Authors' unofficial revision.
To appear in the ACM SIGCHI Conference on Human Factors in Computing Systems, May 5-10, 2012, Austin, Texas, USA.

"Yours is Better!" Participant Response Bias in HCI

Nicola Dell† **Vidya Vaidyanathan**‡ **Indrani Medhi**§ **Edward Cutrell**§ **William Thies**§

†University of Washington
nixdell@cs.washington.edu

‡San Jose State University
vidya.vn@gmail.com

§Microsoft Research India
{indranim,cutrell,thies}@microsoft.com

ABSTRACT
Although HCI researchers and practitioners frequently work with groups of people that differ significantly from themselves, little attention has been paid to the effects these differences have on the evaluation of HCI systems. Via 450 interviews in Bangalore, India, we measure participant response bias due to interviewer demand characteristics and the role of social and demographic factors in influencing that bias. We find that respondents are about 2.5x more likely to prefer a technological artifact they believe to be developed by the interviewer, even when the alternative is identical. When the interviewer is a foreign researcher requiring a translator, the bias towards the interviewer's artifact increases to 5x. In fact, the interviewer's artifact is preferred even when it is degraded to be obviously inferior to the alternative. We conclude that participant response bias should receive more attention within the CHI community, especially when designing for underprivileged populations.

Author Keywords
HCI4D, ICTD, demand characteristics, interviewer effects, bias, culture, social status, methods

ACM Classification Keywords
H5.m. Information Interfaces and Presentation (e.g., HCI): Miscellaneous;

General Terms
Design, Human Factors

INTRODUCTION
The rapid proliferation of technological devices throughout the world has allowed a diverse range of previously unreached user groups to gain access to digital technology. The discipline of human-computer interaction (HCI) has embraced the study of these diverse user groups and HCI researchers have proposed a variety of methodologies targeting their specific needs. For example, a growing number of researchers are investigating the ways in which disabled people interact with computer systems [15]; researchers in child-computer interaction explore how computer systems are used by children, particularly in the context of education [24]; cross-cultural HCI looks at what happens when designers and users come from different cultural backgrounds [12] and the relatively new field of human-computer interaction for development (HCI4D) looks at the relationship between humans and technology in the context of international development [1]. Although the methods and objectives of these research domains may vary significantly, they share the characteristic that there are increasingly large differences between the investigators and the people under investigation. These differences may stem from variations in ethnicity, education, age, income and other sociodemographic characteristics.

Figure 1. Interviewing an auto rickshaw driver in Bangalore, India. When shown two technologies by a foreign interviewer (with translator), rickshaw drivers preferred the one they believed to be developed by the interviewer, even when it was obviously inferior.

In this paper, we discuss the increasingly common situation in which the investigators have higher social status and social power than the people they investigate. An increasing amount of anecdotal evidence [1] [18] suggests that in such situations, participants may be particularly susceptible to a type of response bias known as *demand characteristics*. Demand characteristics refer to aspects of a study that may convey the investigator's hypothesis to participants who then adjust their behavior in relation to what they perceive to be the investigator's expectations [22]. Demand characteristics are an important consideration in any research that involves

CHI'12, May 5–10, 2012, Austin, Texas, USA.

human participants and may have a large effect on the study of humans and computer systems.

Although the psychology community acknowledges the effects of demand characteristics, there has been little examination of their effects on the design and evaluation of HCI systems. Qualitative research presented by Brown et al. highlights some of the ways that participants may alter their usage of a system to fit the investigators' expectations [5], but there is a lack of research in HCI that quantifies the extent to which demand characteristics may affect participant behavior. As a result, the significance of the response bias that may result from investigator demand characteristics in HCI remains unknown. In addition, there is a scarcity of research that explores how the social and demographic profiles of investigators and participants influence the ways in which participants respond to demand characteristics.

This paper presents a quantitative analysis of demand characteristics in an HCI setting. As shown in Figure 1, we focus on the developing-world context, in which there are frequently large differences between researchers and participants, and investigate the impact of varying social and demographic factors on the observed effects. For our experiments, we recruited a total of 450 male participants from two distinct population groups in Bangalore, India, and employed two different interviewers to record participant preferences as they interacted with technological devices.

Our work makes four core contributions to the CHI community. First, we survey existing literature to bring demand characteristics and their known effects to the attention of HCI researchers. Second, we show that if participants believe that a particular technological artifact is favored by the interviewer, their responses are biased to favor it as well. Third, we demonstrate that if the interviewer is a foreign researcher who requires a translator, responses are even more biased towards the technology favored by the interviewer. Finally, we show that for a foreign interviewer with translator, participants report a preference for an obviously inferior technology. Our intention is to sensitize researchers regarding the critical and under-appreciated role of demand characteristics, especially when interacting with underprivileged populations.

BACKGROUND AND RELATED WORK

Before discussing in detail the nature of demand characteristics, it is important to note that this is only one of several biases that merit attention in studies with human participants. Other biases include social desirability bias, which occurs when a participant tends to respond in ways that make her look good [23] [25] and evaluator bias, which occurs when the results of a study vary depending on the evaluator analyzing the experimental data [13]. There are also a variety of biases that may be attributed to participant survey methodologies, such as acquiescing or extreme responding [23].

We focus on three categories of related research. First, we draw on psychology literature to discuss the nature of demand characteristics and their effects as well as work that focuses on social influence. Then, we examine several public health studies that look at participant response bias resulting from interviewer characteristics. Finally, we discuss relevant work within the HCI and HCI4D communities.

Psychology

Demand characteristics were first defined by psychologist Martin Orne in 1959 as "the scuttlebutt about the experiment, its setting, implicit and explicit instructions, the person of the experimenter, subtle cues provided by him, and, of particular importance, the experimental procedure itself. All of these cues are interpreted in the light of the subject's past learning and experience" [22]. In a series of psychological experiments performed with undergraduates, experimenters found that participants would willingly perform almost any task, regardless of how meaningless, boring or uncomfortable it was, since they had knowingly and willingly placed themselves under the control of the experimenter [20]. In addition, participants in an experiment often share with the experimenter the hope that the study will be successful [22]. Frequently, a participant will want to ensure that she makes a useful contribution to the study and so will strive to be a 'good' participant and provide the experimenter with the 'right' results. Alternatively, a participant may resent the experimenter and actively work to disprove the hypothesis. In either case, a participant should not be viewed merely as a passive responder, but rather as an active agent with a real stake in the outcome of the study.

Although researchers have acknowledged that there may be a connection between demand characteristics and the tendency for participants to respond in ways confirming the experimenter's hypothesis, few have designed studies specifically to quantify this effect. One notable exception is a 2008 study by Nichols and Maner in the US that investigated the extent to which possessing knowledge of the study's hypothesis affected participant behavior [21]. The findings suggest that demand characteristics may make experimental effects appear more substantial than they actually are.

Demand characteristics cannot be eliminated from any study. In the absence of obvious demand characteristics, participants will guess the experimental hypothesis and alter their behavior accordingly [22]. Thus, instead of trying to eliminate demand characteristics, it is better to take them into account, study their effect, and manipulate them if necessary. Psychologists have proposed several techniques to disguise the purpose of the study or detect participants that guess the real hypothesis. These techniques include using the post-experimental inquiry [26], non-deceptive obfuscation [31] and a so-called red herring technique [17]. Nevertheless, few studies have directly examined the effects of demand characteristics or sought to identify factors that may increase or decrease the likelihood that participants' succumb to demand characteristics [21]. Additionally, different participant populations are likely to respond to demand characteristics in different ways, and it is important to study under what circumstances, in what kind of experimental contexts, and with what kind of participant populations, demand characteristics become significant in determining participant behavior [22].

However, most research investigating demand characteristics has been performed with undergraduates in the United States. To the best of our knowledge, no psychology experiments have specifically investigated the extent to which demand characteristics might affect studies performed in developing countries or with disadvantaged communities.

A number of psychological studies examine the influence of social status on decision-making and social conformity. Strodtbeck and Lipinski [27] found that jury members of higher socioeconomic status were more likely to be elected as jury foremen than members of lower socioeconomic status. Kirchler and Davis [16] studied the effects of status differences on group consensus and found that participants of higher status changed their individual opinions and preferences less often than those of lower status. Finally, Asch [2] performed experiments to study the social and personal conditions that cause participants to resist or to yield to group pressures, and found that participants conformed with answers given by other people in the group even though the answers chosen were objectively and noticeably wrong.

Public Health

Public health programs frequently define their target populations by ethnicity, gender, age and other sociodemographic characteristics, and there are several relevant studies exploring how interviewer characteristics might affect public health data. Davis et al. review studies in the US that show race, ethnicity and gender effects [8]. In this context, response bias appears to be most likely to occur when survey items query attitudes about sociodemographic characteristics or respondents' engagement in sensitive behaviors [14] [30].

Several other studies analyze public health survey data in developing countries. Bignami-Van Assche et al. examined data collected by local interviewers in Kenya and Malawi and conclude that interviewer gender may affect participant responses to sensitive questions [4]. Weinreb found that respondents in Kenya admitted telling untruths to stranger interviewers because the interviewers were not known by the community and their motives were therefore suspect [29]. Bernhart et al. found that the tendency of respondents to withhold critical comment hampered the collection of patient satisfaction data by government workers in health centers in Indonesia [3]. The paper suggests that more useful information might be obtained by asking participants about events and behaviors, rather than for their opinions.

HCI and HCI4D

Given the importance of usability studies in HCI, it is surprising that there has been so little attention paid to the effects that demand characteristics may have on their reliability. In both field trials and laboratory testing, users are frequently aware of the researcher's role in the study and the hypotheses under investigation. Although we acknowledge that it might be impossible to hide a study's true purpose, participant comments and suggestions are frequently taken at face value and the potential for participant response bias is ignored. A notable exception is the field of child computer interaction in which papers stress that even where there is no deliberate intervention the interviewer has an effect [24].

We found only one study that specifically addresses demand characteristics in HCI. In a 'trial of trials' Brown et al. found that participants changed their system usage partly to give researchers 'good' data [5]. The authors argue that demand characteristics are a part of what makes field trials possible and may be exploited to encourage participant usage. The paper also suggests that the need for researchers to present their systems as successful is problematic, and that it would be better to postpone the evaluation of technologies until they can be better understood by users. While our findings are aligned with Brown et al., we go beyond qualitative observations and contribute a rigorous measurement of the influence of demand characteristics in an HCI setting.

There is a growing body of work that explores the role of culture in HCI. Most of this work addresses the design of global interfaces that can accommodate users' cultural differences [9] [10]. One particularly relevant study by Vatrapu et al. examines the effects of culture in structured interviews in the US [28]. Two groups of Indian graduate students were asked to evaluate a website, and the group with an Indian interviewer provided more feedback and identified more culturally sensitive materials than the group with a US interviewer. While this study is similar to ours at a high level, there are a number of key differences. First, in our study participants interact with two technological artifacts rather than a single website. Since we know the full extent of the differences between the artifacts, we are able to compare participant responses between different interviewers as well as in relation to ground truth information. Second, in the study by Vatrapu et al., the differences between the two participant groups relate mainly to the identification of culturally sensitive materials. In contrast, our experiments relate to purely technological artifacts. Finally, their study involved 16 students while we interview 450 participants.

There are also a number of papers that discuss the role of culture in the developing world. Irani mentions that cultural differences between usability evaluators and participants can affect evaluation outcomes [12]. Ho et al. find that the hierarchical structure of some societies causes users to withhold criticism from researchers [11]. Chavan encourages participants to submit critical feedback by situating user studies within dramatic storylines [6].

The relatively new subfield of HCI4D targets the design and evaluation of systems that promote international development. Several recent papers anecdotally mention that foreign researchers may affect the results of HCI4D studies. Anokwa et al. had difficulty eliciting negative feedback from users and discuss the importance of gathering data from multiple sources [1]. Ledlie discusses how projects can be hampered by a lack of cultural insight and suggests methods for obtaining critical feedback from participants [18]. Table 1 summarizes the papers discussed in this section and highlights the scarcity of research that quantifies participant response bias due to demand characteristics in an HCI setting. Our paper targets this gap, providing rigorous (and sobering!) experi-

	Psychology	Public Health	HCI
Qualitative and/or Anecdotal	Orne '62 Sawyer '75 Laney '08 Zizzo '08	Williams '68 Johnson '94 Bernhart '99 Weinreb '06 Davis '10	Read '05 Chavan '05 Vatrapu '06 Anokwa '09 Ho '09 Irani '10 Ledlie '10 Brown '11
Quantitative	Milgram '63 Rosnow '73 Sawyer '75 Laney '08 Nichols '08	Williams '68 Johnson '94 Bernhart '99 Bignami-Van Assche '03 Weinreb '06	Vatrapu '06

Table 1. Summary of related research on participant response bias. We target the highlighted scarcity of quantitative research exploring participant response bias due to demand characteristics in HCI.

mental data to guide the design of future studies and to help interpret studies that have already been completed.

EXPERIMENTAL DESIGN

Frequently, the aim of a research project in HCI is to introduce a new technological artifact into a target community, explore the design issues associated with the new technology, and evaluate the potential for the technology to impact the community. The nature of HCI research often requires researchers to spend considerable time in the field interacting with users. However, in many cases, researchers are not members of the target community and may differ from users in ethnicity, language, culture and socioeconomic status. As discussed in the previous section, researcher demand characteristics have the potential to impact the responses obtained from users, and we wanted to quantify this effect. Specifically, we formulated the following hypotheses:

H.1 If participants believe that the interviewer favors a technology, their responses will be biased to favor it as well.

H.2 If the interviewer is a foreign researcher requiring a translator, participants' responses will be even more biased towards the technology favored by the interviewer.

H.3 Participants will express a preference for an obviously inferior technology if they believe it is favored by the interviewer.

To test our hypotheses, we recruited a total of 450 participants and conducted a field study in Bangalore, India that comprised two main experiments. In Experiment 1, which was designed to test H.1 and H.2, participants were shown an identical video clip on each of two identical smartphones, one after the other. We purposely introduced demand characteristics by having the interviewer clearly associate herself to one of the phones by telling participants that she was working to improve the video player on that phone. Within this scenario we investigated if changing

	Experiment 1: Identical Videos	
	Rickshaw Drivers	Univ. Students
Foreign Interviewer†	50	50
Local Interviewer	50	50
	Experiment 2: Degraded Video	
	Rickshaw Drivers	Univ. Students
Without Association‡	50	50
Foreign Interviewer†	50	50
Local Interviewer	50	0*

† The foreign interviewer interacted with rickshaw drivers with the aid of a translator.

‡ This condition represented a baseline that minimized demand characteristics by removing phrases from the script (in **bold**) that associated one video to the interviewer.

* Because the results obtained from Experiment 1 showed no significant differences between the foreign and local interviewers with university students, we performed this experiment with only one of the interviewers.

Table 2. Number of people interviewed for each experimental condition.

the social and demographic profiles of the interviewers and the participants affected the extent to which participants succumbed to demand characteristics. To do this, interviews were conducted with two interviewers (a foreign, Caucasian interviewer and a local, Indian interviewer) and two sets of participants (auto rickshaw drivers and university students).

Experiment 2 tested H.1, H.2 and H.3 by obviously degrading one of the video clips and seeing if participants stated a preference for the degraded video clip when it was associated with the interviewer. For quick reference, Table 2 summarizes all the experimental conditions that we tested. The rest of this section discusses the general experimental procedure and the characteristics of the different interviewers and participant populations. In subsequent sections we discuss additional details and variations in procedures that were specific to each experimental condition.

Experimental Procedure

Data collection was performed over a period of 5 weeks in July and August 2011. Our experiments utilized a between subjects design with a sample size of 50 for each experimental condition. Individual participant interviews were conducted from Monday to Saturday, between 12pm and 4pm, with each interview lasting between two and three minutes. We employed the same general interview procedure across all experimental conditions. In advance of the interviews, we uploaded a 21 second video clip of a popular local music video to each of two identical Windows smartphones. The video clip had a resolution of 640 x 480 pixels per frame and 30 frames per second. The phones were set to use exactly the same video player, as well as identical volume and brightness levels. Individual interviews were administered by

reading the following script to participants (the exact phrases introducing demand characteristics are highlighted in bold):

"Thank you for participating in my experiment. I am a computer science researcher and I'm trying to improve video players on mobile phones. I want you to watch a short video on these two phones and tell me which one looks better, or if they look the same. The same video will play on both phones, **but this phone uses my new player** *[indicate phone]. Please tell me your honest opinion and please concentrate because I will play each video only once. Do you have any questions? Ok, watch this one first.* **This one uses my new player** *[play video]. Now watch this one [play video]. Which one do you think looks better or do they look the same? Why? Thanks very much!"*

In each experimental condition, the order in which the video associated to the interviewer was played was randomized to prevent any bias due to ordering effects. The interviewer recorded participant responses and comments on paper for later analysis and aggregation. Responses were coded into three distinct classes: those that favored the video associated with the interviewer, those that favored the video not associated with the interviewer, and those that said the two videos looked the same. We included the option of "same" because we expected that it would provide more nuanced data than a forced-choice paradigm in which participants were required to state a preference for one video. However, this paper focuses on responses that preferred one video to the other, and leaves detailed analysis of "same" responses for future work.

Interviewers

Since we wanted to vary the social status of the interviewers relative to the participants, we conducted interviews using two different female, graduate student interviewers: a 29-year-old, English-speaking Caucasian, referred to from now on as the foreign interviewer, and a 33-year-old, Kannada- and English-speaking Indian referred to from now on as the local interviewer. The local interviewer grew up in the same neighborhood in Bangalore in which the interviews were conducted. As a result, in addition to speaking the local language, she was identifiable as a local member of the community through her dress and knowledge of the customs of the area. In contrast, the foreign interviewer was not born in India, and had spent approximately one month in Bangalore at the time that the experiments were performed. Thus, she was distinguishable as an outsider by her ethnicity, language, dress and unfamiliarity with the local customs.

Since participants in one of the groups (auto rickshaw drivers) spoke limited English, the foreign interviewer required a translator to interact with them. For consistency, we utilized the local interviewer as the translator. The need for a translator necessarily required the presence of two interviewers for the interactions with the foreign interviewer but only one for the interactions with the local interviewer. It is well known that the presence of multiple interviewers may have an affect on participant conformity [2]. However, since the presence of a translator is a common occurrence in many HCI4D projects, and since part of our goal is to emulate a realistic HCI4D setting, any response bias resulting from the presence of two interviewers, rather than a single interviewer, would also be a factor in HCI4D projects and as such is part of the effect that we are trying to measure.

The social status of local and foreign interviewers differed in the eyes of low-income individuals in India. Although foreigners are perceived differently in different countries, in India Caucasians are usually perceived as having a high social status. This owes partly to India's history as a colony under British rule. Also, independent of its past, Caucasian visitors are likely to have an education and income that is higher than the local mean, and are usually fluent in English, a language associated with prestige and opportunity. In addition, during interviews with rickshaw drivers, the presence of the translator further elevated the social status of the foreign interviewer.

Participants

Participants were recruited from two distinct social groups that we chose on the basis of availability and social status relative to the interviewers. The first group consisted of male university students from the Indian Institute of Science (IISc), an elite scientific graduate institute in India. We restricted participation to male students since the other participant population (rickshaw drivers) is composed of males. Since both of the interviewers were graduate students, the social status of this population was relatively well-matched to that of the interviewers. In addition, all IISc students speak English and typically have experience using and understanding sophisticated technology. We recruited a total of 200 male university students aged 19 to 41 (M=25 years, SD=3.8 years). Recruitment was performed on campus at IISc by approaching individuals and asking them to participate in a research project. Participants that agreed were then interviewed immediately. Individual interviews were done in English by either the foreign interviewer or the local interviewer. Participants were not compensated, other than being thanked for their time.

The second participant group consisted of local auto rickshaw drivers. Auto rickshaws are 3-wheeled vehicles that provide cheap transportation in India. In Bangalore, rickshaw drivers are men who usually have some high-school education and a daily income of between US $5 and $10. Rickshaw drivers typically possess cheap mobile phones but do not have extensive experience with sophisticated technology. As a result, the socio-demographic difference between the rickshaw drivers and the interviewers was greater than it was between the university students and the interviewers. Most rickshaw drivers in Bangalore speak Kannada, the local language in the Indian state of Karnataka.

All rickshaw drivers were recruited by the local interviewer on a single street in Bangalore. The local interviewer stood on the side of the road and hailed passing auto rickshaws. Rickshaw drivers that stopped were then invited to participate in the experiment. Depending on the condition, the foreign interviewer would step up at this point or the local interviewer

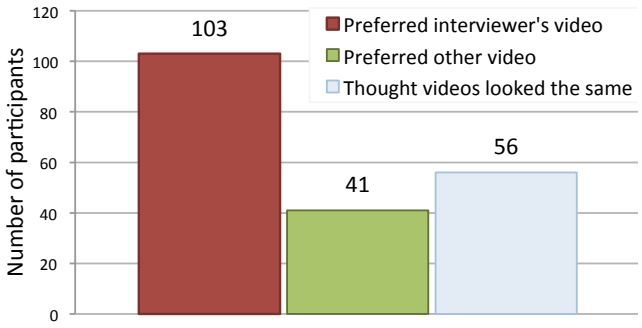


Figure 2. Results of Experiment 1: Preferences stated by participants when shown identical video clips (combined across all conditions).

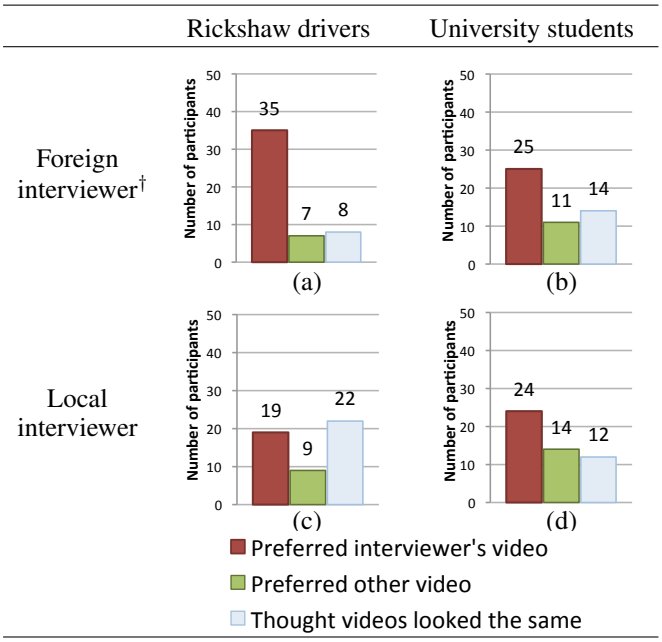
would begin the experiment. In both cases, the rickshaw driver remained seated in the vehicle. We simplified the interview script slightly to make it easier for rickshaw drivers to understand. Specifically, in the second sentence of the script, the phrase *"I am a computer science researcher"* was changed to *"I am a computer scientist"*. Other than this small change, the interview script was identical to that used for the university students. A total of 250 male rickshaw drivers aged 19 to 72 (M=37 years, SD=11.2 years) were interviewed. Participants were compensated for their time with a small gift worth about $1. However, a large number of participants refused to accept compensation, and as a result we ceased compensation for the rickshaw drivers in Experiment 2.

EXPERIMENT 1: RESPONSE TO IDENTICAL VIDEOS

Experiment 1 recorded participants' preferences when shown an identical video clip on each of two identical smartphones, with the interviewer associating herself to one of the video clips as described in the previous section. The experiment utilized a 2x2 factorial design in which we varied both the interviewer (foreign and local) and the participants (rickshaw drivers and university students).

The results of the experiment are summarized in Figure 2, which shows participant responses aggregated across all conditions. As expected, participants who expressed a preference for either video were more likely to choose the video associated with the interviewer. For this paper, we define the *response bias* as the ratio between the number of participants who preferred the interviewer's video to the number of participants who preferred the other video. Averaging across all conditions, the response bias was 2.5x.

Detailed results for each condition appear in Figure 3. While there was a response bias in all cases, the magnitude of the bias varied with the interviewer and participant group. The largest bias occurred when the foreign researcher interviewed rickshaw drivers: there was a 5x bias in favor of the interviewer's video (Figure 3a). The smallest bias, 1.7x, occurred in the opposite configuration, when the local researcher interviewed university students (Figure 3d). The other conditions showed intermediate response biases of 2.3x with the foreign interviewer and university students (Figure 3b) and 2.1x with the local interviewer and rickshaw drivers (Figure 3c).



† The foreign interviewer interacted with rickshaw drivers with the aid of a translator.

Figure 3. Participants' responses when shown two identical videos.

Testing H.1: Presence of Response Bias

To evaluate Hypothesis 1, we compare to the null hypothesis that interviewer association does not impact participant responses, i.e., that the same number of people choose the interviewer's video as the other video, and the response bias is 1. The null hypothesis is strongly rejected for the aggregate responses, as reflected in Figure 2 ($\chi^2(1, n = 144) = 26.7, p < 0.001$). At a finer granularity, the bias is also significant in the case of the foreign interviewer, interacting either with rickshaw drivers ($\chi^2(1, n = 44) = 18.7, p < 0.001$) or university students ($\chi^2(1, n = 36) = 5.4, p = 0.02$). The bias observed with the local interviewer was borderline-significant in the case of rickshaw drivers ($\chi^2(1, n = 28) = 3.6, p = 0.06$) and not significant in the case of university students ($\chi^2(1, n = 38) = 2.6, p = 0.10$). However, pooling across both participant groups does reveal a significant bias in response to the local interviewer ($\chi^2(1, n = 66) = 6.1, p = 0.01$).

Testing H.2: Impact of Foreign Interviewer

Hypothesis 2 states that the response bias increases when the interviewer is a foreign researcher requiring a translator. The only condition satisfying this criterion is that of the foreign interviewer with rickshaw drivers. Thus, to test this hypothesis, we compare the response bias observed with the foreign interviewer and rickshaw drivers to conditions with a different interviewer (local instead of foreign) or different participants (university students instead of rickshaw drivers).

Our results suggest a trend that is consistent with the hypothesis: the bias between foreign interviewer and rickshaw drivers is 5x (Figure 3a), but it decreases to 2.1x when replacing the foreign interviewer with a local one

Figure 4. A single frame from the high quality video clip (left) and the low quality, degraded video clip (right).

(Figure 3c), or to 2.3x when replacing rickshaw drivers with university students (Figure 3b). To evaluate the significance of this trend, we utilize 2x2 contingency tables, in which the variables are video chosen (*Interviewer's, Other*) and, depending on the test, interviewer (*Foreign, Local*) or participant (*Rickshaw Driver, University Student*). For Experiment 1, we do not find a significant relationship between the video chosen and the interviewer ($\chi^2(1, n = 70) = 2.28, p = 0.13$) or the participant group ($\chi^2(1, n = 78) = 2.11, p = 0.15$)). However, this effect is significant in Experiment 2, as described in the next section.

EXPERIMENT 2: RESPONSE TO A DEGRADED VIDEO

We designed Experiment 2 to measure the extent of participant response bias in the face of an obviously poor technological artifact. To do this, rather than showing participants identical video clips, we made one of the video clips noticeably worse than the other and had the interviewer associate herself to the degraded clip. Specifically, the resolution of one of the clips was decreased from 640 x 480 to 120 x 90 pixels per frame (the media player scaled both videos to the full screen width of 800 x 480.) Additionally, the video frame rate was halved, from 30 to 15 frames per second. The audio, brightness, content and length of the video clips remained unchanged. Sample video frames from the original, high-quality video clip and the degraded, low-quality clip are shown in Figure 4. For our experiments, we loaded the low-quality video clip on one smartphone and the high-quality clip on the other.

To ensure that the video clip had been sufficiently degraded so as to be noticeably different from the original, high-quality clip, we performed an experiment in which participants were shown the two video clips one after the other without the interviewer associating herself to either clip. To achieve this, we modified the interview script by removing the phrases from the script (emphasized in **bold**) that associated the video to the interviewer. The rest of the interview script remained unchanged. The order in which the low-quality clip was played was randomized to avoid order effects.

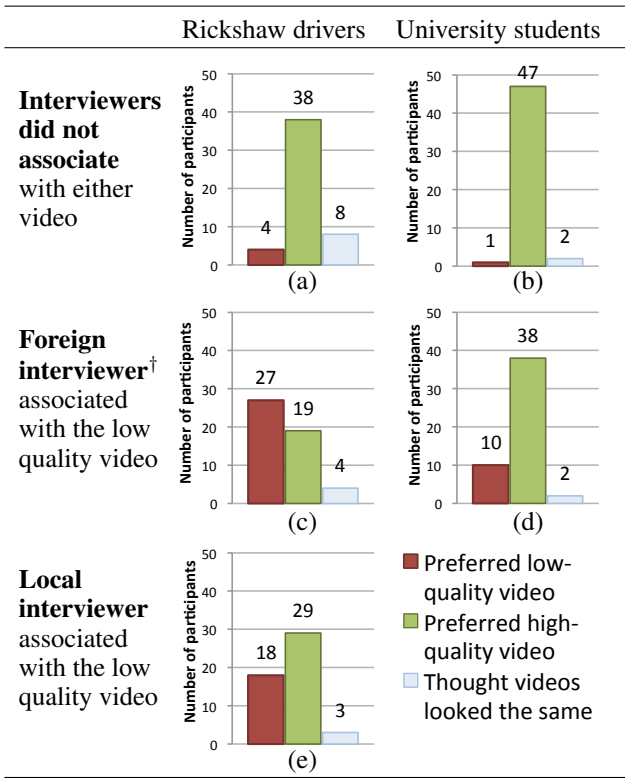
Following this, interviews were conducted in which the interviewer associated herself to the low-quality clip. The experimental procedure and interview script used were identical to those described for Experiment 1. As in Experiment 1, interviews were conducted for two sets of interviewers (foreign and local) and two sets of participants (rickshaw drivers and university students). However, we did not evaluate the condition in which the local researcher interviews university students. Since the results obtained from Experiment 1 showed no significant differences between the foreign interviewer and the local interviewer with the university students, we felt that performing this experiment with only one of the interviewers was sufficient.

Overview of Results

The results of Experiment 2 appear in Figure 5. When the two videos were presented without the interviewer associating to either video, university students (Figure 5b) overwhelmingly chose the high-quality clip (about 95%). Rickshaw drivers usually chose the high-quality clip (about 75%), though they said they looked the same about 15% of the time (Figure 5a).

When the interviewer associated herself to the degraded video, the degree of response bias varied by participant group. In the case of university students (Figure 5d), most participants still selected the high-quality clip (about 75%). However, almost 20% of participants chose the low-quality (associated) clip, an order of magnitude more than when it was unassociated.

The response bias is more dramatic in the case of rickshaw drivers. When the interviewer was associated with the low-quality clip (collapsing across foreign and local interviewers), rickshaw drivers were almost as likely to choose the interviewer's (low-quality) clip as the other (high-quality) clip, 45% and 48%, respectively. However, as in Experiment 1, there was a dramatic difference between how participants responded to the local interviewer alone compared to the foreign interviewer with local translator. When the low-quality clip was associated with the local interviewer alone, participants were 4.5 times more likely to select it than when it was unassociated (Figure 5e), though they still usually preferred the high-quality video (36% low-quality vs. 58% high-quality). But when the low-quality video was associated with the foreign interviewer (Figure 5c), this ratio flipped:



† The foreign interviewer interacted with rickshaw drivers with the aid of a translator.

* Because the results obtained from Experiment 1 showed no significant differences between the foreign and local interviewers with university students, we performed this experiment with only one of the interviewers.

Figure 5. Participants' responses with the degraded video.

participants were even *more* likely to choose the low-quality video over the high-quality one (54% vs. 38% respectively).

This is an important finding since in a normal HCI4D scenario, it is likely that only one of the two interviewers would be utilized, yet they lead to opposite conclusions. Responses submitted to the foreign interviewer suggest that the low-quality video is preferred, but responses submitted to the local interviewer suggest that the high-quality video is preferred! While the trial without association eliminates the bias, in practice it is rarely possible to remove all elements of association in evaluating a new system. Thus the choice of interviewer determines the outcome of the experiment.

Testing H.1: Presence of Response Bias

Like Experiment 1, we define response bias in terms of the number of participants that preferred one video to the other. However, because the videos are no longer identical, the bias is measured relative to the preferences stated without interviewer association. For a given interviewer and participant group, we test for response bias using a 2x2 contingency table in which the variables are video chosen (*Interviewer's*, *Other*) and interviewer association (*Foreign or Local Association, Without Association*). Since some of the counts in the table are small, we use Fisher's exact test to improve upon the accuracy of the chi-squared test. We found a significant relationship between video chosen and interviewer association across all conditions: foreign interviewers with rickshaw drivers ($p < 0.001$), foreign interviewers with university students ($p = 0.008$), and local interviewers with rickshaw drivers ($p = 0.003$).

Testing H.2: Impact of Foreign Interviewer

To test for the impact of the foreign researcher (with translator) on the response bias, we compare the responses of rickshaw drivers across the local and foreign interviewers. (Unlike Experiment 1, we do not compare the results of the foreign interviewer across different participant groups, because these groups demonstrated different preferences even without interviewer association.) We utilize a 2x2 contingency table, in which the variables are video chosen (*Interviewer's*, *Other*) and interviewer (*Foreign*, *Local*). We find that the interviewer has a significant association with the video chosen, and hence with the response bias ($\chi^2(1, n = 93) = 3.87, p = 0.049$, Cramér's $V = 0.288$).

Testing H.3: Preference for Inferior Technology

Hypothesis 3 states that participants may express a preference for an inferior technology if they believe it is favored by the interviewer. This hypothesis is consistent with our results: in the case of foreign interviewers with rickshaw drivers the participants select the low-quality video 54% of the time (selecting the high-quality video 38%, and neither video 8% of the time). While the preference for the low-quality video over the high-quality video is not statistically significant ($\chi^2(1, n = 46) = 1.39, p = 0.24$), it is alarming that the foreign interviewer is unable to confirm the superior technology in this scenario. Furthermore, even the local interviewer would be unable to reject the hypothesis that the videos were of equal quality ($\chi^2(1, n = 47) = 2.57, p = 0.11$), as only 58% of rickshaw drivers responded with a preference for the high-quality video (versus 36% for the low-quality video, and 6% saying same).

DISCUSSION

Explaining Participant Response Bias

To further understand why participants responded in the ways that they did, we collected and analyzed comments that explained their preferences. Many comments were surprisingly detailed and thoughtful. After watching two identical video clips, one university student told us, *"You are having a better frame rate, which is reducing the blur affect that is there in the other player. The resolution is very clear, so I think if you improve a little bit more, then it will be a great player"*. A large proportion of participants also believed that they saw a clear difference between the two video clips: *"I feel that in the newer version which you have coded, whenever there was a significant color contrast between two parts of an image, your version was somewhat smoother and less pixelated"*. The rickshaw driver participants also provided convincing reasoning to support their choices: *"The quality of the background color and figures is too light in*

that player while the quality of the color and graphics is better in [your] one. Will [your] new player be introduced in the market?" Additionally, several participants seemed anxious to give us a genuine answer: *"I've given you my honest opinion, so please don't be cross with me if it wasn't the right one"*. These comments suggest that participants did not just tell the interviewer the 'right' response while secretly thinking otherwise, but rather that participants seemed to genuinely believe the interviewer's artifact to be superior and identified convincing reasons to justify their choice. These findings have important implications for researchers and indicate that even detailed and convincing participant opinions cannot be taken at face value.

Recommendations

Our primary recommendation is that researchers pay more attention to the types of response bias that might result from working with any participant population and actively take steps to minimize this bias. To do this, it is important that interviewers dissociate themselves as much as possible from any particular design or solution. Our findings indicate that if participants are aware of the interviewer's personal stake in the outcome of the study, the results are more likely to be affected by demand characteristics. Additionally, collecting and reporting subjective information from participants as a primary method of evaluation is problematic and should be avoided. We have shown that even though participant comments might be detailed, well thought out and delivered with conviction, they do not necessarily reflect the merit of the solutions at hand. As far as possible, the focus of participant interviews and feedback should be on obtaining factual, rather than subjective, information [3]. Using implicit metrics [7] or triangulation [19] to validate the data collected could further increase confidence in the results of the study.

Our findings also suggest that minimizing the differences between the interviewer and the participants could help to mitigate the response bias resulting from interviewer demand characteristics. A large number of existing of HCI4D research papers extol the practice of 'field work' in which researchers spend time with potential users in those circumstances in which the technology might take hold [1] [18]. While we do not dispute the value of field work or the benefits of establishing rapport with users, we stress that care must be taken to understand the complications and error that may result from the influence of researchers working in communities that are vastly different from their own.

Generalization and Limitations

Our experiments focus on the ways in which social and demographic factors may affect participant response bias due to demand characteristics. However, this is only one aspect of demand characteristics, and there are undoubtedly many more that could play an influential role on the outcome of a study. For example, research suggests that gender might be an important factor that could influence participant responses [8]. However, in this study we have specifically avoided examining the extent to which gender may play a role in any bias observed. All of our participants were male and all of the interviewers were female. A separate study would be required to understand the ways in which response bias might be affected by participant and interviewer gender.

Additionally, social status and ethnicity are influential social characteristics in many cultures, but their exact effects are likely to vary from culture to culture. Our experience in India has been that Westerners are often afforded special or preferential treatment, but research performed in rural Kenya suggests that outsiders there may be treated with hostility and suspicion [29]. Furthermore, our experiments only deal with situations in which the social status of the interviewer is either the same as or higher than the participants. Further research is required to investigate the nature and magnitude of participant bias in other contexts.

The analysis in this paper focused on the case in which participants expressed a preference for one video or another, without paying much attention to the cases in which the videos were judged to be the same. The "same" responses contain valuable information and represent a fruitful opportunity for follow-up analysis. For example, in the case of the local interviewer and the rickshaw drivers, there is a strong preference for "same", which could be evidence of a partial response bias amongst participants who otherwise would have chosen the video that was not associated with the interviewer. More sophisticated analysis tools could quantify this effect.

Finally, the experiments presented in this paper have been conducted within a particular culture and city, and with two specific participant populations. Since we wanted to emulate a realistic HCI4D setting, one of the chosen participant populations required the presence of a local language translator for the interactions with the foreign interviewer. As a result, we are unable to determine whether the response bias we observed was caused primarily by the large difference in social status between the foreign interviewer and the rickshaw drivers, by the addition of a translator, or both. Further research is required to tease apart the extent to which these different factors individually influence participant responses.

CONCLUSIONS

As the field of HCI embraces the globalization of technology, researchers and practitioners are increasingly working with groups of people that differ significantly from themselves. This paper brings the notion of demand characteristics to the attention of the CHI community and explores the effects that they may have on participant responses. Via experiments with 450 participants in Bangalore, India, we showed that (1) if participants believe that a particular technological artifact is favored by the interviewer, their responses are biased to favor it as well, (2) the bias due to interviewer demand characteristics is exaggerated much further when the interviewer is a foreign researcher requiring a translator, and (3) in response to a foreign interviewer with a translator, participants of lower social status report a preference for an obviously inferior technology, which they otherwise do not prefer. Until now, the significance of demand characteristics

in HCI has remained largely unexplored and undervalued. We have demonstrated that it is crucial for researchers and practitioners to pay more attention to the role of social status and the effects that demand characteristics may have in the design and analysis of studies involving human participants.

ACKNOWLEDGEMENTS

We would like to thank Sara Kiesler for going above and beyond the call of duty to improve this paper. We also want to acknowledge Mary Czerwinski, Jonathan Donner, Jonathan Grudin, Brandon Lucia, Anne Oeldorf-Hirsch, Nimmi Rangaswamy and all of our participants for their valuable contributions to this research.

REFERENCES

1. Anokwa, Y., Smyth, T., Ramachandran, D., Sherwani, J., Schwartzman, Y., Luk, R., Ho, M., Moraveji, N., and DeRenzi, B. Stories from the Field: Reflections on HCI4D Experiences. *ITID 5*, 4 (2009).
2. Asch, S. Effects of Group Pressure Upon the Modification and Distortion of Judgements. *In H. Guetzkow (ed.) Groups, Leadership, and Men* (1951).
3. Bernhart, M., Wiadnyana, I., Wihardjo, H., and Pohan, I. Patient Satisfaction in Developing Countries. *Social Science and Medicine 48* (1999).
4. Bignami-Van Assche, S., Reniers, G., and Weinreb, A. An Assessment of the KDICP and MDICP Data Quality. *Demographic Research S1*, 2 (2002).
5. Brown, B., Reeves, S., and Sherwood, S. Into the Wild: Challenges and Opportunities for Field Trial Methods. In *CHI* (2005).
6. Chavan, A. Another Culture, Another Method. In *CHI* (2005).
7. Czerwinski, M., Horvitz, E., and Cutrell, E. Subjective Duration Assessment: An Implicit Probe for Software Usability. In *IHM-HCI Conference* (2001).
8. Davis, R., Couper, M., Janz, N., Caldwell, C., and Resnicow, K. Interviewer Effects in Public Health Surveys. *Health Education Research 25*, 1 (2010).
9. Evers, V. Cross-Cultural Understanding of Metaphors in Interface Design. In *Proc. Cultural Attitudes towards Technology and Communication* (1998).
10. Evers, V., and Day, D. The Role of Culture in Interface Acceptance. In *INTERACT'97* (1997).
11. Ho, M., Smyth, T., Kam, M., and Dearden, A. Human-Computer Interaction for Development: The Past, Present, and Future. *ITID 5*, 4 (2009).
12. Irani, L. HCI on the Move: Methods, Culture, Values. In *CHI Extended Abstracts* (2010).
13. Jacobsen, N. The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments. In *Proc. of the Human Factors and Ergonomics Society 42nd Annual Meeting* (1998).
14. Johnson, T., and J, P. Interviewer Effects on Self-Reported Substance Use Among Homeless Persons. *Addict Behavior 19* (1994).
15. Kane, S., Wobbrock, J., and Ladner, R. Usable Gestures for Blind People: Understanding Preference and Performance. In *CHI* (2011).
16. Kirchler, E., and Davis, J. The Influence of Member Status Differences and Task Type on Group Consensus and Member Position Change. *Personality and Social Psychology 51*, 1 (1986).
17. Laney, C., Kaasa, S., Morris, E., Berkowitz, S., Bernstein, D., and Loftus, E. The Red Herring technique: A Methodological Response to the Problem of Demand Characteristics. *Psychological Research 72* (1962).
18. Ledlie, J. Huzzah for my Thing: Evaluating a Pilot of a Mobile Service in Kenya. *Qual Meets Quant, London, UK* (2010).
19. Mackay, W. Triangulation within and across HCI disciplines. *Human-Computer Interaction 13*, 3 (1998).
20. Milgram, S. Behavioral Study of Obedience. *Abnormal and Social Psychology 67*, 4 (1963).
21. Nichols, A., and Maner, J. The Good Subject Effect: Investigating Participant Demand Characteristics. *General Psychology 135* (2008).
22. Orne, M. On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications. *American Psychologist 17* (1962).
23. Paulhus, D. Measurement and Control of Response Bias. *J. P. Robinson, P. R. Shaver and L. S. Wrightsman eds. Measures of personality and social psychological attitudes* (1991).
24. Read, J., and Fine, K. Using Survey Methods for Design and Evaluation in Child Computer Interaction. In *Workshop on Child Computer Interaction: Methodological Research at Interact* (2005).
25. Rosnow, R., Goodstadt, B., Suls, J., and Gitter, G. More on the social psychology of the experiment: When compliance turns to self-defense. *Personality and Social Psychology 27*, 3 (1973).
26. Sawyer, A. Detecting Demand Characteristics in Laboratory Experiments in Consumer Research: The Case of Repetition-Affect Research. *Advances in Consumer Research Volume 02, eds. Mary Jane Schlinger: Association for Consumer Research* (1975).
27. Strodtbeck, F., and Lipinski, R. Becoming First among Equals: Moral Considerations in Jury Foreman Selection. *Personality and Social Psychology 49*, 4 (1985).
28. Vatrapu, R., and Pérez-Quinones, M. Culture and Usability Evaluation: The Effects of Culture in Structured Interviews. *Usability Studies 1* (2006).
29. Weinreb, A. The Limitations of Stranger-Interviewers in Rural Kenya. *American Sociological Review 71* (2006).
30. Williams, J. Interviewer Role Performance: a Further Note on Bias in the Information Interview. *Public Opinion Quarterly 32* (1968).
31. Zizzo, D. Experimenter Demand Effects in Economic Experiments. Available at SSRN: http://ssrn.com/abstract=1163863 (last accessed 09/03/2011), 2008.