follow-up interview.

- **Prevent users from going too fast.** Because some questions included an explicit confirmation ("press 1 to confirm your answer, or press 2 to change"), one participant wrongly assumed that all questions were followed by confirmation. Thus, he pressed "1" after every response, causing him to answer some questions without even listening to them. We responded by prefacing each new question with a prompt that says, "next question", and has barge-in disabled.

- **Define all terms, leaving no room for interpretation.** Because users do not have the opportunity to ask for clarification during the IVR interview, all potentially ambiguous terms need to be explained fully. For example, a participant did not know if "night shift" implied working continuously overnight, or work-

I. Personal information	
1. Age	multi-digit*
2. Marital status	yes/no
3. Education	multiple choice*
II. Professional information	
4. Own commercial permit	yes/no
5. Years as a driver	multi-digit*
6. Years with license	multi-digit*
7. Number of hours willing to work (per day)	multi-digit*
8. Open to working night-shifts	multiple choice
9. Open to working part-time or short-term	yes/no
10. Latest salary (Rs / month)	multi-digit*
11. Expected salary (Rs / month)	multi-digit*
12. Own vehicle for commute to work	yes/no
13. Carry mobile phone	yes/no
14. Knowledge test: Is Lenin Sarani one-way	yes/no
15. Knowledge test: Is MG Road one-way	yes/no
16. Knowledge test: Landmark near Esplanade	multiple choice
17. Comfortable working outside Kolkata	yes/no
18. Comfortable wearing uniform to work	yes/no
19. Comfortable driving foreigner	yes/no
20. Willing to do odd jobs in addition to driving	yes/no
21. Number of traffic tickets received	multi-digit*
22. Smoking habits	multiple choice*
23. Drinking habits	multiple choice*
III. Free response	
24. Languages understood, spoken, written, read	free response*
25. How to find an unknown place	free response*

** indicates questions with replay and confirmation of response*

Table 1: Data collected in our IVR questionnaire for those seeking employment as drivers. All questions were worded in the local language (Bengali).

ing just a few hours at night. As another example, for a question asking for years of education, a participant did not understand that their 3-year diploma after course 12 counted as "more than 12 years" of education. We clarified all such language to be more explicit.

Incorporating the feedback from initial trials, we arrived at the final design for our IVR system, which is illustrated in Table 1. Our final questionnaire had 25 questions: 7 questions required a multi-digit response, 5 were multiple-choice questions, 11 were yes/no questions, and 2 required open, spoken responses. Three questions (#14-#16) tested drivers' knowledge of local roads, while the others gathered basic personal and professional information. Multiple-choice questions had either three or four choices.

To improve data accuracy, the system required confirmation for all of the multi-digit responses, as well as two multiple-choice responses and both of the oral responses. The questionnaire was developed in Bengali, recorded by a native Bengali speaker and administered only to native speakers of Bengali.

The IVR system was implemented using a variant of IVR Junction [33], based on Voxeo Prophecy, Classic ASP, and IIS Server. We used a GSM modem (Matrix ATA 211G) with a mobile SIM card for the telephony interface.

4. STUDY METHODOLOGY

The broad goal of our user study is to assess the viability of an IVR job interview for large-scale deployment in India. In support of this goal, we are interested in three specific questions: 1) Do users provide similar answers using IVR as they do with a live phone interview, 2) How long does it take users to complete an IVR survey versus a live phone interview, and 3) Is it more cost-effective for an organization to offer an IVR interview or a live phone interview?

To answer these questions, we performed a study that compares answers collected via IVR with those collected via a live phone interview. The basic procedure was to administer the same interview twice to each participant, once using IVR and once with a live phone conversation. Comparisons within subjects enable assessment of data accuracy: if a participant provides the same answer in both interviews, then that answer is deemed "accurate", i.e., it is an accurate reflection of the participant's intended response. If an answer differed between the two interviews, we checked in a follow-up conversation whether the participant intended to give different responses. As some answers were intentionally changed between interviews (e.g., because participants had more time to think about the question), we did not count these discrepancies as errors.

To assess the time required for each interface, a within-subjects comparison is inadequate; participants proceed more quickly during their second interview, because they are already familiar with the questions. Thus, for evaluating elapsed time we rely instead on a between-subjects comparison, examining only the first interview of each participant.

4.1 Experimental Protocol

To administer the survey, we followed the following protocol. We started by calling a participant's mobile phone, based on a number we obtained previously (details below). We explained the purpose and mechanics of the study, being careful to explain that we were not actually hiring drivers, and their responses would be used for research purposes only. In exchange for completing the study, which would take about 30 minutes of their time, we offered participants ₹50 (about one dollar) via a mobile top-up. While this amount is a significant incentive for our target audience, we believe it is also consistent with the real-world usage scenario, in which job seekers are highly motivated to complete a survey to improve their employment prospects.

After obtaining a participant's consent, we administered the interview twice in succession, once with IVR and once via a live phone conversation, with the order randomized (but balanced) across participants. In advance of both interviews, we explained the concept of IVR to users and warned them that a live operator would not be available to answer any of their questions during that interview. Thus, to replicate our results, an organization should expect to initiate each call with a live human conversation, even if IVR is used for the main survey. Following both interviews, a short follow-up survey was administered in order to gather additional demographic data and to clarify any discrepancies between the participant's two sets of survey responses. All calls were recorded for later analysis.

The role of the live operator was fulfilled by the first author, a Ph.D. student in computer science. This operator was a native speaker of the local language (Bengali), but did not have any experience conducting phone interviews. This operator also provided the live introduction, and recorded all prompts for the IVR system. The script for the live operator conversation was identical to the script for the IVR.

4.2 Participants

We made contact with 31 participants, of whom 20 completed the study. All participants were actual drivers, and many were currently seeking a job. We restricted enrollment to native speakers of Bengali, to ensure that varying language skills did not distort the results. Of the 20 participants who completed the study, 10 were identified in cooperation with Babajob, using their internal database of job seekers. Each of these participants had previously registered with Babajob as a driver seeking employment, though they had never provided Babajob with the detailed data collected by our survey. Upon exhausting the Babajob contacts, we utilized snowball sampling to recruit 9 additional participants. We offered participants a referral bonus of ₹50 (about one dollar) for each driver that they referred to us. Based on our phone interviews, we are confident that each person referred is a legitimate driver. Finally, we recruited one participant based on his advertisement in an online jobs portal (http://click.in/).

The demographics of participants were revealed as part of our survey. All participants were male, with an average age of 31 (min=19, max=47). Three quarters of participants were married. Most had received 10 or fewer years of education (n=12), though some received 11-12 years (n=5) or more than 12 years (n=3). Participants had an average individual income of ₹8300 (about $150) per month (min=$54/month, max=$358/month) and an average of 8 years experience as a driver (min=1, max=21). The majority of participants owned a feature phone (n=11), though many had basic phones (n=8) and one had a smart phone. The vast majority (n=17) of participants had used an IVR system before, typically in interacting with customer care.

Participant ID	Question	Question Type	IVR Given 1st or 2nd	IVR Answer	Live Answer	Cause of Error
15	Latest salary (Rs / month)	multi-digit*	2nd	80000	8000	known accident
11	Willing to do odd jobs in addition to driving	yes/no	1st	no	yes	known accident
11	Age	multi-digit*	1st	36	40	re-typed example
11	Years as a driver	multi-digit*	1st	5	17	re-typed example
11	Years with license	multi-digit*	1st	12	16	re-typed example
3	Willing to do odd jobs in addition to driving	yes/no	1st	no	yes	lapse of hearing/understanding
6	Own commercial permit	yes/no	1st	yes	no	lapse of hearing/understanding
15	Knowledge test: Is Lenin Sarani one-way	yes/no	2nd	no	yes	lapse of hearing/understanding
5	Knowledge test: Is MG Road one-way	yes/no	2nd	no	yes	lapse of hearing/understanding
15	Knowledge test: Is MG Road one-way	yes/no	2nd	no	yes	lapse of hearing/understanding
19	Comfortable driving foreigner	yes/no	2nd	no	yes	lapse of hearing/understanding
16	Years as a driver	multi-digit*	2nd	5	4	lapse of hearing/understanding
16	Years with a license	multi-digit*	2nd	4	5	lapse of hearing/understanding
19	Number of traffic tickets received	multi-digit*	1st	0	1	lapse of hearing/understanding
13	Languages understood, spoken, written, read	free response*	2nd	silence	(detailed)	lapse of hearing/understanding
14	Languages understood, spoken, written, read	free response*	1st	silence	(detailed)	lapse of hearing/understanding
4	Languages understood, spoken, written, read	free response*	1st	silence	(detailed)	lapse of hearing/understanding
20	Languages understood, spoken, written, read	free response*	1st	silence	(detailed)	lapse of hearing/understanding
4	How to find an unknown place	free response*	1st	silence	(detailed)	lapse of hearing/understanding
20	How to find an unknown place	free response*	1st	silence	(detailed)	lapse of hearing/understanding

Table 2: Errors observed in the IVR interviews, using the live interviews as a baseline. Errors are sorted by cause, then by type, then by question. Questions denoted (*) request confirmation of the user's response.

The majority of participants (n=11) reported having at least one traffic violation; the average was 2 violations, and the maximum was 12 violations. The majority of participants also reported smoking regularly (n=11), though only one person reported drinking regularly.

5. RESULTS

5.1 Task completion

Of the 31 people that we contacted for the study, 20 people went on to complete both surveys. Of the 11 cases that did not complete, 5 were unrelated to the usability of the IVR system:

- Two people rescheduled their calls for a later time, but when we called them back, they did not pick up.
- One person was interrupted by another call (from his boss).
- One person was dropped due to a software malfunction.
- One person declined to participate.

However, the remaining 6 cases of non-completion were due to interface challenges. Each one of these participants disconnected the call during the IVR interview:

- Five people apparently did not understand a question or did not know how to answer it; after the question repeated several times, they eventually hung up. In two of these cases, we recorded significant background noise from the caller's environment, which may have prevented them from hearing or understanding the question.
- One person could not figure out how to navigate to a numeric dial pad (while the call was in progress) on his touch-screen phone.

In practice, an organization would have to follow-up with these participants to conduct a live phone interview. Alternately, it may be possible to resume the IVR interview at a later time, though we do not evaluate such functionality in this paper.

The remainder of the paper focuses on participants who completed the study. It is possible that these participants are systematically different from the participants who dropped out; for example, they might be more educated, or more technologically savvy. However, this fact does not interfere with the goal of our study, which is to assess the accuracy of an IVR interview *for users who are capable of completing such an interview*. There will always be other users who require a live phone interview instead; however, the data above suggest that this represents a small minority (6 out of 26 people who had the opportunity to take the survey).

5.2 Accuracy

For the purpose of our discussion, we define an *error* to be a case in which a user's response on the IVR interview differed from their response on the live interview, and they later acknowledged that their response to the IVR interview was incorrect. Note that we rely on the user's responses, both during the live interview and the follow-up interview, as the only indications of the ground truth. It remains possible that users are consistently mis-representing the truth during these conversations; we are unable to detect these as errors, nor are we interested in doing so.

In addition, our study does not attempt to characterize the operator's error in transcribing call data into a database. In practice, an operator (especially with limited training) may make several mistakes in this process. Our operator was highly trained, and moreover, recorded values were checked (in cases they differed from IVR) via follow-up with the participant. Audio recordings of the call were also used to clarify any ambiguities.

Across 500 questions administered via IVR, we recorded a

total of 20 errors, for an overall error rate of 4.0% (95% C.I. 2.5% – 6.1%). Nine participants had error-free reports, and five participants logged only one error. Four participants logged two errors, and two participants logged three or four errors. On average, each 25-question survey contained one error.

Sources of Error

The observed errors are enumerated in Table 2. Based on follow-up interviews, we determined that the vast majority of errors (15/20) were due to general difficulties in hearing or understanding the IVR questions. One participant (#11) made three consistent errors on multi-digit responses: instead of keying in his own entry, he repeated the last entry that was used as an example by the IVR prompt. Also, two participants recognized their errors immediately after making it, but were unable to correct it; these errors are labeled "known accident" in Table 2. Of these participants, one (#15) mistakenly thought that he corrected his salary from 80000 to 8000 by pressing a 'cancel' key on his phone.

We do not observe any statistical correlation between the demographic characteristics of participants and their likelihood of making an error. Overall, the values of age, education, prior IVR experience, ordering of IVR versus live interviews, source of recruitment and other indicators were distributed similarly across participants who did (n=11) and did not (n=9) commit any errors. However, as our sample size is small, we have limited power to detect such trends even where they might exist.

Certain questions were more error-prone than others. Questions eliciting a free-form oral response from participants were often met with silence, accounting for 6 errors across 4 participants. Multi-digit responses also remained problematic, accounting for 7 errors. Interestingly, there were no errors observed on multiple-choice questions, even though five of them appeared in the interview.

Impact of Confirmation Prompts

The presence of confirmation prompts helped to avert errors in many cases. We analyzed the recorded calls to understand what the IVR results would have been in the absence of confirmation prompts. This was done by re-executing the call flow as if users had always confirmed their first response, instead of re-entering responses they judged to be problematic.

Without confirmation prompts, there would have been 32 errors instead of 20. In other words, introducing confirmation prompts reduced the error rate by a factor of 1.6. All of the errors averted correspond to multi-digit questions, in which the confirmation prompt enabled users to better understand their own response. For example, a user might misunderstand the question about age and enter the same value that is used as an example in the prompt. However, when the system replies, "you said that your age is 36", the user can more easily recognize the error and move to correct it.

Interestingly, confirmation prompts had no impact on the other question types. For multiple choice questions, no one elected to change the answer originally submitted. For free-response questions, four participants re-recorded their response, but it was only different in tone and not in content. For the multi-digit questions, no participant ever changed a correct response to an incorrect response; however, two changes were made that failed to correct an incorrect response (i.e., the second attempt was incorrect as well).

	IVR Interview	Live Interview
Average	14:12*	6:03
Std. Dev.	2:14	2:14

** IVR interviews are also preceded by a 30-second live introduction*

Table 3: Time taken for participants to complete their first interview. Times do not include the general overview of our study, or the wrap-up questions.

Deliberate Discrepancies

In three cases, discrepancies between the IVR interview and the live interview did not count as errors, because the participant judged the IVR entry to be correct. In two of these cases, participants deliberately changed their answer in the time between interviews, and the IVR interview was most recent. One participant reduced his number of work hours to leave room for a personal commute, while another remembered a more accurate date for obtaining his driver's license. In the third case, a participant indicated a willingness to work outside Kolkata, but only for extra pay; he later summarized this position as "not willing" to the IVR. While the IVR collected less information than the human in this case, we believe the answer collected by the IVR was an accurate reflection of the participant's intent, given the options available.

These cases demonstrate that even live operators do not collect perfect data from participants, as they may arrive at a more accurate answer at a later time.

5.3 Speed

The time taken for participants to complete each interview[1] is illustrated in Table 3. On average, 14:12 was required for the IVR interview, while 6:03 was required for the live interview: a difference that is statistically significant ($t(16)=7.74$, $p<0.001$). The IVR interview also received a 30-second live introduction, in which the operator explains the concept of IVR and cautions users that a human will not be available to answer their questions. Including this introduction, the IVR interview was almost 2.5x slower than the live interview, on average.

The main explanation for this difference is the conservative pacing of prompts in the IVR interview. The prompts alone require 12:20 to play; thus, a user who listens to every prompt is spending less than two minutes total (6 seconds per question) thinking about and submitting their replies, on average. However, by barging in with responses early, some users complete the IVR interview in less time than it would take to play all of the prompts; the fastest completion time is 11:44, about 20% faster than the average. Conversely, some users require much longer to complete the IVR survey, because they change their answers when asked for confirmation. The slowest participant required 18:42, about 30% slower than average.

Note that these figures represent only the first interview that participants complete. If the live interview is adminis-

[1]It was not possible to measure speed for 5 out of the 40 conditions tested, due to interruptions to the interview.

tered second, then it goes faster (average=3:50, stddev=2:08) than if it is administered first (average=6:03, stddev=2:14), a difference that is borderline significant (t(14)=2.00, p = 0.06). The second interview goes faster because participants have already familiarized themselves with the questions. However, speeds on the IVR interview do not decrease when the IVR interview is given second: the average time remains at 14:16 (stddev=2:16). This is because the duration of the prompts remains the bottleneck to completing the survey.

We did not observe any statistical correlation between participants' demographic characteristics and their speed on the interviews. We examined indicators such as age, education, error rate, and others, but our sample size is likely too small to identify meaningful patterns.

5.4 Cost

There are three primary costs involved to organizations seeking to collect data via phone calls: the cost of the phone operator, the cost of the phone calls, and the cost of setting up and maintaining any technology required for IVR. In this discussion, we quantify only the prior two costs: that of operators and air time. The cost of technology setup and maintenance is very important, but could also be variable across organizations. Organizations with existing technical infrastructure and expertise could set up an IVR relatively quickly, while smaller organizations may need more up-front investment. Moreover, the up-front investments are amortized to a varying degree, depending on the ultimate scale of the phone survey.

To estimate the costs of operators and air time, we use a simple model. We assume a fixed per-minute cost $C_{operator}$ of hiring a live operator, and a fixed per-minute cost C_{phone} for phone calls. We assume that the IVR interview requires an introduction by a live operator; while some introduction may also be needed for the live interview, this can be accounted for as part of the normal interview time. Then the total cost TC of performing each interview under either IVR or a live operator can be represented as follows:

$$TC_{IVR} = (C_{operator} + C_{phone}) * \textit{Intro-Length}_{IVR} + C_{phone} * \textit{Interview-Length}_{IVR}$$

$$TC_{Live} = (C_{operator} + C_{phone}) * \textit{Interview-Length}_{Live}$$

Dividing one equation by the other, we can solve for the cost of the IVR interview normalized to the cost of the live interview:

$$\frac{TC_{IVR}}{TC_{Live}} = \frac{\textit{Intro-Length}_{IVR}}{\textit{Interview-Length}_{Live}} + \left(\frac{C_{phone}}{C_{operator} + C_{phone}}\right)\left(\frac{\textit{Interview-Length}_{IVR}}{\textit{Interview-Length}_{Live}}\right)$$

For the interview protocol in this paper, we can substitute our measured times to arrive at the following:

$$\frac{TC_{IVR}}{TC_{Live}} = 0.083 + \frac{C_{phone} * 2.35}{C_{operator} + C_{phone}}$$

This equation implies that, for the interviews evaluated in our study, the IVR interview will be cheaper than the live interview whenever $C_{operator}/C_{phone} > 1.56$. In other words, each minute of the operator's time needs to be at least 1.56 times more expensive than a minute spent on the phone, in order for the IVR system to offer any cost savings.

How does this compare to real-world costs in India? The live operator at Babajob is paid about $180 per month, and spends about 24 hours per week on the phone. Thus, $C_{operator} \approx \$0.031$/min. A reasonable phone plan charges about 60 paise per minute, which implies that $C_{phone} \approx \$0.011$/min. Thus, the cost of a live operator is currently about 2.9x more expensive than the cost of phone calls in India. Under these parameters, the IVR interview currently offers a cost savings of about 1.46x. We emphasize that this estimate does not include the cost of technology setup and maintenance, and is highly dependent on labor and airtime costs, which are taken from the Indian context in 2012.

Overall, it seems reasonable to conclude that IVR is at least cost-competitive with a live operator, though in India it does not yet offer the 10x savings that have been quoted in a Western context [3]. In the future, however, we would expect the cost benefits of IVR to increase, as the cost of labor is rising in the Indian context as well. Thus, it will remain an important and relevant question to understand and improve the accuracy of IVR for administering phone surveys.

6. FUTURE IMPROVEMENTS TO IVR

Based on the experience gained during our study, we offer some hypotheses on how to improve the error rates beyond what we achieved. As opposed to the recommendations in Section 3, these are speculative guidelines that would need to be evaluated in future systems.

R1: Whenever a user is asked to press a key, provide a self-contained description of the meaning of that keypress.

For example, instead of asking "Are you married? Press 1 if yes, press 2 if no", the system should ask, "If you are married, Press 1; if you are not married, press 2." This advice parallels best-practices developed in computer interfaces. For example, a dialog box asking "Save the document?" should label its buttons "Save" and "Don't Save" instead of labeling them "Yes" and "No". The reason is that it is much simpler to affirm a specific action that one intends to take, as opposed to examining the phrasing of a question and formulating an appropriate response.

Several results in our study are in support of this recommendation. First, the multiple choice questions were very successful: not a single error was observed across five multiple choice questions. Furthermore, though three multiple-choice questions requested confirmation from the user, participants never elected to change their answer. In contrast, even relatively "simple" yes/no questions registered 7 errors across our sample. We conjecture that the phrasing of these questions made the difference: the multiple-choice questions succeeded because every keypress was directly associated with a response. This was not the case for other question types.

The second piece of evidence in support of this recommendation is the success of the confirmation prompts for multi-digit responses. Participants did not recognize that they were making an error *until their choice was played back to them*: in other words, until they heard a clear statement of what they were about to submit. This suggests a gen-

eral principle: it may be fundamentally easier for users to check their replies than to construct them. Consequently, users should be given every opportunity to hear the exact response that will be submitted, whether it is a fixed option in the system or a flexible field that they have entered.

R2: Consider a dedicated "undo" button for IVR.

One interesting class of errors we observed is when users immediately recognized they had made a mistake, but were unable to undo that mistake within the IVR. As mentioned previously, one user attempted to correct a mistake using the cancel button on his phone.

This leads to a natural question: why not provide a dedicated "undo" button on IVR systems? For example, the '*' or '#' keys could function as undo in many scenarios. While an undo key could have many possible semantics, perhaps the lowest-hanging fruit is to simply roll back the last keypress made by the user, whether that was in response to a menu option or a multi-digit entry.

One limitation of this feature is that it would still take time to learn. First-time users of an IVR would likely have as much difficulty understanding or remembering the undo function as they would with the rest of the system. However, for users who were familiar with a different IVR system, the presence of a standard undo functionality could perhaps help them adapt to an unfamiliar IVR.

R3: If the user is stuck, provide help via an automated assistant or a dedicated key.

As simple as it sounds, this recommendation could go a long way towards improving both task completion and data quality on IVR systems. We observed that users often displayed signs of being stuck – remaining silent while a prompt repeats – in advance of dropping the call (Section 5.1) or failing to answer an individual question (Section 5.2). While commercial IVR systems routinely default to a live operator if the user is stuck, in the absence of such human resources, our recommendation is to provide either automated assistance or a dedicated "help" button to help a user to successfully complete the survey.

For example, if the user is not responding to a question, the system could automatically invoke a different set of prompts that either explains the question in more detail, or perhaps skips the question to retain the user's interest. For the specific case of users who fail to understand and answer the free-response questions, the system should explain in more detail what is expected of the users. Note that this case will require detecting silence on the part of the users, which could be tricky in the presence of background noise.

An alternative to providing automatic help could be to provide a standard "help" key that a user can always press to either skip the current question or to obtain more information on how it should be answered. Perhaps some users who felt stuck, or did not hear the question properly, would invoke such a key in advance of giving up or responding with an incorrect keypress. The help key could even be explained by a human during the introduction to the IVR, to ensure that it was properly understood by callers. Nonetheless, as per the undo button, a dedicated help key could take some time to learn.

7. DISCUSSION

Overall, our IVR platform compares favorably with a live operator. We were encouraged that users could successfully navigate a complex 25-question interview with no intervention from an operator. The overall error rate using IVR is 4.0%, a figure that is comparable to error rates reported previously for mobile data collection via SMS (4.5%) and electronic forms (4.2%) [24]. However, unlike in [24], our IVR platform did not require any prior training of participants, making it more suitable for large-scale deployment.

Should online job portals such as Babajob adopt an IVR interview in their daily operations? The answer depends on the priorities within that organization. The IVR system has several points in its favor, in particular the ability to rapidly scale-up to many users without recruiting, training, or managing a large fleet of operators. (To replicate the results described here, however, operators would need to be maintained for the live introduction to each call.) The IVR prompts can be recorded in numerous languages and dialects, potentially making it more accessible to populations who might have trouble understanding the accent of a given operator. There could be modest cost savings, as well: approximately 1.5x in the current Indian context, not counting the cost of technology setup and maintenance.

On the other hand, the IVR also introduces new challenges, including: 1) an error rate of 4.0%, which may or may not be tolerable for a given application, 2) some technical sophistication is required for an organization to configure and host the IVR, and 3) an operator may need to follow a more complex workflow to manage interviews across both live and IVR scenarios. Whether the benefits of IVR outweigh the drawbacks will depend on the organization. Are the gains in management overhead worthy to justify a small number of errors in the data collected, as well as technical overhead of running an IVR? For organizations conducting phone surveys at a sufficiently large scale, we expect the answer could be 'yes'.

Before closing, we emphasize that our study considers only a specific socio-demographic context and may not be generalizable across other low-income environments. In particular, we focus on the population of drivers, who are often more technologically savvy than peers in the same income bracket. Our participants were further distinguished by being urban residents, half of whom had already taken the initiative to register with an online jobs portal. The monetary incentive provided may have captured their attention and improved their performance relative to an unpaid interview. Finally, some of our participants were referred by others, and it is possible they may have received tips or other advice from their referrer in advance of taking our survey. While real-world job seekers may also share tips amongst each other, we did our best to discourage such behavior. Also, we do not see any systematic bias in error rate or time taken for those recruited via referral.

8. CONCLUSIONS

This paper demonstrates the viability of an automated IVR interview for low-income job-seekers. Our system is designed and evaluated in an ecologically valid setting, with a full-length, real-world interview administered to actual job seekers. In this process, we also quantify the accuracy, speed, and cost of an IVR survey relative to a live operator.

Our system contains several elements that boost the accuracy of IVR-collected data to an acceptable level: 4.0% error across all questions asked. We showed that it is important to request users' confirmation for responses requiring multiple

keypresses, as it reduces the error rate by a factor of 1.6. We also describe several other design recommendations, gleaned from iterative prototyping as well as the formal evaluation, that help to improve users' experience on IVR.

Finally, we hypothesize that providing a live introduction in advance of the IVR had a large impact in retaining participants throughout the survey. While the industry standard is to start with an IVR and fall back to an operator if needed, perhaps the operator's time is better spent introducing the caller to the IVR, thereby preventing the confusion in the first place. In the future, it would be interesting to test this hypothesis in a real-world system, in which users are initiating the phone call and would not need a live operator to explain the purpose of the study.

9. ACKNOWLEDGMENTS

This study was conceived jointly with Sean Blagsvedt and Maya Chandrasekaran of Babajob. We are very grateful for their help in recruiting participants from their database, as well as their ongoing feedback and support. We also thank Aditya Vashistha for help configuring the IVR system and for feedback on drafts. Finally, we thank the anonymous reviewers for thoughtful comments that helped to improve the paper.

10. REFERENCES

[1] Gram Vaani: http://www.gramvaani.org.

[2] S. K. Agarwal, A. Kumar, A. A. Nanavati, and N. Rajput. User-generated content creation and dissemination in rural areas. *ITID*, 6(2), 2010.

[3] K. Buckstaff, D. McLain, and T. Szybalski. Benchmarking customer service. In *Public Power*, June 2008.

[4] B. Clark and B. Burrell. Freedom fone: dial-up information service. In *ICTD*, 2009.

[5] R. Corkrey and L. Parkinson. Interactive voice response: Review of studies 1989-2000. *Behavior Researc Meth.*, 2002.

[6] W. H. Curioso, B. T. Karras, P. E. Campos, C. Buendia, K. K. Holmes, and A. M. Kimball. Design and Implementation of Cell-PREVEN: A Real-Time Surveillance System for Adverse Events Using Cell Phones in Peru. *AMIA Annu Symp Proc*, 2005.

[7] D. A. Dillman, G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B. L. Messer. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, 38(1):1 – 18, 2009.

[8] D. Gardner-Bonneau, H. E. Blanchard, and B. Suhm. Ivr usability engineering using guidelines and analyses of end-to-end calls. In *Human Factors and Voice Interactive Systems*, Signals and Communication Technology. Springer, 2008.

[9] A. S. Grover, K. Calteaux, E. Barnard, and G. van Huyssteen. A voice service for user feedback on school meals. In *ACM DEV*, 2012.

[10] A. S. Grover, M. Plauché, E. Barnard, and C. Kuun. HIV health information access using spoken dialogue systems: touchtone vs. speech. In *ICTD*, 2009.

[11] C. Hartung, Y. Anokwa, W. Brunette, A. Lerer, C. Tseng, and G. Borriello. Open data kit: Tools to build information services for developing regions. In *ICTD*, 2010.

[12] Z. Koradia, C. Balachandran, K. Dadheech, M. Shivam, and A. Seth. Experiences of deploying and commercializing a community radio automation system in india. In *ACM DEV*, 2012.

[13] Z. Koradia and A. Seth. Phonepeti: exploring the role of an answering machine system in a community radio station in india. In *ICTD*, 2012.

[14] F. Kreuter, S. Presser, and R. Tourangeau. Social Desirability Bias in CATI, IVR, and Web Surveys. *Public Opinion Quarterly*, 72(5):847–865, 2008.

[15] A. Kumar, D. Chakraborty, H. Chauhan, S. K. Agarwal, and N. Rajput. Folksomaps - towards community driven intelligent maps for developing regions. In *ICTD*, 2009.

[16] A. Lerer, M. Ward, and S. Amarasinghe. Evaluation of IVR data collection UIs for untrained rural users. In *ACM DEV*, 2010.

[17] I. Medhi, S. Patnaik, E. Brunskill, S. N. Gautama, W. Thies, and K. Toyama. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Compututer-Human Interaction*, 18(1), May 2011.

[18] P. Mudliar, J. Donner, and W. Thies. Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *ICTD*, pages 159–168, 2012.

[19] F. Oberle. Who, why and how often? key elements for the design of a successful speech application taking account of the target groups. In *Usability of Speech Dialog Systems*, Signals and Commmunication Technologies. Springer, 2008.

[20] B. Odero, B. Omwenga, M. Masita-Mwangi, P. Githinji, and J. Ledlie. Tangaza: frugal group messaging through speech and text. In *ACM DEV*, 2010.

[21] N. Patel, S. Agarwal, N. Rajput, A. Nanavati, P. Dave, and T. S. Parikh. A comparative study of speech and dialed input voice interfaces in rural india. In *CHI*, 2009.

[22] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *CHI*, 2010.

[23] N. Patel, S. R. Klemmer, and T. S. Parikh. An asymmetric communications platform for knowledge sharing with low-end mobile phones. In *UIST Adjunct*, 2011.

[24] S. Patnaik, E. Brunskill, and W. Thies. Evaluating the Accuracy of Data Collection on Mobile Phones: A Study of Forms, SMS, and Voice. In *ICTD*, 2009.

[25] M. Plauché and M. Prabaker. Tamil market: a spoken dialog system for rural india. In *CHI EA*, 2006.

[26] A. A. Raza, M. Pervaiz, C. Milo, S. Razaq, G. Alster, J. Sherwani, U. Saif, and R. Rosenfeld. Viral entertainment as a vehicle for disseminating speech-based services to low-literate users. In *ICTD*, 2012.

[27] N. Sambasivan, J. Weber, and E. Cutrell. Designing a phone broadcasting system for urban sex workers in india. In *CHI*, 2011.

[28] K. Schroder, C. Johnson, and J. Wiebe. Interactive voice response technology applied to sexual behavior self-reports: A comparison of three methods. *AIDS and Behavior*, 11, 2007.

[29] A. Sharma Grover and E. Barnard. The lwazi community communication service: design and piloting of a voice-based information service. In *WWW*, 2011.

[30] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. Healthline: Speech-based access to health information by low-literate users. In *ICTD*, 2009.

[31] J. Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali, and R. Rosenfeld. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *ICTD*, 2009.

[32] M. Thulasingam and P. K. Cheriyath. Telephone Survey as a Method of Data Collection in South India. *Indian Journal of Community Medicine*, 33(4), 2008.

[33] A. Vashistha and W. Thies. IVR Junction: Building Scalable and Distributed Voice Forums in the Developing World. In *NSDR*, 2012.