CHARM

Choosing Human-Computer Interaction (HCI) Appropriate Research Methods

Controlled Experiments

Liz Atwater
Department of Psychology
Human Factors/Applied Cognition
George Mason University
lizatwater@hotmail.com

Ketan Babaria
Institute for Systems Research
Human Computer Interaction Lab
University of Maryland, College Park
kbabaria@cs.umd.edu

1. Introduction

Experimentation supports from the scientific method, which is a process a proposing and verifying hypotheses [1]. The first step of the scientific method is to develop a theory. This can be something from your own observations or a conjecture borrowed from previous research. A hypothesis is formed next. This is a statement that can be tested. For example, the statement "Users can read a document faster when a small font is used versus a large font" is a testable hypothesis. The statement "Large screens are good" is not a testable hypothesis. What is meant by good and how would it be measured? The experimenter forms the hypothesis and carries out the experiment. Independent and dependent variables are chosen and the method in which they will be manipulated and observed is decided upon. Then the observations are made on the behavior, the results are analyzed, and the original theory is refined. You may have found that your results are contrary to what has been found previously, or that the results were the same except for in one condition. Either way, the results of an experiment are always meaningful and can further the theory [1].

Controlled experiments are considered to be the most rigorous of the research methods [2]. The goal of a controlled experiment in Human-Computer Interaction (HCI) is to support a theory and make predictions about behavior. This is done by manipulating an independent variable, like screen size or input method, and observing the outcomes on performance, like scanning speed or information input speed. Independent variables will have at least two treatments, for example a 3 x 4 cm screen and a 4 x 5 cm screen. The outcomes are also known as the dependent variables; things that depend on the treatment of the independent variable [2].

The procedures used can be summarized with the phrase MaxMinCon, which is shorthand for
(1) Maximize systematic variance between the treatment groups,
(2) Minimize error variability within each treatment group, and
(3) Control extraneous systematic variance [3].
In other words, the two treatment groups should be as different as possible, the control and experiment groups should have few, if any, differences between them, and ensure that the measures used are reliable. While many controlled experiments share similar characteristics there are different types of experiments that are used in different situations. These will each be discussed with their respective advantages and disadvantages, examples of previous research, and links that provide further information.

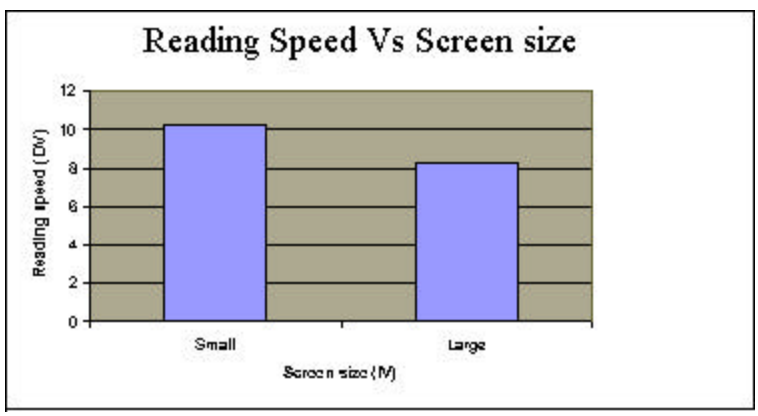
2. Different methods

2.1. Single Factor

2.1.1. Description:

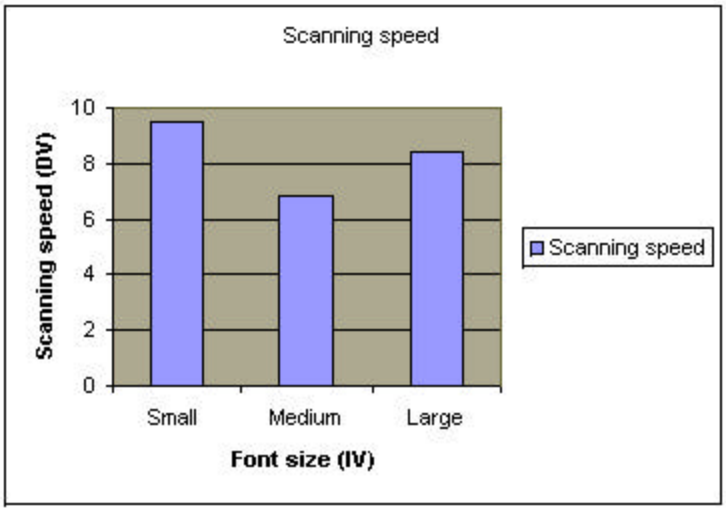
Single factor controlled experiments are characterized by having only one independent variable that is manipulated [3]. This is the simplest form of a controlled experiment. Once the experiment has been planned, participants should be randomly selected for participation and then randomly assigned to the two groups. If participants are not selected and assigned randomly, there is an element of experimenter bias present. The next step is to have the participants complete the tasks that were planned. Care should be taken that each participant receives the same instructions and procedure; otherwise there will be extraneous error that weakens the confidence in the results. Once the data has been collected, statistical analyses are conducted to determine if the treatments groups caused a significant difference in the outcome measures [2]. Depending on the number of treatments there are for the independent variable, either a t-test or a one-way ANOVA can be conducted. If there are only two levels a t-test can be used, but if there are more than two levels a one-way ANOVA must be used [2]. Here are some examples of what a table and chart of a single factor experiment would look like. The independent variables are on the top of the table and the dependent variable is on the left.

	Small Screen	Large Screen
Reading Speed	10.2 s	8.3 s



This single factor experiment has two treatments of the independent variable: small and large screen size. By looking at the graph there appears to be a difference in reading speed with the two screen sizes. A t-test should be conducted to determine if this difference is statistically significant.

	Small Font	Medium Font	Large Font
Scanning Speed	9.5 s	6.5 s	8.4 s



By examining this graph, it appears that the medium font size has the fastest scanning speed. Because there are three treatments of the independent variable (small, medium, and large font size), a one-way ANOVA should be conducted. One-way ANOVAs are used when there is only one independent variable and it can determine if there is a significant difference between one of the treatment means and the overall mean. If there is a significant difference, post-hoc tests should be conducted to determine which treatment mean is different from the others [2].

2.1.2. Advantages/Disadvantages

The single factor experiment is considered to be statistically and structurally elegant because it is easy to control and simple, robust statistics can be performed on the data [3]. It is also easy to extend the independent variable to more groups. The major drawback of this type of controlled experiment is that it is not very efficient. Experiments can take a lot of time, money, and effort to complete so it is desirable to examine more variables in one experiment rather than look at each one separately in many experiments. . By using more than one independent variable in an experiment, the resources can essentially be cut in half compared to conducting several experiments.

2.1.3. Examples

Example 1:
Icons in graphical user interfaces can be used for many different ways and for many different purposes. But are icons always better than text or a mix of icons or text or just the text. The question is one of generality. At one extreme it might be argues that icons are inherently more attractive than text and should be used whenever possible. At he

other extreme one might argue that icons can be useful only when performing functions that are inherently visual or for systems that include a complete graphical metaphor such as virtual desktop.

Douglas et al [4] tried to get answer to this question by performing a controlled experiment for the three different interfaces 1. Icons only 2. Icons with command name and 3. Command names only. The experiment measured the user preference for a particular type of interface.

Independent Variable: Interface. Three treatments 1. Icons only 2. Icons and Command name 3. Command names only

Dependent variable: User preference rating

Results:

	Icons only	Icons & Command name	Command names only
User Preference Rating	3.92	5.55	5.67

Figure 1. User preference ratings on each of the interfaces

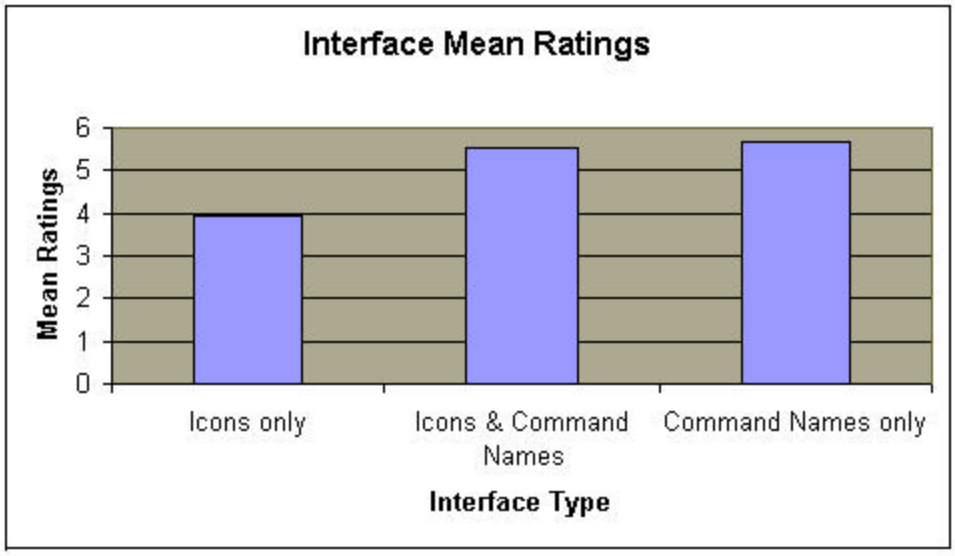


Figure 2. User preference ratings Vs Interfaces

Example 2:

Author's Interactive Design Dialogue Environment (AIDE) is an interactive tool for human computer interface implementation. The user of this tool could implement an interface by directly manipulating and defining objects rather than by the traditional method of writing source code. A controlled experiment study was done by Deborah Hix [5] to empirically evaluate the usefulness of such a tool.

Hypothesis: Creation and modification of an interface is faster and easier by using AIDE than by writing source code in programming language.

Subjects: Group of 3 expert AIDE users and group of 3 expert programmers in 'C' were chosen. Choice of expert subjects was to avoid the issue of training.

Tasks: 1. Interface creation 2. Interface modification.
Both group of subjects were given the same written description of the interface including the sketches and textual explanation. They were told that the task had two parts, creation and modification. They were told they would be given first the creation task and then when it when it had been verified by the experimenter for correctness they would be given the modification task.

Independent Variable:
Mechanism for creating user interface Two treatments 1. AIDE 2. Programming Language

Dependent Variables:
1. Length of time taken to create the user interface
2. Length of time taken to modify the user interface.

Results:

	AIDE	Programming Language
Creation Task mean time	43 min	168 min
Modification Task	29 min	63 min

Figure 3. Table summarizing the time taken for completing each task using two different mechanisms

The difference between the two groups was significant for the creation task $t(4) = 11.9$, $p<0.005$ and for modification $t(4) = 9.9$, $p<0.005$

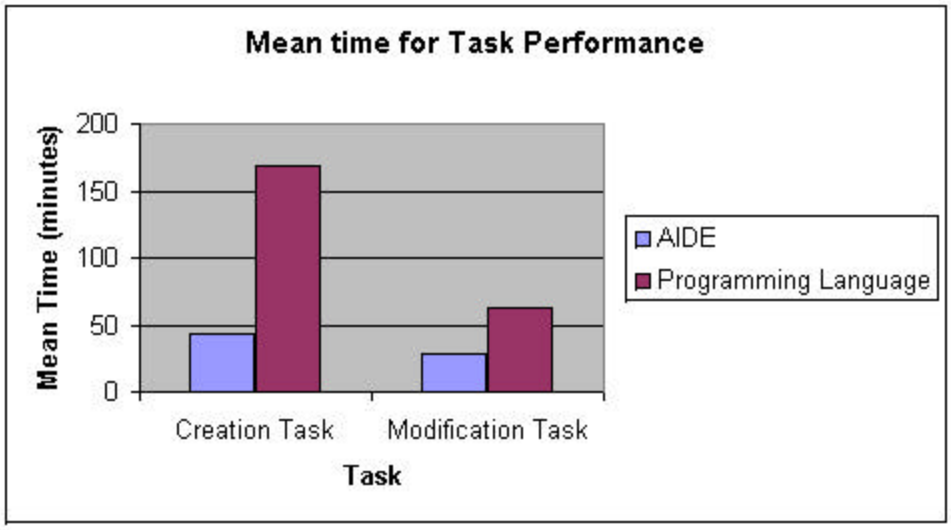


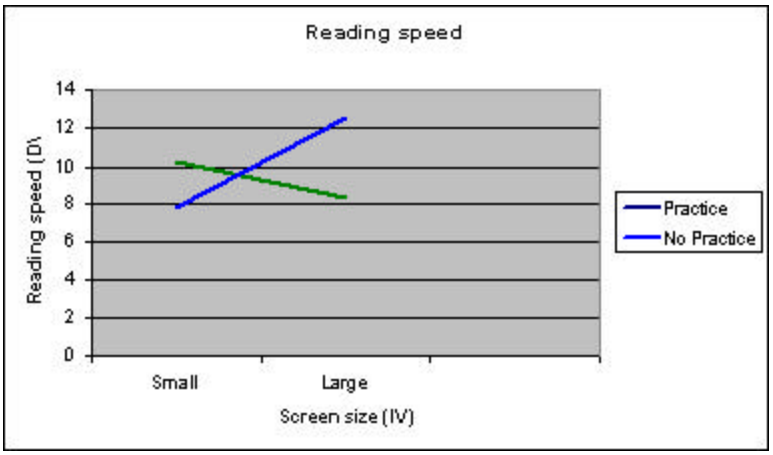
Figure 4. Mean time vs. task for each mechanism

2.2. Multi Factor

2.2.1. Description

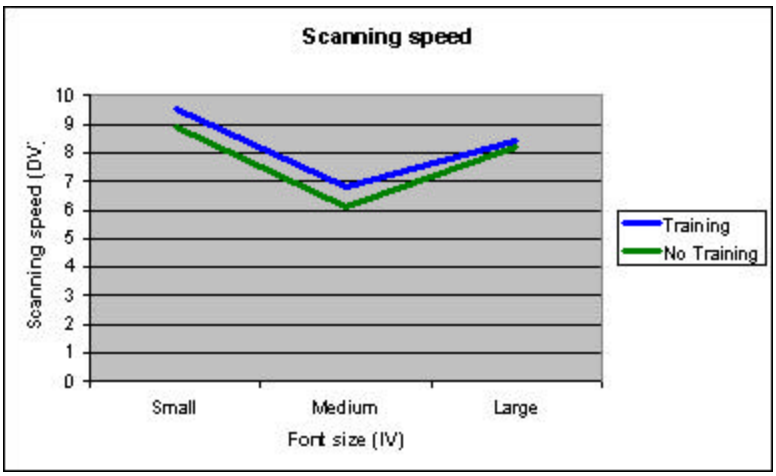
Multi-factor controlled experiments involve two or more independent variables [3]. The same procedures for the single factor experiment are used for the multi-factor. This design is used more often than the single factor because of the greater variety of research questions that can be answer by it. With this design, the data will always have to be analyzed with ANOVA [2]. Whether a two-way, three-way, etc. is used depends on the number of independent variables. Here are some examples of what tables and charts for a 2 x 2 and a 2 x 3 multi-factor experiment would look like. The independent variables are on the top and on the left, while the dependent variables are inside the cells.

	Small screen	Large screen
Practice	10.2 s	8.3 s
No practice	7.8 s	12.6 s



In this example, it appears that the practice group read faster when they use a large screen and the no practice group read faster when they use a small screen. This is known as an interaction effect, i.e. when two groups perform oppositely on different treatments [2]. An interaction effect can always be seen in graph form when the two lines cross. If the lines were parallel, it would mean that both groups performed better on the same treatment of screen size. A two-way ANOVA can determine if the two groups perform significantly differently from each other.

	Small font	Medium Font	Large Font
Training	9.5 s	6.8 s	8.4 s
No Training	8.9 s	6.1 s	8.2 s



In this example, both groups performed similarly on all three treatments of font size. Both groups had the fastest scanning speed when they saw a medium font and slower speeds when they saw small and large fonts. As can be seen from the graph, there is no interaction effect (the lines never cross). A two-way ANOVA can determine if the differences in performance are significant [2].

2.2.2. Advantages/Disadvantages

The greatest benefit of multi-factor experiments is the ability to analyze interaction effects between the independent variables[3]. Not only can you see if there was a difference overall because of the screen size (i.e. a main effect), but you can see if genders performed differently from each other with respect to the screen sizes (i.e. an interaction effect). As can be seen from the first example, women had a faster reading speed when they used the small screen and men had a faster reading speed when they used the large screen. In general, if the lines of the two groups cross on the graph, there is an interaction effect. As can be seen in the second example, the two lines do not cross and therefore there is not an interaction effect. Both age groups had the fastest scanning speeds when the medium size font was used. It appears that there is a main effect for font size because the medium font always produced the fastest scanning times. It is unclear just by examining the graph if there is a main effect for age because there is not a great gap between the two age lines (i.e. they seem to be performing similarly).

Another benefit of multi-factor designs is that they are considered to be more efficient than single-factor designs[3]. Not only can you examine more independent variable, but the combination of them through the various interactions allows for greater generalization to other situations.

A disadvantage to this design over the others is that it can become too complicated if too many independent variables are explored at once[2]. Also, any design with over three independent variables becomes very difficult to analyze if you do not have a statistics computer program to help with the analyses.

2.2.3. Examples

Example 3. A menu is a list with a limited number of options. Gary Perlman [6] conducted a controlled experiment to study how menu length, menu ordering and menu items affect the search time.

<u>Null Hypothesis:</u> Menu length, menu ordering and menu items has no effect on the search time.

<u>Independent Variable:</u>
1. Menu List. Two treatments a. Numbers from 1 to 20 b. Names starting from letter 'a' through letter 't'
2. Menu Length. Four treatments a. 5 b. 10 c. 15 d. 20 3.
List type. Two treatments a. Sorted b. Random

<u>Dependant Variable:</u>
1. Search time

<u>Results:</u>
1. Finding words took longer than numbers $F(1,28) = 11.7$, $p<0.01$
2. Sorted list were easier to search than random $F(1,28) = 10.05$, $p<0.001$
3. It took longer to find items in longer lists than sorter lists $F(3,84)=113.86$, $p<0.001$

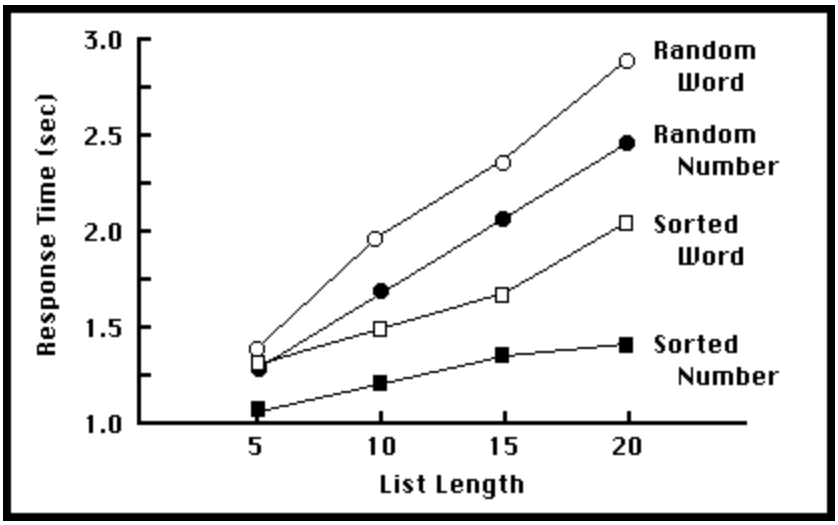


Figure 5. Response time vs. List Length

Example 4.

Menus have been popular method for accessing information in computer systems but human short term memory limitations reduces performances efficiency for deeper hierarchies. Scrolling offers an alternative option for accessing information. Sarah J Swierenga [7] conducted controlled experiment to find the relative efficiency of a scrolling and menus as alternative access methods.

Independent variable:

1. Access methods (Four treatments) a. Menuing (Previous menu, next menu, main menu etc) b. Line by line c. Half computer screen (12 lines) d. Full screen (24 lines)
2. Word familiarity (Two treatments) a. Familiar words b. Unfamiliar words

Dependent Variable:

1. Mean total task time.

Results:

1. For familiar words the effect of access method was significant $F(3,40) = 18.27$, $p<0.0001$
2. For unfamiliar words the effect of access method was significant $F(3,40) = 69.38$, $p<0.0001$
3. Menuing was the fastest technique, followed by line-by-line, full screen and half screen scrolling

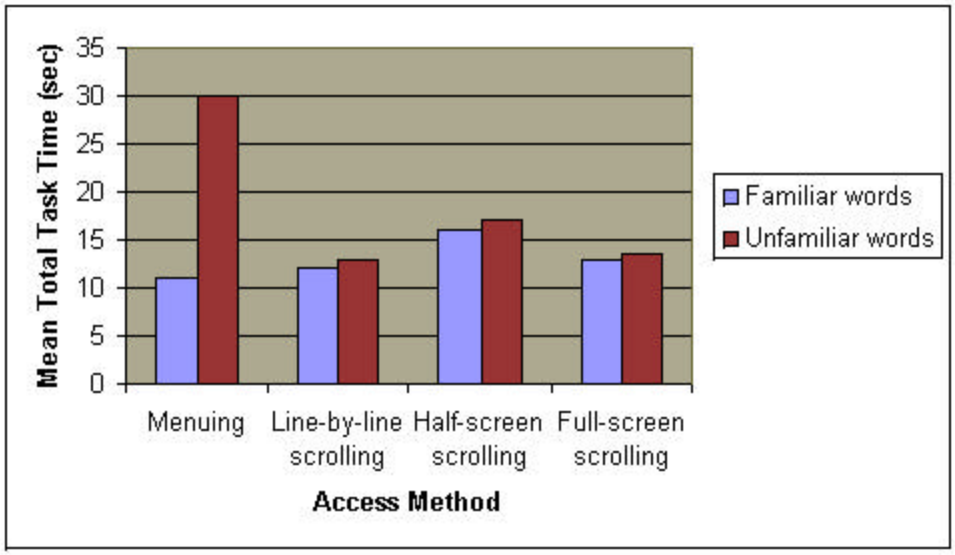


Figure 6. Means for access method by word familiarity on Mean Total Task Time

2.3. Quasi-experimental

2.3.1. Description:

A quasi-experimental design is a controlled experiment without all the control[3]. In essence, what is lacking is random assignment to groups. Quasi-experiments are very similar to true experiments but use naturally formed or pre-existing groups. For example, if you wanted to study the performance differences of two age groups on a certain interface, it would be considered a quasi-experiment because the age groups are naturally formed. It is impossible to assign people randomly to young and old age groups because it is already predetermined. Another characteristic of quasi-experimental designs is that the testing environment may not be as controlled. Instead of having the testing done in a lab setting, it may be done on a job site or in someone's home.

2.3.2. Advantages/Disadvantages

The advantage of quasi-experiments is that they are easier to implement[3]. It is much easier to use groups that are already formed than to have to worry about randomization. The main disadvantage to this design is that it is inferior in terms of internal validity. The threats to internal validity inherent in this design are selection bias and the interaction of selection and maturation of the participants. Because the participants were not randomly assigned, it is impossible to know if the changes that occurred were due to the treatment or changes in the individual. Further, you must be careful in making statements of causality because of the lack of total control.

2.3.3. Example

For example with Quasi-experiment we can find the trend in effect of technology in schools. Select a school that does not have computers. Evaluate students' performance in the school. Now provide the students in the same school access to computers, but at the

same time allow students to choose whether or not to participate in computer classes. Evaluate the performance of students who use the computers compared to those who don't.

2.3.4. Links

http://trochim.human.cornell.edu/kb/quasiexp.htm - Complete description of quasi-experiment design
http://www.alleydog.com/101notes/methods.html - General description of Research Methods
Click Here - Short tutorial on quasi-experiment design

2.4. Case Studies

2.4.1. Description:

Case studies are somewhat different than traditional controlled experiments, but they still can fall under the same category. This design typically involves one person and many observations are made[3]. The experimenter chooses one behavior (the dependent variable) and measures it repeatedly. This is usually accomplished by what is called a time-series design, i.e. the participant is observed before, during, and after the independent variable is introduced. The goal is to examine one person and his/her behavior very closely. More complicated designs involve turning the independent variable on and off several times to observe the effects. In these situations the independent variable is usually something like a different font size or different organization of information.

2.4.2. Advantages/Disadvantages

The great advantage to case studies is the amount of control the experimenter has over the situation because there is only one person. Another advantage is that there is the potential to get ample, detailed data[3]. A challenge in the case study design is to get a good baseline of behavior before the independent variable is introduced. The behavior must be stable before the different font or different organization is introduced, otherwise you will not be able to determine if the change in behavior is due to chance or to the independent variable. Other disadvantages include poor generalization to other people or groups, experimenter bias in the selection of the individual to be observed, and the lack of robust statistics generally used. In many cases, experimenters will examine the data on a graph and determine if a significant change has occurred [2]. This is known as "eye-balling" the data and is not a reliable or valid statistical method.

3. Overall Advantages/Disadvantages

Controlled experiments are not appropriate for every research situation. However, there are several advantages if a research question can be tested in a controlled environment. Controlled experiments have greater internal validity than non-experimental methods; i.e. there is greater confidence that the results came from the manipulation of the independent variable and not from extraneous variance[3]. Another advantage is the ability to use robust statistical analyses like ANOVA, t-tests, and regression on the data. ANOVA and t-tests are used when there are two or more treatments groups and can reveal if there is a significant difference on the dependent variable between them. Regression is used when there is more than one independent variable and can determine which one was the most responsible for any differences between the treatment groups[2]. Finally, controlled experiments are the best for identifying causal relationships between the independent and dependent variables.

However, there are critics who say that the disadvantages outweigh the advantages [3]. One drawback is that the results are less generalizable to other situations than non-experimental research. Another disadvantage is that fewer research questions can be explored because of the amount of control involved. Not every research question can be brought into a lab and broken down into independent and dependent variables. There are critics who say that the artificial environment of a controlled experiment can distort participant behavior and that the findings are not applicable to real-world situations.

4. Guidelines

These are some suggestions for planning and running a successful controlled experiment:

1. Choose participants so they are as homogeneous as possible with respect to the independent variable
2. Design, plan, and conduct the research so that the experimental treatments are as different as possible
3. Reduce threats to internal validity
4. Make your hypotheses before you start collecting data
5. Run a few participants and collect pilot data to find any problems before conducting the full experiment
6. Choose the appropriate statistical tests depending on the amount and kind of independent and dependent variables
7. Set the confidence interval at $p < .05$

5. References

1. Myers, David G. *Psychology, Fourth Edition*. (1995), p. 2-39.
2. Howell, D. C. *Statistical Methods for Psychology, Fourth Edition*. (1997). Boston: Duxbury Press.
3. Kerlinger, F. N. & Lee, H. B. *Foundations of Behavioral Research, Fourth Edition*. (2000). New York: Harcourt College Publishers.
4. Brems Douglas J , Whitten II William b. "Learning and Preference for Icon Based Interface" *Proceedings Human Factors and Ergonomic Society Annual Meeting* 1987 pp 125-129.
5. Hix Deborah . "Assessment of an Interactive Environment for Developing Human Computer Interfaces" *Proceedings Human Factors and Ergonomic Society Annual Meeting* 1986 pp 1349-1353.
6. Perlman Gary. "Making the Right Choices With Menus" *INTERACT '84: First IFIP Conference on Human-Computer Interaction* 1984
7. Swierenga Sarah J Menuing and Scrolling as Alternative Information Access Techniques Proceedings Human Factors and Ergonomic Society Annual Meeting.1990 pp 356-359

This paper was written as a class project for Advanced Usability (Computer Science 838S) (Fall of 2001). The courses were led by Prof. Ben Shneiderman Founding Director of the University of Maryland Human-Computer Interaction Lab.

Authors: Liz Atwater & Ketan Babaria

Last updated October 30, 2001