Usage of Subjective Scales in Accessibility Research

Shari Trewin
IBM T.J. Watson Research Ctr
Yorktown Heights, NY, USA
trewin@us.ibm.com

Diogo Marques
University of Lisbon
Lisbon, Portugal
dmarques@di.fc.ul.pt

Tiago Guerreiro
University of Lisbon
Lisbon, Portugal
tjvg@di.fc.ul.pt

ABSTRACT

Accessibility research studies often gather subjective responses to technology using Likert-type items, where participants respond to a prompt statement by selecting a position on a labeled response scale. We analyzed recent ASSETS papers, and found that participants in non-anonymous accessibility research studies gave more positive average ratings than those in typical usability studies, especially when responding to questions about a proposed innovation. We further explored potential positive response bias in an experimental study of two telephone information systems, one more usable than the other. We found that participants with visual impairment were less sensitive to usability problems than participants in a typical student sample, and that their subjective ratings didn't correlate as strongly with objective measures of performance. A deeper understanding of the mechanism behind this effect would help researchers to design better accessibility studies, and to interpret subjective ratings with more accuracy.

Categories and Subject Descriptors

H.5.2 [**Information Systems and Presentation**]: User Interfaces—*Evaluation/Methodology*; K.4.2 [**Computers and Society**]: Social Issues—*Assistive technologies for persons with disabilities*

General Terms

Measurement, Experimentation, Human Factors

Keywords

Likert, accessibility, response bias

1. INTRODUCTION

Accessibility researchers often use Likert-type items (Figure 1) to gather feedback on proposed accessibility technologies and to compare technologies. This type of measurement is a well-established tool in HCI research [13, p. 270]. However, Likert-type items can also be prone to positive response bias. Indeed, unrelated accessibility studies with different populations have reported participants giving high usability ratings for a task or system, when comments, observations or objective performance measures suggest otherwise [7, 18, 26].

This system was easy to use

strongly disagree	*disagree*	*neither agree nor disagree*	*agree*	*strongly agree*
(1)	(2)	(3)	(4)	(5)

Figure 1: A Likert item.

Positive bias in accessibility research could arise from many sources. For example, people who volunteer for accessibility research studies may have a very positive attitude, or verbal presentation of questions may make it harder to give negative opinions. If accessibility research studies are prone to positive response bias, this is a phenomenon that accessibility researchers need to be aware of, seek to avoid whenever possible, and allow for when interpreting such data.

This paper investigates positive bias in two ways: 1) by analyzing subjective responses reported in accessibility research papers, and 2) by presenting an experiment that compares subjective responses given by a group of participants with visual impairment, and a group of students with no visual impairment, when performing the same phone-based task using two different systems, one more usable than the other.

We find that Likert-type items are widely used in accessibility research, but Likert scales with balanced sets of items are not. Within-subjects designs with target users are typical, and the use of a comparison group is rare. User ratings of accessibility research technology are higher than typical user ratings in usability studies, while user ratings of existing technologies or situations are lower.

Our study found that participants with visual impairment were less sensitive to usability problems than the student group. We point to several possible explanations for this phenomenon and suggest ways that researchers could reduce the likelihood of bias in accessibility studies.

ASSETS'15, October 26–28, 2015, Lisbon, Portugal.
© 2015 ACM. ISBN 978-1-4503-3400-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2700648.2809867.

2. BACKGROUND

Likert scales [10, 14] measure attitudes by presenting a set of statements and asking the respondent to indicate their level of agreement or disagreement with each statement by selecting from a fixed scale of responses. Each statement and response is a Likert item, illustrated in Figure 1.

The underlying assumption is that peoples' opinions will vary on a continuous dimension, from strongly negative to strongly positive, and survey participants will select the response option that most closely matches their strength of opinion. Each option is assigned a numeric code, typically 1-5. In Likert's original formulation, each response also has a verbal label, but the practice of labeling only the endpoints of the scale is also widespread. In this paper, we will use the term 'Likert-type item' to include items that use partially labeled response scales, and also items that prompt with a question rather than a statement, such as "How easy to use was this system?"

Likert-type items are widely used in human-computer interaction research to measure subjective user experience [13, 11, 15]. Established usability measures such as ISO-9241-9 [9], SUS [3], QUEST [6], and NASA-TLX [8] use Likert-type items. They are appropriate for smaller sample sizes, easy to learn and quick to execute [20]. They can be presented verbally, on paper, or digitally. Sauro and Lewis [21] found strong correlations between prototypical usability metrics, including Likert-type user satisfaction ratings, task times and errors.

In an early meta-analysis of usability studies, Nielsen and Levy [15] also found that subjective preferences gathered with Likert-type items generally corresponded well to objective performance measures. In an analysis of 127 studies they found a mean subjective rating of 3.55 +- 0.12 on a 5-point response scale. Suggesting that this value could be used to calibrate scores in future studies, they noted that "users tend to be polite and give fairly high ratings unless they are really upset with an interface" [15].

As this comment implies, responses can be subject to positive bias, especially in contexts such as laboratory studies where it may seem rude to give a negative or contrary opinion. Being based on a statement rather than a question, Likert-type items are known to be subject to acquiescence bias, in which respondents tend to agree with a statement, regardless of the content. In a classic study, Schuman and Presser [22] found that two groups of individuals showed 57% and 60% agreement with opposite statements, suggesting that the effect size of acquiescence bias can be considerable. In a Likert scale containing many individual items, the impact of acquiescence bias can be reduced by balancing the items, so that the statements do not all represent the same point of view.

The desire to give a response that presents the participant in a favorable light (social desirability) is a related well-known source of bias in psychology questionnaires. Social desirability bias is influenced by the way a question is presented [23, 16]. Computer-based presentation reduces this bias, especially if users are anonymous, alone, and able to backtrack [17].

Characteristics of the experimenter or experimental setup (demand characteristics) can also introduce significant bias. Social status and cultural differences between experimenters and participants have been shown to influence participants' stated preferences to such an extent that two different experimenters performing the same comparison of two systems can find opposite results [5].

In accessibility research studies, Likert-type items are often used to gather opinions on proposed technologies. Some researchers have reported apparent contradictions between subjective ratings and observations made during a session. For example, Gerling, Mandryk and Kalyn [7] studied older adults playing wheelchair-based games. They observed:

"In contrast to the results of the NASA-TLX, items of the ISO questionnaire investigating physical fatigue report low levels of physical demand. These findings need to be interpreted along with observations that were made during the study, which contradict subjective ratings for fatigue that were made through the ISO questionnaire; many participants seemed to be challenged by the system and became increasingly tired throughout the gaming session (heavier breathing, individual comments on physical exertion)."

Trewin et al. [26] measured the time for a group of individuals with cerebral palsy to perform a set of point-and-click tasks using a mouse or trackball, and then asked for a subjective rating of how easy the task was, on a 5-point scale. Three participants who rated the task as 'Easy' or 'Very easy' had average pointing times greater than 10 seconds, and one had an average time of 52.4s.

Another study with older users testing a voice-augmented web browsing system reported that while older adults believed themselves to have completed the tasks faster with voice augmentation, in fact they took longer [18].

Similar apparent contradictions have been noted in usability testing, where 14% of users who fail a task will still give it the highest possible satisfaction rating [19]. Proposed reasons for this mismatch include users giving a value that indicates a strong positive, but intending it to be a strong negative; and users being unaware that the task was not successfully completed.

Accessibility research, by its very nature, may be more prone to this positive bias phenomenon, for a number of reasons:

- Questions are often presented verbally by the researchers themselves, often because a paper-and-pencil presentation would not be accessible. This form of presentation may not only increase demand characteristics, social desirability and acquiescence biases, but it may also result in confusion over the meaning of a numeric value on a scale, if there is no visual prompt to refer to.
- Selection of representative participants for studies is difficult [24] - participants volunteering for accessibility research studies may be individuals with a 'can-do' attitude, or a desire to encourage and support young researchers. The user pool at the University of Dundee [4] is a notable exception, offering researchers a rare opportunity to select participants based on criteria such as age, gender and cognitive abilities. However, some bias may still be present in the willingness of individuals to join a user pool in the first place.
- Accessibility barriers may make it difficult for participants to know whether they have completed a task successfully. For example, in searching for information on a web page using screen reading technology, users may be unaware of important information they have missed.

This paper investigates positive bias in accessibility research studies by examining the accessibility literature, and exploring whether the effect can be observed in a formal experimental setting.

3. REVIEW OF LIKERT-TYPE ITEMS IN ASSETS PAPERS

3.1 Goals

To provide a deeper understanding of the use of Likert-type items in accessibility research, we examined accessibility research studies presented as full papers at ASSETS within the last 5 years. Our goals were to learn:

- How frequently such items are used
- What target populations they are used with
- How they are typically administered
- What experimental designs they are used in
- Whether there is evidence of positive bias exceeding the typical bias identified in HCI literature

3.2 Method

We examined full papers published in ASSETS between 2010 and 2014, inclusive, and identified papers containing target user evaluations of new or existing technologies. We included studies where participants evaluated artifacts produced by different technologies (for example sign language avatars, or videos created with different compression techniques), and excluded studies that only surveyed opinions without presenting any artifact for evaluation, such as a web-based survey of experiences with computer game accessibility. We did not count studies targeted exclusively at non-disabled users (e.g. crowdworkers). We included studies with older adults, and studies of sign language users, although the latter often include hearing individuals who are fluent in sign language.

For each study, we identified instances of the use of Likert-type items to gather subjective judgments, and recorded details of the experimental methodology, including the meaning of the question, whether the question was drawn from an established standard measurement instrument, how it was administered, the study design, the number of participants, participant demographics, and responses to the questions. Online studies were assumed to have anonymous response, while studies conducted in person by researchers were non-anonymous.

In analyzing the wording of items, we included all items that were described, even if no results were reported. In four papers, responses to individual questions were only given in a figure. We estimated the response values for 45 items from figures.

We normalized all response values to a 5-point response scale where 1 indicates a very negative opinion and 5 a very positive opinion. This consisted of: scaling 7-point scale values (and other scales) to the equivalent 5-point value; inverting values for reverse coded questions (where a higher value meant a more negative opinion); and shifting values on scales where the neutral point was coded as 0. We used mean response values where available, falling back on median values where no mean was provided.

Table 1: Characteristics of 51 studies published at ASSETS 2010-2014 that used Likert-type items.

Population	No. of studies	Mean/ median pts.	% with anon. responses	% within subjects	% between subjects
Visual impairment	22	10/08	0	50	14
Sign language users	8	63/19	38	88	0
Deaf or hard of hearing	4	53/23	25	100	0
Physical impairment	5	12/14	0	40	0
Older adults	6	10/10	0	66	0
Other populations	6	22/13	0	50	0
All	51	23/11	6	59	4

The number of questions within a study is dependent on the number of conditions being compared, and the range of opinions sought. This means that some studies have more influence on the analysis than others. We did not combine questions within a study, because many studies ask a range of very different questions. Thus, each individual question has equal weight.

3.3 Findings

As in HCI studies, subjective Likert-type items are widely used in accessibility research: 42% of papers included the use of Likert-type items to gather subjective opinions from participants.

Of the 135 papers published between 2010 and 2014, 101 included some form of study involving people facing accessibility challenges. Of these, 51 (50%) used Likert-type items to gather subjective opinions from participants. Table 1 summarizes the characteristics of these 51 studies, organized according to the target user population. One study that included participants with visual and physical impairments is represented as a physical impairment population.

Sample sizes ranged from case studies of three individuals up to an online study with 317 respondents [12], with a median value of 12 participants (mean 23 and standard deviation 48).

Only 8% of these studies (4 studies) had anonymous submission of user ratings, and all of these were studies involving sign language users, or deaf or hard of hearing individuals. In the remainder of studies, participants are assumed to have met with researchers in person.

Many papers did not explicitly state how the questions were presented. Nine studies described paper or electronic administration, all others are assumed to be verbal or signed. In studies with visually impaired participants (almost half of the studies), verbal presentation is the most efficient and natural way to gather such feedback, since paper and pencil questionnaires are inaccessible, and online questionnaires would require a system set up with the assistive technologies that the participant is familiar with.

Most of the studies using Likert-type items (59%) are within-subjects comparative studies. In our sample, only

6 studies collected Likert-type responses from populations with and without access issues responding to the same technology. As Sears and Hanson [24] observe, when testing an intervention designed for a specific population, a comparison group may not be appropriate.

In the 51 papers that reported using Likert-type items, 417 individual items were described, 21 of which were from control group respondents. Response values were reported for 369 of these items. Response scales varied from 4-point scales up to 21 points (NASA-TLX). 66% of items used a 5-point response scale, and 16% used a 7-point scale.

Most items (94%) were custom-designed. Only 27 (6%) were part of a previously established usability measure. The validated measures used were NASA-TLX (7 items) [8], ISO-9241-9 (4 items) [9], QUEST (1 item) [6], DEEP (8 items) [27], SUS (2 items) [3], and Browsing experience (2 items) [1], though the whole measure was not always used.

It was not possible to rigorously analyze the potential for acquiescence bias, or the use of balanced sets of items (both positively and negatively worded), because many papers did not include the precise wording of their Likert-type items. Overall, 75% of the items were described with positive language (e.g. "X is easy to use"). 9% were presented with neutral language (e.g. "experience using X"), and 16% were described with negative language (e.g. "X was confusing"). One online study [25] explicitly considered the effect of valence in wording, using two groups who were asked to respond to the items "I found the video easy to comprehend" and "I found the video difficult to comprehend". Consistent with acquiescence bias, the results for the two groups did not show an inverse relationship, and were analyzed separately.

Excluding responses from control groups, the overall mean of Likert ratings normalized to a 5-point scale where 5 is the most positive response was 3.64 for studies with non-anonymous response modes, which is slightly higher than the mean value of 3.55 reported by Nielsen and Levy [15]. Anonymous responses from the four online studies are lower, at 2.61. This low average rating is unsurprising, because three of these studies (56 items) explored sign language intelligibility in video and included conditions with deliberately degraded video quality. We cannot therefore make a fair comparison between anonymous and non-anonymous responses.

Responses to questions that were about a proposed innovation were rated even more highly (Mean = 3.74), in comparison to a mean rating of 2.9 for other questions.

These findings are consistent with positive response bias for in-person studies and items relating to proposed innovations. The lack of anonymity, verbal format and use of predominantly positive language are likely to introduce positive bias in responses. However, these papers represent the best 25-30% of papers submitted to the ASSETS conference, and are very likely to describe successful solutions to access challenges. It is possible that the proposed innovations were simply very effective.

4. USER STUDY

4.1 Design

Based on the analysis of papers of the last 5 years of ASSETS proceedings, coupled with our own observations while conducting user studies, we reasoned that if there is, in fact, an inflation in positive reports, it could be revealed by an experiment. Hence, we designed a study where we asked people to evaluate two systems with subjective scales, but purposefully introduced usability problems in one of them. We recruited two convenience samples typical of accessibility research, one of 16 people with visual impairments, and another of 16 students with no visual impairments, as a control. The experiment we will describe next has, therefore, a 2x2 design, with one factor being the population from which the sample is drawn (between-subjects), and the other the system that was evaluated (within-subjects). We will henceforth refer to the levels of factor Population as VI and Student, and to the levels of factor System as Usable and Degraded, with Degraded being the one engineered to have usability problems.

The systems we created are automated call-answering services that people dial in to on the telephone and then navigate the options recited by a recorded voice using the key buttons. We chose this kind of system for two reasons. First, because it is usable by both user groups. Second, because it is something that every participant was likely to have encountered before, and yet could be understood as an innovation proposed by the researchers conducting the study - as a new and improved call-answering service. Had we tried to evaluate an innovation that was very unfamiliar to participants, other confounds could have been introduced, such as exaggerated excitement for new technologies, or, conversely, a resistance to change.

4.2 Research questions

If there is stronger positive response bias in the visually impaired group, it should be true that it rates the systems more positively when asked. It should also be true that, in this group, the ratings are less sensitive to usability problems. We hence formulate the following research questions:

RQ1 Do visually impaired participants offer higher subjective scores?

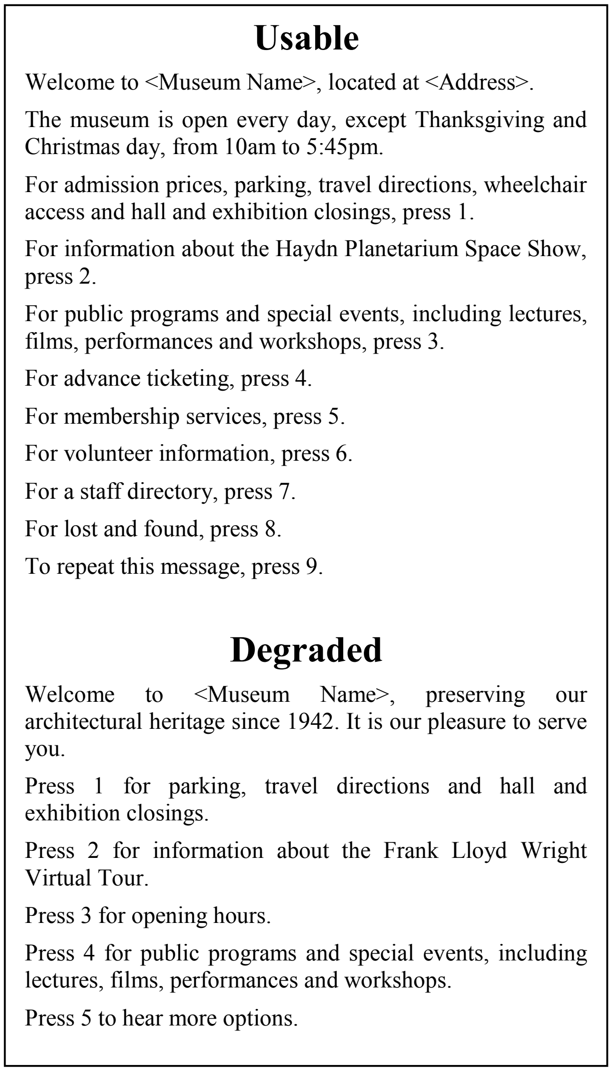
RQ2 Is the difference between the subjective scores they assign to the two systems smaller than the difference assigned by students?

RQ3 Is the relationship between the subjective scores and objective usability measures weaker for the group with visual impairment?

4.3 Participants

We recruited 16 participants with visual impairment (the VI group) from a vocational learning center for blind people. There were 9 males and 7 females, and their average age was 45.2 years old, standard deviation (henceforth SD) 11.2, and range 23-64. Thirteen participants were legally blind, and three had low vision (the criteria for inclusion being 1. usually using cane for outside mobility and 2. usually using screen readers in computing devices). Among the blind participants, the average time of onset was 27.5 years ago (8-52, SD 16.0). Six participants reported using telephone-based information systems weekly, 6 monthly, 6 rarely, and 3 never. When participants reported never having used such a system, an expanded explanation was offered.

For the control group, we recruited 16 graduate school students with no declared visual impairment from a university campus. There were 12 males and 4 females, and their average age was 27.2, SD 7.1, range 23-53. One participant



Usable

Welcome to <Museum Name>, located at <Address>.

The museum is open every day, except Thanksgiving and Christmas day, from 10am to 5:45pm.

For admission prices, parking, travel directions, wheelchair access and hall and exhibition closings, press 1.

For information about the Haydn Planetarium Space Show, press 2.

For public programs and special events, including lectures, films, performances and workshops, press 3.

For advance ticketing, press 4.

For membership services, press 5.

For volunteer information, press 6.

For a staff directory, press 7.

For lost and found, press 8.

To repeat this message, press 9.

Degraded

Welcome to <Museum Name>, preserving our architectural heritage since 1942. It is our pleasure to serve you.

Press 1 for parking, travel directions and hall and exhibition closings.

Press 2 for information about the Frank Lloyd Wright Virtual Tour.

Press 3 for opening hours.

Press 4 for public programs and special events, including lectures, films, performances and workshops.

Press 5 to hear more options.

Figure 2: Main menu of Usable and Degraded telephone systems.

reported using telephone-based information systems weekly, and 6 monthly, and 10 rarely. The lower frequency might be explained by this sample using internet services instead of calling, as was mentioned by some of the participants.

There is considerable difference in the demographic make-up of the groups. We are aware that in some studies in which the population is a factor, the control sample is matched to the sample of interest. As we have not done that, we cannot exclude the possibility that differences between population groups are explained by demographics. However, we point out that even if we matched age and gender, other personal characteristics, which we didn't collect or measure, could be the explaining variable. Indeed, if an effect exists, it follows that it is caused by something, or some things, like personality traits, which perhaps could be measured. As we were interested in first finding evidence that the effect exists, we opted for what is a typical student sample of HCI research, and we leave the question of its causes open.

4.4 Apparatus

We sought to define tasks that could be performed in the same manner by both groups, and would be equally easy for both groups, We selected a set of auditory tasks that would be familiar to both user groups: calling an automated telephone information system for specific information. Our two designs were both modeled directly on the main menu of the American Museum of Natural History's telephone information system. The Usable system closely follows the content, structure, prompt style and navigation method of the original system. User actions are always described AFTER describing the option ('For xxx, press 2'). Navigation in this system centers around a single main menu, and users are automatically returned to this main menu if they make no selection in a submenu.

The Degraded system was an edited version of the same system, with the following changes:

- Prompts were reorganized so as to provide the user action BEFORE describing the option. ("Press 2 for xxx"). This means that users must retain the number in working memory while listening to the rest of the prompt. Since some options are long (up to 17 words), this should make the system more difficult to use.

- Additional preamble was added before presenting the menu options.

- One item of useful information (opening times) was moved from the preamble into a submenu.

- Menu options were reordered.

- One option ('buy a ticket') was moved into a submenu.

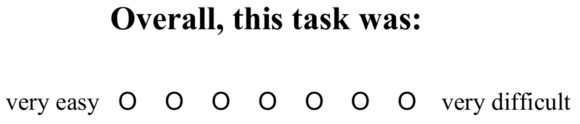
- One option ('lost and found') was removed, and replaced with a generic option in a submenu ('all other enquiries')

- Navigation was hierarchical, and information about how to return to the previous menu was added at the end of the sublevel menus.

The main menu for each system is shown in English in Figure 2. For the purposes of the experiment, both systems were translated into Portuguese and local museum names were used in place of the terms Usable and Degraded.

Each system was simulated through a desktop application, with which a user could 'call' the museum, and navigate the menus using the keyboard numpad. The menus were presented as pre-recorded human speech, to reduce effects due to the blind participants being more familiar with TTS voices than the comparison group.

We used Sauro's Single Ease Question [20], shown in Figure 3, to gather subjective feedback. This question has been validated in comparative studies [20] and performed well. In contrast to many items in the literature, it is neutrally worded. Responses are given on a 7-point scale where one extreme (1) is 'very difficult' and the other (7) is 'very easy'. For verbal administration, the question was phrased "From 1 to 7, do you think this task was easy or hard to perform, where 1 is very difficult and 7 is very easy".



Overall, this task was:

very easy ○ ○ ○ ○ ○ ○ ○ very difficult

Figure 3: Sauro's Single Ease Question.

4.5 Procedure

Sessions with participants with visual impairment took place in the learning center. Sessions with students took place at their university. After giving informed consent, participants provided demographic information and were then asked to complete the same four tasks with each system. The objective for each task was:

Task 1 Find out where the wheelchair access is located.

Task 2 Find out if the museum is open today.

Task 3 Get to the Lost and Found department.

Task 4 Find out the admission price for a senior.

After stating the objective, a moderator dialed the number. Participants were instructed that, from that moment on, the moderator would not intervene at all, until the participant hung up.

They were further instructed that they could try calling as many times as they wanted, with the assistance of the moderator, until quitting or achieving the stated goal, as this is a reasonable strategy on real call-answering systems. To verify that the goal was achieved, participants were instructed to, after hanging up by themselves, verbally recount the information they were tasked to retrieve. For task 3, where there was no information to retrieve, success was assessed based on participants reaching the right option. Each participant completed all four tasks, first on one system, then on the other. The order of systems was counterbalanced. The order of tasks was kept constant, as the research questions do not call for comparisons among them. The systems were referred to using the museum names, so as not to imply any ordering between them. After each task, the experimenter asked the Ease question, presented verbally. The participant's spoken responses were recorded. Participants were asked to give a numerical response to the 'Ease' question. Immediately after they responded, the corresponding meaning of the value was reflected back to them, to confirm their intention. This was to eliminate any possible confusion about the meaning of the numerical values. The moderator used form fields in the desktop application to record if the task was completed successfully and the subjective score. The application automatically recorded, for each task, the times at which calls were dialed and hung-up, and the sequence of options selected by the participant.

4.6 Results

In the following analysis, we use standard t-tests, which are paired when comparing within-subjects measurements, and unpaired otherwise. We adjust the significance level with the Bonferroni correction when comparing *Population x System* combinations.

Table 2: Average answer to the Single Ease Question per group and per system.

	VI	Student	
Usable	6.6 (SD=1.0)	6.8 (SD=0.6)	6.7 (SD=0.8)
Degraded	6.4 (SD=0.9)	6.0 (SD=1.2)	6.2 (SD=1.1)
	6.5 (SD=0.9)	6.4 (SD=1.1)	

Table 3: Average task completion time per group and per system.

	VI	Student	
Usable	56s (SD=37)	45s (SD=17)	52s (SD=30)
Degraded	79s (SD=49)	62s (SD=28)	71s (SD=41)
	68s(SD=45)	54s (SD=24)	

4.6.1 Subjective ease

All participants ultimately completed all tasks successfully. Table 2 summarizes the responses to the 'Ease' question. From the overall ratings, and our observations during study sessions, there is a ceiling affecting the effect sizes. We had hoped that the usability problems in the Degraded system would cause more difficulties than they actually did. Overall, participants in the VI group offered only slightly higher subjective scores (6.5 vs. 6.4 on average). The effect size being small (Cohen's d=0.01, with 0.2 being considered a standard small effect), our sample size doesn't provide adequate power to detect a difference, and hence can't say that the scores are significantly higher (t=0.72, p>0.05). We can't therefore answer the first research question conclusively, as to whether VI participants offer higher subjective scores.

Still, the differences in subjective scores by system are informative as to possible differences in sensitivity to usability problems. As expected, taking the two groups of participants together, tasks on the Usable system were on average rated as being easier to perform than in the Degraded system (0.47 mean difference, t = 4.70, p<0.00001, d=0.48). As would be expected if there was a response bias, despite the fact that the Degraded system had usability problems, it wasn't rated much worse by visually-impaired participants (0.13 mean difference), and this difference is not significant (t=1.07, p>0.025, d=0.14). The students, in contrast, offered significantly lower scores to the Degraded system (0.86 mean difference, t=5.84, p<0.0001, d=0.88). Taking the difference in ratings that participants assigned to the same task in the two systems as a measurement of sensitivity to usability problems, we find that indeed the visually-impaired group was less sensitive (t=3.79, p=0.0001) and that this effect is sizable (d=0.68). We thus find support to respond positively to the second research question: the difference between the subjective scores visually-impaired participants assign to the two systems is smaller than the difference assigned by students.

4.6.2 Relationship with objective measure

Table 3 summarizes the task completion times. We define task completion time as the interval between the first time the phone line was opened, and the last time a call was ended, within a task. Although we are not testing differences in measured performance, we can observe that, as expected, tasks in the Degraded system took on average more time than in the Usable system. We can also observe that the VI group took on average more time than the student group, which can be at least partially explained by the added effort of seeking the right key on a keyboard using only touch. We computed the correlation between the task completion times and the subjective score for each group. We find that for the student group, the relationship between the two measures was very strong (Pearson's r (117 d.f.) = 0.70, 95% CI [-0.78,-0.59]). For the VI group, despite the relationship also being strong, it was less so, as would be predicted by the thesis of a positive bias (Pearson's r(134 d.f.) = 0.50, 95% CI [-0.62,-0.36]). The difference in correlation coefficients, despite the ceiling effect in subjective scores, leads us to answer positively to the third and final research question: the relationship between the subjective scores and objective usability measures is weaker for the group with visual impairment.

4.6.3 Examples

To further illustrate the positive bias in VI participants, and as a sanity check to the statistical analysis, we present some instances of mismatch between what we perceived as being users struggling with the system, and the subjective scores they assigned to the task. We represent dialing with the letter D, and hanging-up with the letter H.

- For task 1 in the Degraded system, the optimal sequence of steps was D-1-H. One VI participant executed the sequence D-4-H-D-5-5-1-H, took more than 200 seconds to complete the task, and then rated the easiness at 6 out of seven.
- For the next task, the same participant executed the sequence D-2-5-3-5-3-H, when the optimal was D-3-H. and rated easiness at 7 out of seven.
- For task 3 in the Degraded system, another participant executed the sequence D-2-5-3-5-3-9-8-H, when the optimal would be D-5-8-H, and rated easiness at 7.
- For task 4 in the Degraded system, yet another participant took 117s to complete the task, having taken two more steps than necessary, and rated easiness at 7.

4.7 Discussion

Although we couldn't establish that those in the population of interest consistently rate systems higher, because of a ceiling effect, we did find that their subjective ratings aren't as sensitive to usability problems, and do not correlate as well with objective measures of performance. Plausible explanations for this finding could include:

- The task was easier for the participants with visual impairment
- Demand characteristics, acquiescence or social desirability bias affected the two groups differently
- The experience of accessibility barriers creates a low usability baseline
- The participants with visual impairment had a generally positive outlook
- The participants with visual impairment wanted to encourage the researchers
- The difference is the result of other unmatched factors such as participant age, study location and recruitment methods.

Further work is needed to unpack these possible factors. Our experiment provides some initial insights. Although we strove to select a task that would be equally easy for the two groups, it is possible that any auditory task might be easier for a person with long-established visual impairment who is used to receiving information auditorially. However, although the VI group did report greater use of telephone information systems, they did not complete the tasks more quickly than the students. They had a lower correlation between their task performance and subjective evaluation, suggesting that this was not the case here. It is still possible that the VI group's familiarity with such systems led to their being less affected by the usability problems we introduced.

Finally, participants' conception of what is 'easy' is inevitably shaped by their daily life experiences. People living with a disability face access barriers that others do not. The tasks here were about accessing information, which can involve access barriers for people with visual impairment. Thus, the participants' prior experiences may have created very different baselines from which they estimated ease of use. Our Usable and Degraded systems both provided the information needed to complete the tasks in an accessible form. Thus, the VI group may have rated them both as 'very easy' in comparison to other tasks they have encountered. Similar arguments would apply to other accessibility groups. For example, for individuals with dexterity impairment, a selection task may be 'very easy' if the target is much larger than typical targets, even if it takes a long time to make the selection and errors are made.

When subjective user ratings seem at odds with other measures, this can be an important indicator that the priorities of individuals with a disability differ from those assumed by researchers who may not have experience of that disability. However, to properly interpret such ratings, researchers need an understanding of how the factors above can also influence responses.

5. IMPLICATIONS FOR ACCESSIBILITY RESEARCH

Our experimental study followed the design of a typical accessibility study - a within-subjects comparison of technologies conducted in person with verbal question administration and a small number of participants. We added a second factor of two different participant groups, so as to compare ratings for the VI group and the Student group.

If a tendency to provide higher usability ratings than typical HCI users were confirmed in further work, it would suggest that accessibility studies are more likely to run into ceiling effects. This would make it difficult to find significant differences between conditions (as in our experiment), especially with small numbers of participants. It would also mean that baseline values for established usability scales may not be directly transferable to accessibility technologies. For example, on the System Usability Scale (SUS), a score above 70 is considered good, based on experience from hundreds of usability studies [2]. A higher baseline value may be more appropriate for access technologies.

Accessibility researchers can take steps in their experimental designs to avoid conditions that have been shown to lead to biased responses, as discussed below.

5.1 Use validated scales

Very few of the studies in our sample used established measures such as NASA-TLX or SUS. Such measures have often undergone extensive development to select appropriate wording, balance the set of items, and confirm reliability. Accessibility researchers could benefit from this work.

5.2 Use balanced or neutrally-worded prompt statements

Few accessibility studies have enough participants to support the use of two groups responding to oppositely worded statements, as in [25]. However, researchers could use a balanced mix of positive and negative items to mitigate acquiescence bias, or use neutral wording, as in the single ease question ("Overall, this task was:").

5.3 Consider alternatives to verbal presentation

To reduce social desirability reporting, or other bias introduced by verbal presentation, consider whether questionnaires could be administered in a more private fashion, for instance through an online format.

5.4 Use response labels to avoid confusion

Where participants do give responses verbally with no visual prompt, using response labels (e.g. "strongly agree" or "difficult") would eliminate the possibility of confusion over the mapping from labels to numbers. Where numerical responses are requested, researchers could translate the number to an appropriate label and reflect it back to participants, to confirm their intended response.

5.5 Report presentation method precisely

Many papers described that they had "administered a questionnaire", but did not state how the questions were presented. Whatever the experimental design, it is good practice to report the specific verbiage used in subjective response questions, and to explicitly state the mode of presentation and response. This allows other researchers to better understand the potential for bias.

5.6 Interpret positive ratings with care

If steps are taken to protect against bias, researchers can more confidently learn from subjective ratings, and avoid a ceiling effect that masks genuine usability differences between systems. Researchers receiving strong positive ratings for novel technologies from participants with disabilities may be wise to consider potential sources of bias in interpreting the results.

6. CONCLUSIONS

Likert-type subjective ratings are commonly used in accessibility research, in formats that may introduce positive response bias. Our study of two telephone information systems, one more usable than the other, found that a group of participants with visual impairment were less sensitive to usability problems than a group of students without disabilities, and their responses had lower correlation with task completion times. Their high ratings produced a ceiling effect that masked the usability differences found by the student group. Further work is needed to establish whether positive bias is at play. Accessibility researchers may be able to reduce positive bias by using electronic administration methods and balanced sets of Likert-type items, among other techniques. Potential positive bias should be taken into account when interpreting Likert-type responses, and applying subjective measures from HCI in accessibility research.

7. ACKNOWLEDGMENTS

This research was partially funded by FCT (UID/CEC/00408/2013, SFRH/BD/98527/2013). We thank Fundação Sain and Carlos Bastardo for their patronage, Erin Brady, and the study participants.

8. REFERENCES

[1] J. S. Ahuja and J. Webster. Perceived disorientation: An examination of a new measure to assess web design effectiveness. *Interacting with Computers*, 14(1):15–29, Dec. 2001.

[2] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[3] J. Brooke. SUS : A retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[4] M. Dee and V. L. Hanson. A large user pool for accessibility research with representative users. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '14*, pages 35–42, New York, New York, USA, Oct. 2014. ACM Press.

[5] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies. "Yours is better!": participant response bias in HCI. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, page 1321, New York, New York, USA, May 2012. ACM Press.

[6] L. Demers, R. Weiss-Lambrou, and B. Ska. Development of the Quebec User Evaluation of Satisfaction with assistive Technology (QUEST). *Assistive Technology*, 8(1):3–13, June 1996.

[7] K. M. Gerling, R. L. Mandryk, and M. R. Kalyn. Wheelchair-based game design for older adults. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '13*, pages 1–8, New York, New York, USA, 2013. ACM Press.

[8] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, volume 52, pages 139–183. 1988.

[9] ISO. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 9: Requirements for non-keyboard input devices. In *ISO 9241*. International Organization for Standardization, 2000.

[10] R. Johns. Likert items and scales. In *Survey Question Bank: Methods Fact Sheet 1*. 2000.

[11] M. C. Kaptein, C. Nass, and P. Markopoulos. Powerful and consistent analysis of Likert-type rating scales. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, page 2391, New York, New York, USA, 2010. ACM Press.

[12] M. Kipp, Q. Nguyen, A. Heloir, and S. Matthes. Assessing the deaf user perspective on sign language avatars. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '11*, page 107, New York, New York, USA, 2011. ACM Press.

[13] J. Lazar, J. H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. John Wiley & Sons, 2010.

[14] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55, 1932.

[15] J. Nielsen and J. Levy. Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4):66–75, Apr. 1994.

[16] D. L. Paulhus. Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46(3):598–609, 1984.

[17] W. L. Richman, S. Kiesler, S. Weisband, and F. Drasgow. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5):754–775, 1999.

[18] D. Sato, M. Kobayashi, H. Takagi, C. Asakawa, and J. Tanaka. How voice augmentation supports elderly web users. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '11*, page 155, New York, New York, USA, 2011. ACM Press.

[19] J. Sauro. Do users fail a task and still rate it as easy?, 2009. `http://www.measuringu.com/failed-sat.php` [accessed 2015-May-02].

[20] J. Sauro and J. S. Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, page 1599, New York, New York, USA, 2009. ACM Press.

[21] J. Sauro and J. R. Lewis. Correlations among prototypical usability metrics. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, page 1609, New York, New York, USA, 2009. ACM Press.

[22] H. Schuman and S. Presser. The acquiescence quagmire. In *Questions and Answers in Attitude Surveys*. 1981.

[23] N. Schwarz, F. Strack, H.-J. Hippler, and G. Bishop. The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5(3):193–212, May 1991.

[24] A. Sears and V. L. Hanson. Representing users in accessibility research. *ACM Transactions on Accessible Computing*, 4(2):1–6, Mar. 2012.

[25] J. J. Tran, J. Kim, J. Chon, E. A. Riskin, R. E. Ladner, and J. O. Wobbrock. Evaluating quality and comprehension of real-time sign language video on mobile phones. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '11*, page 115, New York, New York, USA, 2011. ACM Press.

[26] S. Trewin, M. Laff, V. Hanson, and A. Cavender. Exploring visual and motor accessibility in navigating a virtual world. *ACM Transactions on Accessible Computing*, 2(2):1–35, June 2009.

[27] T. Yang, J. Linder, and D. Bolchini. DEEP: Design-oriented evaluation of perceived usability. *International Journal of Human-Computer Interaction*, 28(5):308–346, May 2012.