Utilizing Employees as Usability Participants: Exploring When and When Not to Leverage Your Coworkers

Joanne Locascio[1], Rushil Khurana[2], Yan He[3], Jofish Kaye[3]
[1:] PubMatic, Redwood City, CA, USA. jgdemanuele@yahoo.com
[2:] Human Computer Interaction Institute, CMU, Pittsburgh, PA, USA. rushil@cmu.edu
[3:] Yahoo, Sunnyvale, CA, USA. heyan | jofish @yahoo-inc.com

ABSTRACT
Usability testing is an everyday practice for usability professionals in corporations. But, as in all experimental situations, *who* you study can be as important as *what* you study. In this Note we explore a common practice in the corporation: experimenting on the company's employees. While fellow employees can be convenient and avoid issues such as confidentiality, we use two usability studies of mobile and web applications to show that employees spend less time-on-task on competitor websites than non-employees. Non-employees reliably rate competitor websites and apps *higher* than employees on both usability (on the 10-question SUS scale) and ease of use (on the 1-question SEQ scale). We conclude with recommendations for best practices for usability testing in the corporation.

Author Keywords
Usability; User Research; Recruiting; Participants

ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/Methodology

INTRODUCTION
Usability testing has received attention in the literature for not only the rapid growth and demand within organizations [30], but also the ability to improve product development [24] at relatively low cost [4]. While many considerations occur in the course of a usability study, arguably one of the most important is identifying the right participants [7, 14]. Recruiting the wrong participants has been described as "worse than useless" as it instills artificial confidence in the results and may result in improper data and recommendations [15, 1]. Participants within a study must generalize to the population of users [6], which is often a compromise in usability testing [4]. Given the implications of poor sampling and the widespread use of usability testing, further attention to this topic is warranted.

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858047

Many sampling or recruiting methods have been detailed in the literature including non-probability sampling techniques, such as convenience and purposive sampling [25, 11]. At the same time, usability studies need to be executed quickly enough to maintain relevance in agile environments [20] while attempting to adhere to best practices and rigor [31]. Given the fast paced nature of usability testing coupled with the existing practices of convenience sampling in the social sciences and qualitative approaches, it is understandable why convenience sampling is the most widely used sampling technique in usability testing [19, 27].

Recent years have seen several discussions regarding the importance of sound sampling techniques to ensure appropriately diverse users. The field of psychology recently had an active discussion around the validity of choosing subjects from Western, educated, industrialized, rich and democratic ("WEIRD") societies for psychological studies [13; see also the extensive series of follow up articles in the same publication]. Additionally, there has been significant critique of the common practice of the use of undergraduate students – again, frequently at Western universities – as subjects in social science research [21, 9], all of which suggest sampling practices play an important role in the phenomena under study.

Corporate usability studies also require strict adherence to sampling techniques and biases as demonstrated by the teachings and practice of participant recruiting screeners. However, corporate usability studies may have a particular set of requirements and concerns, such as confidentiality and attempts to keep product developments out of the general marketplace until formally announced with strict policies should this information be released. Literature comparing contract based employees and full time employees in a variety of domains is abundant [17, 32] with conclusions that there, are indeed, differences. However, an exhaustive literature review yielded no established findings suggesting that use of employees instead of the general public (non-affiliated) would yield a difference in usability study findings, suggesting a potential avenue of research

To better understand the prevalence of this practice, the authors placed a two-question survey on LinkedIn, Twitter and Facebook, recruiting self-identified usability professionals to answer a two-question survey. The first question asked:

"Some UX professionals perform user testing on their company's own employees for many reasons, including confidentiality and convenience. Thinking about usability studies that focus on products developed for use by external (non-employees) participants, about how much time do you conduct usability tests using employees as participants?"

The second question invited open-ended responses:

"Let us know if you have any other thoughts related to this topic."

We gathered 104 responses and 55 open-ended responses in 9 days. Results are presented in Figure 1. The data suggest this is a fairly frequent practice: 56% of our 106 respondents told us they tested on employees at least some of the time. Looking at the open-ended responses, 40% included discussion to the effect that this is not an ideal practice as it introduces an element of "bias" and lacks representation of actual users, of which many responses applied to those who admitted to using employees. This suggests bias is, at the very least, suspected when using employees. The second most frequent theme (19%) suggested utilizing employees as pilots in preparation for a larger external study. Confidentiality (13%) and convenience (7%) as reasons for using employees were also themes observed in the open-ended responses.

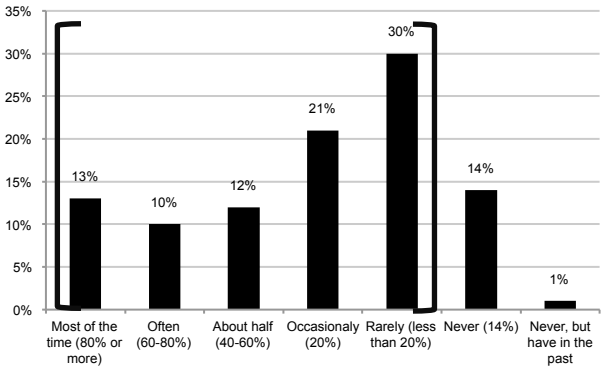

Figure 1. Self-reported frequency of using employees as usability participants for external facing products. The brackets indicate the 85% of usability professionals who sometimes use employees for testing purposes.

The objective of this study is to determine the validity of using employees as a sampling practice. We aim to report investigations uncovering differences, if any, that exist between product ratings of internal and external products from employees and non-employees. While other studies have made similar comparisons such as culture [16], gender [10] and age [18], to date no other study investigating the validity of employees as usability participants exists.

METHODS

Tools

Two standardized questionnaires were utilized in this study, the SUS (System Usability Scale) and the SEQ (Single Ease Question), for two primary reasons. First, these scales have established reliability and validity as established by statistical methods (see below). Second, quantitative measures of usability and self-reported experiences allow for statistically driven group comparisons.

The System Usability Scale (SUS) was designed by John Brooke as a measure to assess the ISO standards of usability, which include effectiveness (ability to execute a task), efficiency (the amount of effort to execute a task) and satisfaction [5]. As Brooke [5] explains, good usability measures require minimal time and mental effort investments from the participants, such that a measure can be completed quickly during a usability evaluation. The SUS is a ten-item questionnaire aimed at "giving a global view of subjective assessments of usability" [5]. Much empirical evidence exists supporting the validity and reliability of the SUS [3, 4, 26].

The Single Ease Question (SEQ) is one question asking the participants to rate the overall difficulty of task on a 7-point Likert scale and is purported to measure ease of use [28]. It performs equally or in some cases better than other standardized measures like the Subjective Mental Effort Questionnaire or the Usability Magnitude Estimation, both of which are more complex in nature [28]. Given the established reliability and validity of the SEQ and its ease of administration (one item), we favored the SEQ as a measure to indicate ease of use.

For study two, where we were not asking for participants to think-aloud, we additionally measured time-on-task: the duration of time it took users to accomplish each pre-determined task on each website. Time was measured via a stop-watch. Sessions were recorded to help validate initial time on task measurements.

Studies

Data collection was carried out in two independent studies looking at mobile applications and at websites, all of which were considered mature products and publically accessible. Both studies were conducted in usability labs located at the company's headquarters with each session lasting approximately 60 minutes. A total of 32 participants (16 employees and 16 externals) contributed to the study. Employees that participated in the study were not part of the product development for the websites and app tested, to eliminate bias.

The first study focused on iOS mobile applications in which 12 participants (6 employees and 6 externals) completed 8 pre-determined tasks on both the company's public app and a competitor's public app, alternating the presentation of apps between participants to remove order bias. These tasks were simple instructions that were designed to be

executable on both the company's and competitor apps, such as "find a recent article from the Entertainment category" or "find the top stories for today". Participants were asked to think-aloud, which is a common usability approach [15, 19]. Error rates were collected and scored success if the participant completed the task accurately. After each of the 8 tasks the SEQ was administered, which generated 8 SEQ scores per participant for each of the company and competitor apps, totaling 16 SEQ scores for each participant. The SUS was completed after all tasks were executed for each of the company's and the competitor's apps, generating 2 SUS scores per participant.

The second study followed a similar design, except for a focus on websites viewed through a desktop computer instead of a phone. 20 participants (10 employees and 10 externals) completed 8 pre-determined tasks on both the company's website and a competitor's. Again, tasks were designed such that that they could be completed on both the company's and the competitors' websites. For example, tasks include "find a score from Wednesday's baseball game" and "find an article featuring NASCAR from yesterday". The think-aloud method is generally not advised when time-on-task measurements are observed, therefore we excluded the think-aloud protocol during this round [1, 33] as no emerging trends were identified from the first study. Following study 1, the SEQ was administered after each task, again resulting in 8 SEQ scores for each company and competitor app (totaling 16 per participant). The SUS was administered after all tasks for each of the company and competitor websites were completed, totaling 2 SUS scores per participant.

Time on task was collected for study 2. Following Sauro's methods [29], time on task was determined from the start of a task until the participant indicated he or she gave up or thought the task was complete. Error rates were scored success if the participant did indeed complete the task accurately, independent of the time on task measurement. The SUS was completed for each the company's and competitor's apps. The SEQ was filled out after each of the 8 tasks for both the company's and competitor's apps.

Error rates were coded as pass (1) or fail (0) for each task completed, constituting a binary repeated-measures variable. Therefore, group differences were determined using a General Linear Mixed Model to account for risks with covariance [34]. No significant differences were found between employees and non-employees for both the company's website ($p = 0.67$) and the competitor's site ($p = 0.62$). This increased confidence that the differences between employees and non-employees for the SUS and SEQ were not due to significant group differences in error rates as they did not exist. Therefore, we do not report these details any further.

Participants

A total of 32 participants (16 employees and 16 non-employees) participated in two usability studies. Employees were recruited using a corporate e-mail list asking for participation. External participants were recruited via a survey posted on Craigslist, which has been cited as a common forum for recruiting [2]. Using Craigslist as a recruiting tool has gained attention recently in the literature with one study suggesting there are differences between Craigslist participants and the general population such as gender and income [2].

For this reason we attempted to match the external participants to employees with respect to device usage, gender, age (see Table 3) and self-reported usage of the app and websites. External participants were given a $100 gift card. Due to corporate policy, employees were given a $30 gift card to the corporate store for their participation.

For the first study, all participants self-reported using an iOS device for a minimum of 3 months and the company app or competitor news app at least weekly. For the second study, all participants self-reported using either the company or competitor website at least weekly.

	Employees	Externals
Females	7	9
Males	9	7
Average age	32.8	37.9

Table 1. Demographic data

No significant differences were found for age ($t(30) = 0.07$, $p. = 0.94$). All participants reported using an iOS device for a minimum of 3 months. Results indicate we can safely conclude that participants were matched appropriately for gender and age.

RESULTS

The Student's *t*-test was used to detect differences between groups for the SUS (Table 4). Results showed employees rated the competitor product significantly lower than non-employees, $t(30) = 2.73$, $p = 0.01$. However, the SUS score comparisons for the company product showed no significant differences, $t(30) = 0.20$, $p = 0.85$.

	Employees	Externals
Company	68.6	70.2
Competitor	**67.2**	**80.8**

Table 2.Mean SUS scores.. Competitor scores are significantly different.

Given that SEQ ratings occurred 8 times for each participant, and thus created a hierarchical model, a two-way nested ANOVA was used to measure statistical differences between the employee and non-employee group [12]. Results (Table 5) shows that employees consistently

rated the competitor product lower than non-employees ($F(30) = 9.43$, $p < .001$.), whereas both groups rated the company product the same ($F(30) = 0.88$, $p = 0.35$).

	Employees	Externals
Company	5.07	5.25
Competitor	**5.07**	**5.66**

Table 3. Mean SEQ scores. Competitor ratings are significantly different.

Consistent with previous research, the time on task data was not normally distributed and showed a positive skew [29]. Therefore, a two-way ANOVA using a log-transformed variable for time on task was used. Employees spent significantly less time on the competitor's website than non-employees ($F(18) = 7.05$, $p = 0.009$), but no significant differences were found between employees and externals for the company's website ($F(18) = 2.34$, $p = 0.13$). Table 4 reports geometric means for time on task following established best practices [29].

	Employees	Externals
Company	29.27	34.58
Competitor	**24.06**	**32.59**

Table 4.Geometric mean for time on task. Employees spent significantly less time on the competitor website.

CONCLUSION

We observed statistically significant differences between otherwise similar employees and non-employees, evaluating the same websites and the same apps, suggesting some preliminary conclusions can be drawn from this study. First, employees rated competitor products differently than non-employees. Second, employees spent significantly less time on the competitor website when attempting to complete tasks. And third, both employees and non-employees rated the company's own products similarly.

Our findings indicate employees are not good substitutes for subjects drawn from the general population when the goal includes comparisons to competitor products, but that it may well be valid to use employees for internal usability testing of company's products.

Our findings may be best explained by the concept of employee brand loyalty. Given that both measures of employee's subjective ratings and time on task were significantly different for the competitor products than non-employees, there is strong evidence that some bias is at play. Previous investigations suggest the existence of employee's loyalty to the brand by which they are employed, and can reach up to 70% of a company's employees exhibiting behaviors consistent with brand loyalty [8]. A positive relationship has been shown between an employee's loyalty and their work performance, suggesting the existence of a connection beyond the traditional 9 to 5 boundaries [22]. Furthermore, research has shown that employees who identify with a brand may be a driving force of brand loyalty [23]. Literature regarding employee's position towards competitor products is non-existent; however, our data not only supports the current literature in that a loyalty to the employer's brand may exist, but also employees may share a critical outlook when it comes to subjective ratings of competitor products.

Clearly, this study is limited by the number of cases observed: it studied engagement with two websites and two apps, with a statistically valid but still comparatively small sample. But, this study was one of a kind with no prevailing data from which to build upon, and as such, makes for a unique contribution. Future studies may consider a correlation of job satisfaction with ratings on measures such as the SEQ and SUS. We did not ask that in this study due to the sensitivity of the matter, but encourage further exploration.

Confidentiality was cited as the purpose for using employees in 19% of the open-ended responses in our survey, suggesting the product is not public knowledge and in early stages of the product lifecycle. Future investigations may also consider similar evaluations along various stages of the product lifecycle.

ACKNOWLEDGMENTS

We would like to thank our colleagues who facilitated with the organization and administration of these studies.

REFERENCES

1. William Albert, Thomas Tullis, and Donna Tedesco. 2009. *Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies*. Morgan Kaufmann.
2. Sarah Anderson, Sarah Wandersee, Ariana Arcenas, and Lynn Baumgartner. 2013. "Craigslist samples of convenience: recruiting hard-to-reach populations." Retrieved August 13, 2015 from*: http://fiesta. bren.ucsb.edu/~ sanderson/CraigslistSurvey.*
3. Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An empirical evaluation of the system usability scale. *Intl. J. of Human–Computer Interaction* 24, 6: 574-594.
4. Carol M. Barnum. 2010. *Usability Testing Essentials: Ready, Set...Test!* Elsevier.
5. John Brooke. 1996. SUS – A quick and dirty usability scale. In *Usability Evaluation in Industry*, Patrick W. Jordan et al. (eds.). Taylor & Francis, 189-194.
6. Mohamed A. Bujang, Puzziawati A. Ghani, Nur A. Zolkep, Shamini Selvarajah, and Jamaiyah Haniff. 2012. A comparison between convenience sampling versus systematic sampling in getting the true parameter in a population: Explore from a clinical database: the Audit Diabetes Control Management

(ADCM) registry in 2009. In *Statistics in Science, Business, and Engineering (ICSSBE), International Conference* (IEEE, '12).

7. Joseph S. Dumas and Janice Redish. 1999. *A Practical Guide to Usability Testing*. Intellect Books.
8. Eugene H. Fram and Michael S. McCarthy. 2003. From employee to brand champion. *Marketing Management* 12, 1: 24-30.
9. Wintre M. Gallander, C. North, and L. A. Sugar. 2001. Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology 42*, 3: 216.
10. David Gefen and Detmar W. Straub. 1997. Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly,* 21, 4: 389-400.
11. Fredrick J. Gravetter and Lori-Ann B. Forzano. 2012. *Research Methods for the Behavioral Sciences* (4th ed.). Wadsworth Cengage Learning.
12. Haredo Sahai and Mohammed I. Ageel. 2000. *The Analysis of Variance: Fixed, Random and Mixed Models.* Springer Science.
13. Joseph Henrich, Steven J. Heine, and Ara Norenzayan, A. 2010. The weirdest people in the world? *Behavioral & Brain Sciences* 33, 2-3: 61-83.
14. Sari Kujala and Marjo Kauppinen. 2004. Identifying and selecting users for user-centered design. In *Proceedings of the third Nordic Conference on Human-Computer Interaction* (ACM, '04), 297-303.
15. Mike Kuniavsky. 2003. *Observing the User Experience: A Practitioner's Guide to User Research.* Morgan Kaufmann.
16. Jung-Joo Lee and Kun-Pyo Lee. 2007. Cultural differences and design methods for user experience research: Dutch and Korean participants compared. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces* (ACM, '07).
17. James R. Lewis. 1986. Power switches: Some user expectations and preferences. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 30, 9: 895-899.
18. Michael G. Morris and Viswanath Venkatesh. 2000. Age differences in technology adoption decisions: Implications for a changing workforce. *Personnel Psychology* 53, 2: 375-403.
19. Jakob Nielsen. 1994. *Usability Engineering.* Elsevier.
20. Lene Nielsen and Sabine Madsen. 2012. The usability expert's fear of agility – An empirical study of global trends and emerging practices. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (ACM, '12).
21. Robert A. Peterson. 2001. On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research* 28:450 – 461.
22. Khanyapuss Punjaisri and Alan Wilson. 2007. The role of internal branding in the delivery of employee brand promise. *Journal of Brand Management* 15, 1: 57-70.
23. Khanyapuss Punjaisri, Evanschitzky Heiner, and Alan Wilson.2009. Internal branding: An enabler of employee brand-supporting behaviors. *Journal of Service Management* 20, 2: 209 – 226.
24. Stephanie Rosenbaum, Janice A. Rohn, and Judee Humburg Rosenbaum. 2000. A toolkit for strategic usability: Results from workshops, panels, and surveys. In *Proceedings of the SIGCHI conference on Human Factors in Computing* Systems (ACM, '00), 327-344.
25. Allen Rubin and Earl R. Babbie. 2010. *Essential Research Methods for Social Work* (2nd ed.).Wadsworth Cengage Learning.
26. Jeff Sauro. 2014. Measuring Usability with the System Usability Scale. Retrieved on August 6, 2014 from https://www.measuringusability.com/sus.php
27. Jeff Sauro. 2010. Do you need a random sample for your usability test? Retrieved August 15, 2015 from http://www.measuringu.com/blog/random-sample.php
28. Jeff Sauro and Joseph S. Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (*CHI '09), 1599-1608...
29. Jeff Sauro and James R. Lewis. 2010. Average task times in usability tests: What to report. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI, '10).
30. Andrew Sears and Julie A. Jacko (Eds.). 2009. *Human-Computer Interaction: Development Process*. CRC Press.
31. Osama Sohaib and Khan Khalid. 2010. Integrating usability engineering and agile software development: A literature review. In *Computer Design and Applications (ICCDA),* (IEEE, '10), V2-33.
32. Anne S. Tui, Jone L. Pearce, Lyman W. Porter and Angela M. Tripoli. 1997. Alternative approaches to the Employee-Organization Relationship: Does Investment in Employees Pay Off? *Academic Management Journal, 40, 5*, 1089-1121.
33. Tom Tullis and Bill Albert. 2013. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics (2nd Ed.). Newnes.
34. Brady T. West, Kathleen B. Welch, and Andrez T. Galecki. 2007. *Linear Mixed Models: A Practical Guide Using Statistical Software.* Taylor & Francis Group.