JUS
JOURNAL OF USABILITY STUDIES

Vol. 3, Issue 1, November 2007, pp. 40-52

Clustering for Usability Participant Selection

Juan E. Gilbert
Auburn University
107 Dunstan Hall
Auburn, AL 36849 USA
gilbert@auburn.edu

Andrea Williams
Auburn University
107 Dunstan Hall
Auburn, AL 36849 USA
willia2@auburn.edu

Cheryl D. Seals
Auburn University
107 Dunstan Hall
Auburn, AL 36849 USA
sealscd@auburn.edu

Abstract

User satisfaction and usefulness are measured using usability studies that involve real customers. Given the nature of software development and delivery, having to conduct usability studies can become a costly expense in the overall budget. A major part of this expense is the participant costs. Under this condition, it is desirable to reduce the number of participants without sacrificing the quality of the experiment. If a company could use a smaller participant pool and get the same results as the entire pool; this would result in significant savings. Given a participant pool of size N, is there a subset of N that would yield the same results as the entire population? This research addresses this question using a data-mining clustering tool called Applications Quest.

Keywords

Usability method, usability data analysis, automated tools, experiment, clustering, participant selection, Applications Quest, data mining, sample size

Introduction

In the software development cycle, developers often hold usability studies to test the accuracy and effectiveness of the software and to retrieve user responses regarding the satisfaction or usability of that software. In practice, usability studies can give developers insight into the mind of the user as well as unveil errors, major and minor, within the system.

Of course, as with anything that involves users and studies, planning and budgeting to assess the cost of usability testing and users in the study must be done. Planning studies can be time consuming because activities, such as designing studies, enlisting participants, and possibly implementing several runs of a study must take place. In planning, developers must consider different methods of usability testing, heuristic evaluation, and observation of tasks done in the study; these activities can become burdensome and intimidating to companies not familiar with this practice or not sure which practices will benefit their company most.

Budgeting for studies within the development cycle is often a tug of war because, although several tests might prove beneficial in the long run of the project scheme,

in the short term, the budget might not allow for testing at all, or it might allow for a single test with a select number of participants. Often, numerous problems occur with planning and budgeting that ultimately cause studies to either be drastically reduced in size or eliminated altogether.

Determining the best factors for a study can be problematic because studies should be designed to fit the particular company, its size, and its goal for the study. Some companies are not familiar with design specifications and often must hire someone to implement their study or they neglect it altogether. In some cases, companies even implement their own study. In all cases, the outcomes can become costly if proper judgment is not used in selecting the type of study, the number of participants, the type of participants, or even the number of runs (trials) needed for that particular study.

The aim of this study is to show that using data-mining clustering software, Applications Quest, one aspect of designing a study, namely the selection of participants, can be done effectively to 1) reduce costs and 2) maintain or improve result quality. An experiment was conducted to evaluate this approach by comparing participant selection using Applications Quest versus random selections and evaluates the significant difference of one group's results over the other.

In usability design, when choosing the number of users for a study, there is a debate about the number of participants to use: 1) the five-user assumption, which says five users is all you need in a study; 2) the idea that five users is not nearly enough because five users will not provide enough feedback about a product. Which idea is correct? In actuality, a number of theories claim to know the number of users for a study.

Five-User Assumption

Usability studies can become expensive when it comes to designing and selecting users. Nielsen says, "The best users come from testing no more than 5 users and running as many small tests as you can afford." (Nielsen, 2000). According to the formula, problems found (i) = $N(1-(1-\lambda)i)$ represented graphically below, one user should be able to uncover a third of the findings and as more users are added, redundancy occurs in the information.

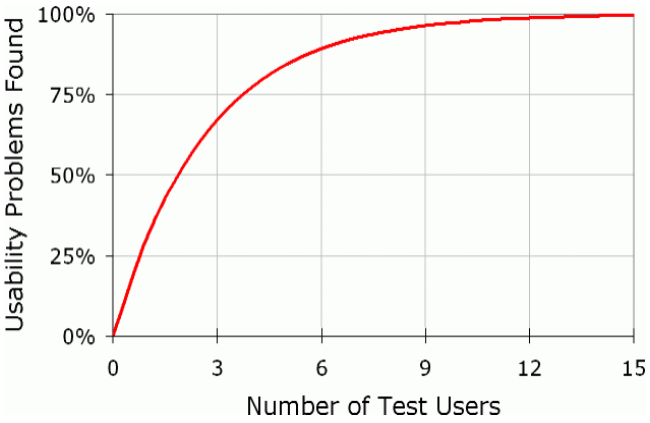

Figure 1: Curve showing relationship between problems found and number of users

Nielsen's study showed that a group of five users were able to find about 80% of the findings in a system and as more users were added, less additional information was found, but more and more money was spent to run tests and to compensate additional users. The idea behind the assumption is that you can learn more from a group of five completing multiple tests than you would on 15 participants completing one test. The study would yield more results and cost the same or less than the study with 15 participants. After reporting these findings, some usability professionals expressed doubts in the five-user assumption.

Five Users And Beyond

After running studies using the five-user assumption, many usability professionals have found that five users are not enough. One study was done where five users were randomly chosen and only uncovered 35% of the findings, while the 13th and 18th user uncovered data that the

original users missed. This result shows that if the study had been discontinued at five, those data would have been overlooked. In this study, users 6-18 were able to find other new data that the original five were unable to find, which shows that, if the right users are not chosen, pertinent data can be left out (Faulkner, 2003). In attempts to describe the confines of the five-user assumption, many professionals neglected the rest of the assumption that recommends running the subjects until the findings meet an "acceptable level," and instead adopted the most minimal number, particularly five. To further examine the theory of five not being enough, Nielsen conducted another, more structured study that took a population of 60 and randomly selected multiple groups of five or more. Each group's findings were then compared against the findings of the entire population to measure how each group's size affected data reliability, confidence, and usability issues. The average percentage of findings by 100 trials of groups of five was 85%, while the average percentage for any random group of five was 55-100%. Adding users increased the percentages, but the most important result showed that 55% was the minimum percentage for a group of five, while a group of 20 produced a minimum percentage of 95% (Faulkner, 2003).

Random Sampling

Random sampling is another method often used in usability studies. When properly done, random sampling contains no bias and can be relatively representative of the targeted population (Arteology, 2005). This method is also used because it requires no prior research or skill in selecting participants and is less expensive. Random sampling allows researchers to make generalizations about the majority of the population, and those claims can be justified by a certain level of certainty (Rosenstein, 2001). Of course, as with any choice made surrounding a usability study and selecting users, companies must choose methods that most benefit their budget and the goal of their study. Samples are chosen in different ways, such as simple random, systematic, weighted (quota), or convenience selection. In the case of simple random selection, participants are chosen from the entire group by the random selection of a unique identifier that can be drawn by hand like a tag drawn from a hat, or mathematically selected by a computer program. Systematic selection is used by dividing the population into partitions, and from each partition, randomly selecting a participant (Arteology, 2005). In some studies, particularly web-based studies where companies are trying to target a specific user group; usability professionals give weight to that particular group to ensure that group's presence in the sample. A convenience selection is where the researcher randomly chooses participants who can be easily found. The participants may or may not be representative of the targeted group at all. Study results have demonstrated that random sampling can be problematic because you can never be 100% certain that the results from the selected sample are representative of the entire population (Arteology, 2005). Random sampling can also give you a false sense of security because in some usability studies, the goal is not to find significant difference, but to find insight into the usefulness of a particular product.

Homogeneous Vs. Heterogeneous Populations

In the article "Eight Users Is Not Enough," authors Perfetti and Landesman found after trying to complete testing on an e-commerce site that the recommendations of four to five users with no more than eight was not enough (Landesman & Perfetti, 2001). The first five users only yielded 35% of the problems in the system; at that rate, it would take 90 tests to uncover the 600 problems in the system. The problem found with this study was that the usability professionals tried to apply a concept that did not quite fit their needs. E-commerce websites contain much more complexity in content versus software and simple websites, and continuously and incrementally change, whereas software only changes with each version release, which does not occur as frequently. With that discovery, they also found that their users varied just as much as the complexity in their system. Their results showed that a sample group could not be used as a representation of the whole because each user who interacted with the system used the system differently. By understanding the type of product they were testing and their users, the authors were able to successfully learn what worked for their system (Landesman & Perfetti, 2001).

Literature Review Summary

When choosing the "right" participants, it is imperative that the users be representative of the population your product is trying to solicit (Heim, 2008). As a sample of the entire group, gathering the relevant demographic information can prove to be helpful in differentiating between the results of individuals in the group (Rosson & Carroll, 2001).

Recruiting these representative participants is another timely and costly activity that creates an intimidation factor for potential usability professionals. Most professionals agree that testing should be done, but some companies just do not have the capability or experience necessary to conduct small tests, let alone multiple tests involving users within deadlines set for the project. On average, it is said to cost $107/user in a study, depending on location and profession, and that is without a recruiting agency's help. Companies who use recruiting agencies must add additional fees, while other companies must spend approximately 1.15 hours per person recruited to obtain participants (Nielsen, 2000).

Even after choosing the "right" participants, it is important for practitioners to understand that there are variables within a study that they have varied control over. The types of participants a usability professional can find, the mission criticality of a system, or usability issues found posing a problem to a system have a deep impact on the number of users a study needs to obtain accurate results.

All things considered, there is still not just one way to select the number of participants to use in studies, nor is there a way to select which user should be used or is most representative. A method that could help usability professionals minimize costs and test group sizes, as well as maximize results would have a significant impact in usability design.

Using Applications Quest To Identify Experiment Participants

"Applications Quest is a data-mining software tool that clusters admission applications based on holistic comparisons" (Gilbert, 2004). The idea for this software came from two landmark court cases, Grutter vs. Bollinger and Gratz vs. Bollinger, where two students challenged the University of Michigan's admissions policies. Because of these cases, the Supreme Court ruled that diversity could be used in admissions policies, but race could not be the determining factor for admission. It was determined that applicants' applications should be reviewed holistically, and not based on a single attribute, such as race or ethnicity (Gilbert, 2004). The notion of holistically reviewing an application means considering each and every attribute of the application, such that no single attribute weighs heavier than another. For admissions committees, the action of holistically reviewing an application is time-consuming and difficult because humans do not possess the ability to effectively compare attributes with reproducibility of the results. Applications Quest achieves the goal of holistically comparing applications and recommending applicants that holistically represent diversity; with diversity not being defined by or giving preference to race or ethnicity. Because the algorithm compares each application with the same rigor, the results are reproducible and justifiable (Gilbert, 2004).

How Applications Quest Works

Applications Quest holistically compares applications one to another and places them in groups or clusters based on their holistic similarity using clustering (Gilbert, 2004). The algorithm uses a novel approach for nominal (non-numeric) attributes and attribute-value pairs to compare each application. The number of values each application has in common determines its placement in a cluster. In other words, attributes, such as major or ethnicity, are paired with their corresponding values and compared amongst all other applications. Those applicants with attributes in common are given a difference/similarity index. Similar applications, those whose indices are in close proximity, appear in the same cluster. With diversity in mind, each cluster is designed to hold similar applications, but from each cluster, the most different or novel applicant is recommended for admission. Given the nature of the Applications Quest approach to solving diversity in university admissions and employee selection; how does this relate to the issue of participant selection for usability studies?

Divisive Clustering: A Possible Solution To Our Problem

Employing a divisive clustering algorithm, Applications Quest recommends applicants who are representative of holistic diversity within an admissions applicant pool. Using this same software, but modifying the context in which it is used, namely for participant selection in usability studies, could possibly help usability professionals select the most representative users of their targeted population. The idea is that users selected by Applications Quest will yield the same, if not better, results as the entire population of potential test users versus those randomly selected. If this premise is true, Applications Quest would pose a solution that has a minimal cost, reproducible recommendations, and quality results.

Methods

Given a prospective participant group of size N, can Applications Quest select a subset of users, particularly a group size of 7 or 15, whose study results would be representative of the population? To determine if the group was representative, it was necessary that their results prove insignificantly different than those of the majority population.

Data

The data for this experiment were selected from a previous study done in a university research lab. The 72 users in the study were members of the targeted user base and represented students from Science, Technology, Engineering, and Mathematics (STEM) majors. Their ages ranged from 19-30 years old. Of the 72 participants, 70 spoke English as their native language, while the other two spoke English as their second language. There were 21 females and 51 males. In the previous study from which these data were collected, pre-experiment surveys were used to collect the participant's demographic data. Post-experiment surveys were used to collect the participant's opinions of the software.

Procedure

The data were imported into a Microsoft Access database. This experiment was conducted in two parts, with the second part being done two different ways:

Test 1

Randomly selected participants were chosen from a group of 72 users. Each participant was given a unique identifier from the original study; that identifier was used in this study as well to maintain their anonymity.

To select users in this approach, a program was written to randomly select groups of participants. The group sizes selected were 7 and 15; approximately 10% and 20%, respectively. For each group size chosen, five trials were run. For each group size, random trials were run five times, meaning that for each trial, new participants were randomly chosen.

Each random participant's answers to questions selected from the questionnaire were queried and placed in a Microsoft Excel spreadsheet where they could be tested for any significant difference from the entire population. The attributes chosen from the questionnaire, wonderful---terrible, frustrating---satisfying, usable---not usable, were based on a 5-point Likert scale.

Significant difference was tested on each group in Microsoft Excel. The t-test can employ three different assumptions, but for this experiment, the two-sample unequal variance assumption was used. (A two-sample unequal variance assumes that the two samples used have come from distributions with unequal variance and is used to determine whether the distributions have equal population means.) Once calculated, the results of the t-test were analyzed to see if the randomly selected group could be considered representative of the population.

Test 2A

All pre-experiment survey demographic data were loaded into a database and run in Applications Quest. Pre-experiment survey data included race, gender, major, age, income, and other attributes.

Applications Quest was given a specified number of clusters to return, and from those clusters, it chose the most representative person of each cluster, e.g., the individual who is least different or closest to the center of the cluster.

Once the participants were chosen, their answers to questions on the post-experiment survey were queried and placed in a Microsoft Excel spreadsheet, where they could be tested for any significant differences from the entire population. The same attributes for the first part of this experiment were employed here as well.

Significant difference was again tested with the t-test and the results analyzed for comparison.

Test 2B

The data loaded into the database for part 2A were used to run Applications Quest again. The algorithm however for part 2B was changed to select the most unique person from each cluster.

e.g., most different/novel or the furthest from the center. Again, the same attributes were used for querying, and results were tested and analyzed using the t-test to determine significant difference.

Results

Once the statistical analysis tools had been applied as described in both approaches to each group size chosen for this experiment, with the following results:

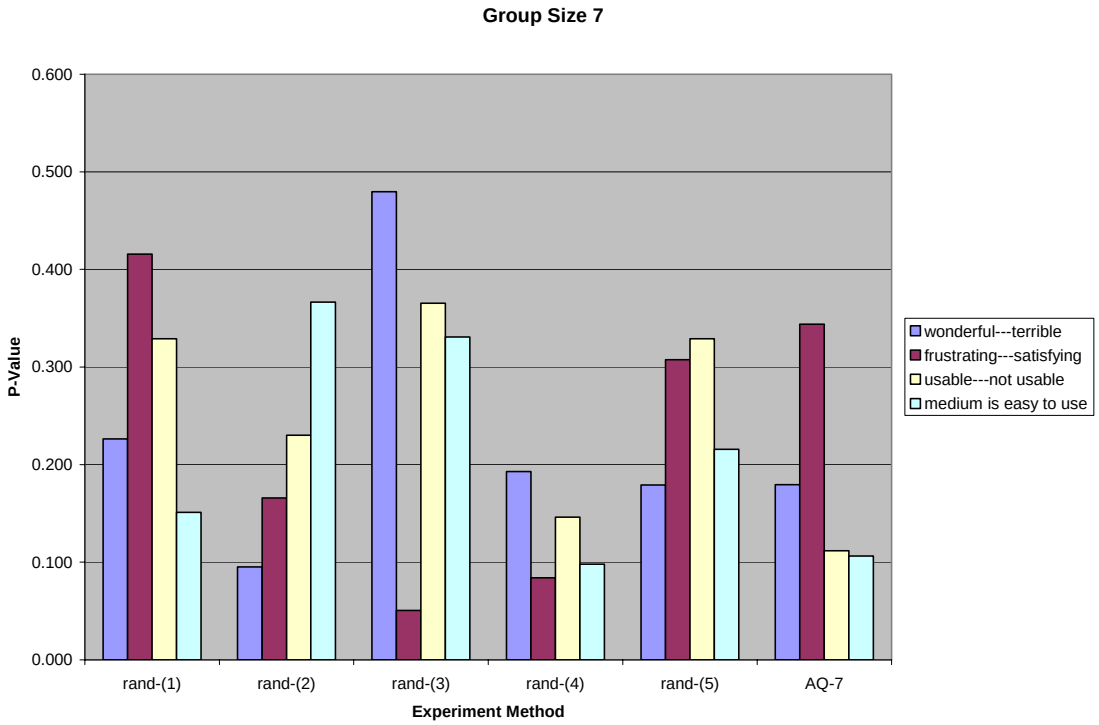

Figure 2: Comparison of statistical differences for group size 7: Random Trial vs. Applications Quest

In Figure 2, trials rand (1), rand (5), and Applications Quest each produced all insignificantly different attributes. The p-values for the random trials were able to surpass those of Applications Quest, but the probability of those trials being selected was only 40 percent. The other random trials were able to generate attributes with insignificant difference, but they still maintained a level of inconsistency. Because the results of Applications Quest compared to the random results initially seemed unaligned with this experiment's hypothesis, approach 2B (Applications Quest with a revised algorithm) was designed and produced the following results:

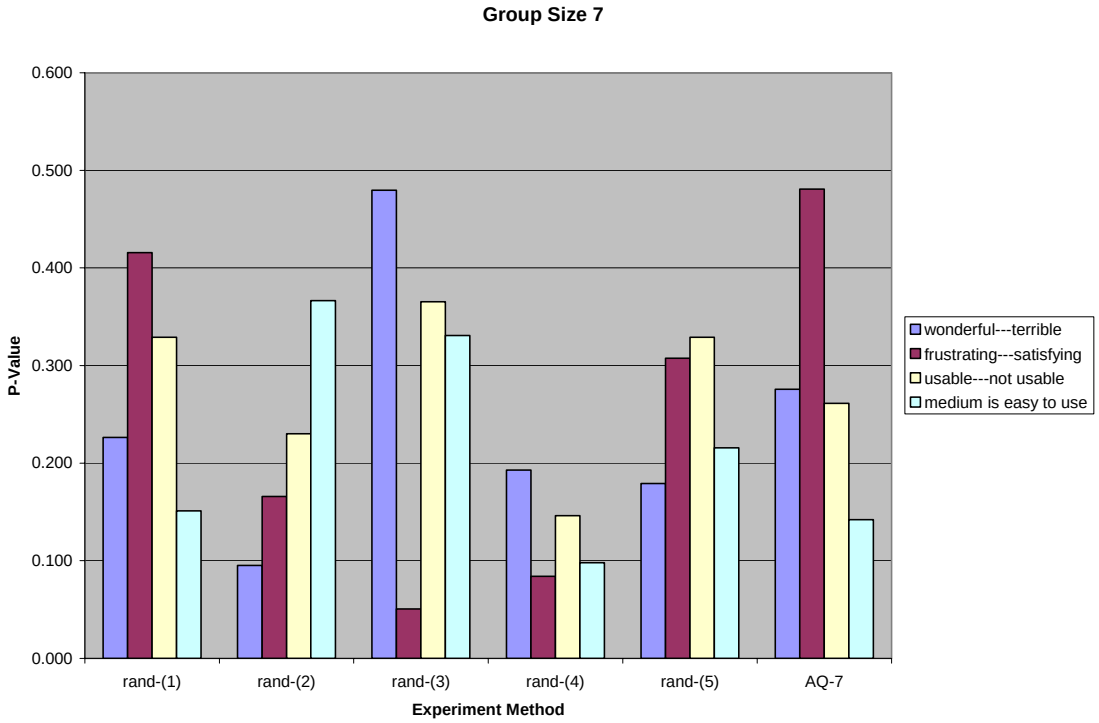

Figure 3: Comparison of statistical difference with revised algorithm for group size 7: Random Trial vs. Applications Quest

Applications Quest was able to produce all attributes with insignificant difference in Figure 3. Random trials one and five also successfully generated a complete set of attributes insignificantly different from the population. Applications Quest group size 7 was able to produce complete sets of attributes insignificantly different from both versions of its algorithm. Group size 15 was unable to demonstrate this. In figure 2 and 3, it can also be seen that the attribute usable-not usable was found insignificantly different in both the Applications Quest trial and the random trials. Below in Figure 4, the graph shows the mean Likert score for each trial, including both versions of Applications Quest. From the graph, it can be seen that all of the trials averaged about the same Likert scores, but the standard deviation among each trial varied tremendously.

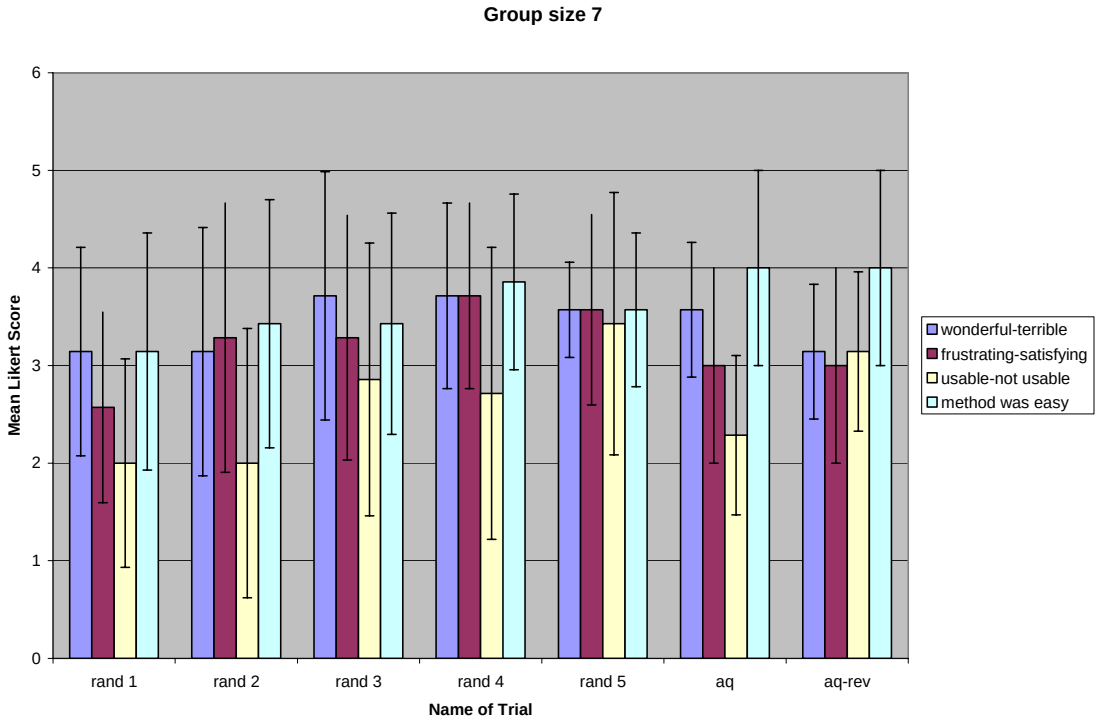

Figure 4: Comparison of mean Likert score for each experiment trial in group size 7

	wonderful-terrible	frustrating-satisfying	usable-not usable	medium is easy to use
rand 1	**1.0690**	**0.9759**	**1.0690**	**1.2150**
rand 2	**1.2724**	**1.3801**	**1.3801**	**1.2724**
rand 3	**1.2724**	**1.2536**	**1.3973**	**1.1339**
rand 4	**0.9512**	**0.9512**	**1.4960**	**0.8997**
rand 5	**0.4880**	**0.9759**	**1.3452**	**0.7868**
aq	**0.6901**	**1.0000**	**0.8165**	**1.0000**
aq-rev	**0.6901**	**1.0000**	**0.8165**	**1.0000**

Figure 5: Table showing the standard deviation for each trial and attribute

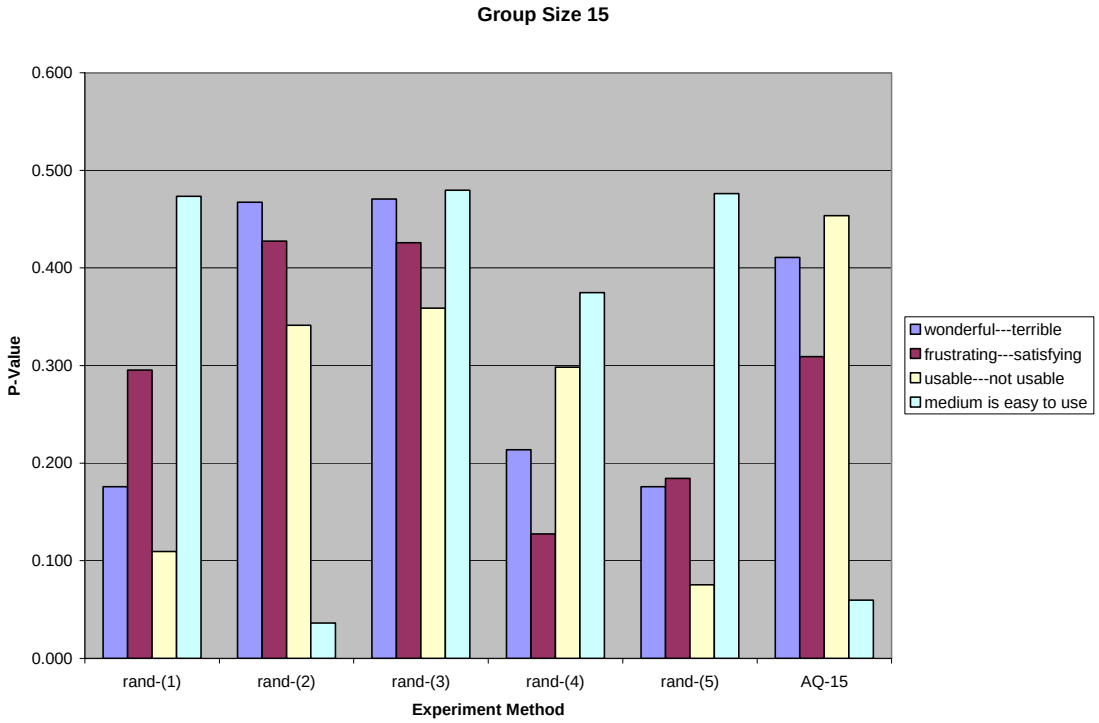

Figure 6: Comparison of statistical differences for group size 15: Random Trial vs. Applications Quest

Figure 6 shows each trial yielding insignificant differences for at least 75% of the attributes tested. Upon first glance, the results seem remarkable, but to a corporation designing a study with very limited funds, this group size is just not workable. When comparing the results of group size 7 and 15, it can easily be seen that group size 7 is the most feasible choice.

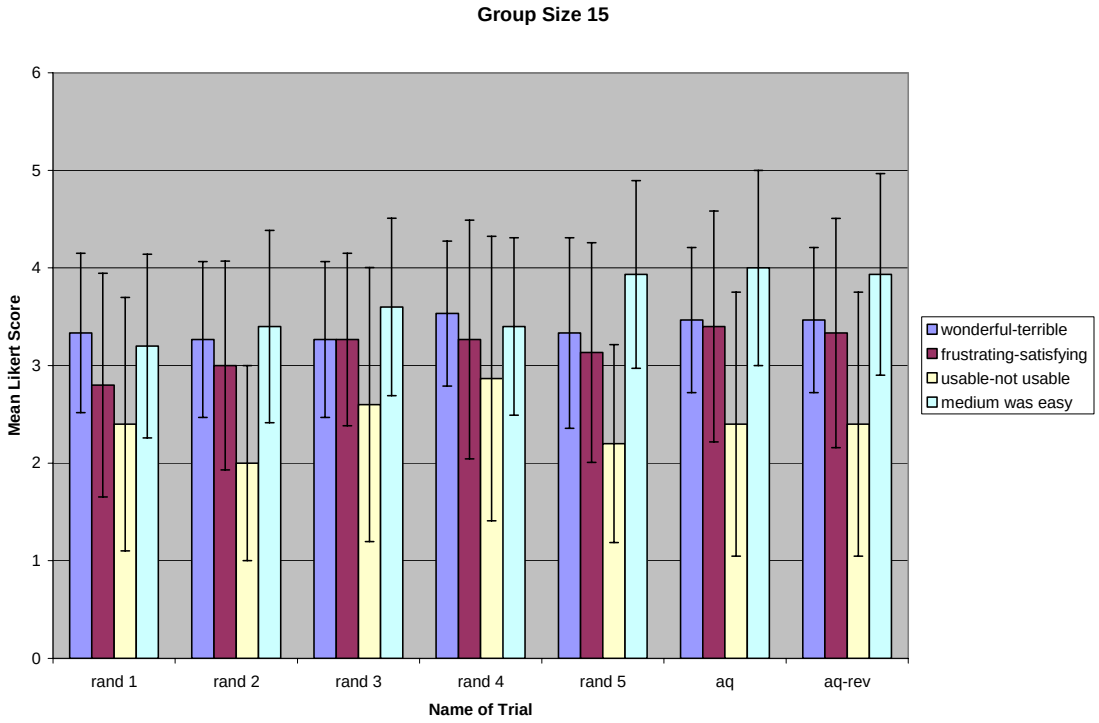

Figure 7: Comparison of mean Likert Score for each experiment trial in group size 15

	wonderful-terrible	frustrating-satisfying	usable-not usable	medium is easy to use
rand 1	0.8165	1.1464	1.2984	0.9411
rand 2	0.7988	1.0690	1.0000	0.9856
rand 3	0.7988	0.8837	1.4041	0.9103
rand 4	0.7432	1.2228	1.4573	0.9103
rand 5	0.9759	1.1255	1.0142	0.9612
aq	0.7432	1.1832	1.3522	1.0000
aq-rev	0.7432	1.1751	1.3522	1.0328

Figure 8: Table showing the standard deviation for each trial and attribute

In Figure 7, the mean scores for each trial were even more similar than those of group size 7; this can be largely accounted for because group size 15 represents a greater percentage of the targeted population. As the sample size increases, the differences between the targeted population and the sample will decrease. Because the goal is to find the lowest number of participants representative of the population with a greater number of certainty, Applications Quest outperformed the random samples and reduced the possibility of group size 15 being selected.

Another interesting finding was that group size 7 from Applications Quest was able to find 100 percent insignificantly different in both approaches. This suggests that if a usability professional were designing a study s/he could be 100 percent certain that the group selected by Applications Quest would provide him the same results as the other 72 participants in the targeted population. With this certainty, the professional could use the 7 users versus the 72 and save money on user testing, stay on budget for the usability design portion, and even stay on schedule for the time allotted to testing.

In the random trials, the results were promising, but the random trials were too unpredictable. In Applications Quest, the algorithm for selection is the same every time; choose the participant that is the most similar or most different. The random trial results returned the larger group as most representative. This becomes a problem because the idea is to find the minimum number of participants. Group sizes of 15 had very high percentages of trials that were insignificantly different, but that does not say much because that is almost one-fourth of the population.

In comparison, the Applications Quest algorithm for selecting the most similar participant was more effective in selecting a better percentage of representative groups. Also in the attribute breakdown, the most similar algorithm returned more attributes with 100 percent certainty of insignificant difference. This result suggests that, although both algorithms were performed in close proximity, choosing which algorithm to use comes down to the goals of the study. If the study aims to find users who are most representative of the targeted population, they would use the most similar algorithm.

Conclusions

The goal of a usability study is to identify and reduce issues with the software and to improve user satisfaction. When the job of designing a study and finding participants becomes too expensive, many designers and researchers cut back on user testing. This experiment was designed to help find a plausible solution to selecting participants for studies by using Applications Quest.

Applications Quest would take a group of size N and, from that group, select participants who would be representative of the population. The selected participants would help reduce costs by reducing the number of participants necessary, while maintaining result quality.

Two approaches were used for comparison, random selection and Applications Quest selection. The random trials produced results worth observing, but they were inconsistent. Applications Quest was able to present results that were insignificantly different and consistently reproducible. The random trials were unpredictable, which does not guarantee the certainty or reliability necessary in selecting participants. Upon comparing Applications Quest to itself when revising its original algorithm, the original version (selects the most representative participant from each cluster) executed more effectively than the algorithm selecting the most different participant from each cluster (which is preferred in admissions). These findings suggest that Applications Quest could be a promising solution to the issue of participant selection. The results are reproducible and consistent, and with more experimentation, it could guarantee a higher percentage of insignificant difference.

Practitioner's Take Away

Conducting user studies can be very expensive when selecting several experiment participants. The approach of collecting participant demographics and other information as input into clustering is a mechanism for selecting ideal participants. As a guideline for practitioners:

1. Identify your target user base.
2. Develop a survey instrument to collect information about them that you deem relevant to the evaluation task. This may include demographic, system experience, etc.
3. Process the completed surveys in a clustering tool using a K-Means approach where you can specify the number of clusters. Ideally, you could use Applications Quest because it will cluster the applicants and make recommendations on which applicants should be used in your experiments. If you don't have Applications Quest, there are alternative methods, which are described next.
4. If you don't have Applications Quest, you can cluster the surveys and then hand pick individuals from each cluster. This could be challenging if the clusters are large, or if you have a large number of clusters.

Future Work

The results suggest the random trials of this experiment yielded insignificantly different results along with Applications Quest, but the hit-and-miss nature of the random trials does not lend

itself to promising certainty. For each group size, there were five random trials and of those five, the best and worst of the results still weren't consistent enough to promote confidence, such as with Applications Quest. Additional research is needed to further evaluate Applications Quest. The same experiment will be run on less homogeneous data, as well as on larger datasets. On a less homogenous, larger distribution of data, it may be found that Applications Quests can select the most representative more efficiently because a larger distribution will lend itself to comparing a less dense cluster. This experiment could also be done with more demographic data. Examining different sets of demographic data with human subjects input into Applications Quest could possibly show a trend in what information usability professionals could use in recruiting participants.

Acknowledgements

This material is based in part on work supported by the National Science Foundation under Grant Number CNS- 0540492. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s), and do not necessarily reflect the views of the National Science Foundation.

Applications Quest is a trademark of Applications Quest, LLC. Microsoft Access and Microsoft Excel are trademarks of Microsoft Corp.

References

Arteology: Sampling. Retrieved March 27, 2007 from www.uiah.fi/projects/metodi/152.htm.

Faulkner, Laura. (2003) Beyond the five-user assumption: Benefits of increased sample sizes in usability testing, *Behavior Research Methods, Instruments, & Computers, 35(3),* 379-383. Retrieved March 9, 2007 from http://www.geocities.com/FaulknerUsability/Faulkner_BRMIC_Vol35.pdf.

Gilbert, J. (2004) *Applications Quest*. Retrieved March 29, 2007 from http://www.ApplicationsQuest.com.

Heim, S. (2008) *The Resonant Interface: HCI Foundations for Interaction Design.* London: Pearson Addison Wesley.

Landesman, L. & Perfetti, C. (2001, June) Eight is not Enough, *User Interface Engineering*. Retrieved March 29, 2007 from http://www.uie.com/articles/eight_is_not_enough/.

Nielsen, J. (2000, March) Recruiting Test Participants for Usability Studies, *Alertbox*. Retrieved March 9, 2007 from http://www.useit.com/alertbox/20030120.html.

Nielsen, J. (2000, March) Why You Only Need to Test With 5 Users, *Alertbox*. Retrieved March 9, 2007 from http://www.useit.com/alertbox/20000319.html.

Rosenstein, A. (2001, September) *Managing Risk with Usability Testing, Classic System Solutions*. Retrieved March 29, 2007 from http://www.classicsys.com/css06/cfm/articlesusability.cfm.

Rosson, M.B. & Carroll, J. M. (2001) *Usability Engineering: Scenario-Based Development of Human-Computer Interaction.* San Francisco: Morgan Kaufmann Publishers.

About the Authors

Juan E. Gilbert
is the TSYS Distinguished Associate Professor in the Computer Science and Software Engineering Department at Auburn University where he directs the Human Centered Computing Lab. He has active research projects in spoken language systems, usability and data mining.

Cheryl D. Seals
is an Assistant Professor in the Computer Science and Software Engineering Department at Auburn University. She conducts research in Human Computer Interaction with an emphasis in visual programming of educational simulations with end user programming, intelligent agents, usability evaluation, computer supported collaborative work, minimalism. She also has projects in software engineering.

Andrea Williams
attends Auburn University where she is a President's Graduate Opportunities Program Scholar and has recently received her Masters of Science in Computer Science for her work on using software to predict group sizes for usability studies. She is currently working on her PhD in Computer Science and her research interests include, but are not limited to, Human Computer Interaction, Data Mining, and Culturally Relevant Computing.