The Effects of Distraction on Usability Testing Results in a Laboratory Environment

by

Brady A. Thrift

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Applied Computer Science

Guelph, Ontario, Canada

© Brady A. Thrift, September, 2012
ABSTRACT

THE EFFECTS OF DISTRACTION ON USABILITY TESTING RESULTS IN A LABORATORY ENVIRONMENT

Brady A. Thrift  Advisor:  Dr. J. McCuaig
University of Guelph, 2012

Hand held device users encounter various forms of distraction in their daily lives. Distractions may prevent them from correctly using the hand held device. Usability evaluations are meant to identify issues prior to end-users experiencing them. However the laboratory environment, which usability evaluations are conducted, may not reflect the real-world conditions that devices are used.

The experiment involved each participant performing tasks in both a quiet and a noisy environment. The noisy environment emulated part of a real-world experience by adding social noise in the background during the participant’s tasks. The goal was to compare how much insight each participant was able to achieve from the data in each of the environments.

It was found that task performance accuracy was higher in a quiet environment as opposed to the noisy environment. The mental demands and frustration of participants were found to be higher during the noisy environment evaluation.
Acknowledgments

I would like to express my gratitude to my supervisor Dr. Judi McCuaig, whose support and guidance throughout the entire duration of my thesis was invaluable. Her valued encouragement, assistance and direction helped me see my thesis through to its conclusion.

I would also like to thank Dr. Dan Gillis, whose statistical expertise helped me comprehend and examine my data. I was able to understand and analyze my non-parametric data with his help.

And lastly, I would like to thank Vassilena Alexandrova for her constant and continual support and motivation. From countless sleepless nights of proof reading to discussing psychological ideas and concepts in my background research, I greatly appreciate her time and effort. Words cannot express my gratitude.
Table of Contents

List of Tables vi
List of Figures vii

1 Introduction 1
1.1 Objective ................................................. 3
1.2 Outline ................................................ 4

2 Literature Review 6
2.1 Usability Testing for Hand Held Devices ................. 6
2.2 Environmental Effects on Performance ................ 10
  2.2.1 Noise ........................................ 13
  2.2.2 Temperature .................................... 15
  2.2.3 Illumination .................................... 17
  2.2.4 Crowding ....................................... 18
2.3 Visualization on Hand Held Computers ................. 19
  2.3.1 Insight ......................................... 20
  2.3.2 Measuring Insight ............................... 26
2.4 Cognitive Load ........................................ 27
  2.4.1 Defining Cognitive Load in a Hand Held Environment 28
  2.4.2 Evaluating Cognitive Load ....................... 29

3 Experiment Design 35
3.1 Overview ................................................ 35
  3.1.1 Experiment Overview ............................ 36
3.2 Rationale .............................................. 37
  3.2.1 Information Visualization ....................... 38
  3.2.2 Evaluation Environment ......................... 38
  3.2.3 Distraction ...................................... 39
  3.2.4 Variables and Measurement ..................... 40
3.3 Experiment Design .................................... 41
  3.3.1 Experiment Details ............................... 41
  3.3.2 Application Design .............................. 43
  3.3.3 Participant Characteristics ...................... 46
  3.3.4 Participant Walkthrough ......................... 47
  3.3.5 Materials ........................................ 48
# Experiment Results

4.1 Results ................................................................. 51
4.2 Analysis ............................................................... 58
   4.2.1 Normality ....................................................... 59
   4.2.2 Significance of Ordering ................................. 61
   4.2.3 Quiet Environment versus the Noisy Environment .. 62
   4.2.4 Performance and Cognitive Load ..................... 63
   4.2.5 Awareness of Environment ............................... 65

# Conclusion

5.1 Discussion .......................................................... 67
   5.1.1 Quiet and Noisy Environments ......................... 68
   5.1.2 Correlation Between Performance with Cognitive Load . 70
   5.1.3 Correlation Between Background Awareness and Accuracy and Cognitive Load ............... 71
5.2 Conclusion .......................................................... 72
5.3 Future Work .......................................................... 73

# References

# A Participant Experiment Documentation

A.1 Ethics Survey ....................................................... 79
A.2 Dataset Tasks .......................................................... 81
A.3 Ted Talk Selection List ............................................. 82
List of Tables

3.1 Dataset and environment ordering for each group of participants . . . 42
4.1 The characteristics of each participant and the group setup . . . . . 53
4.2 Participant task completion accuracy . . . . . . . . . . . . . . . . . 54
4.3 Participant task completion completion time . . . . . . . . . . . . . 55
4.4 Participant mental demands by task . . . . . . . . . . . . . . . . . . 56
4.5 Participant frustration by task . . . . . . . . . . . . . . . . . . . . . 57
4.6 The resulting p-values from the Anderson-Darling Test for Normality 60
4.7 A comparison between of medians between the quiet and noisy envi-
ronments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
4.8 Correlation tests using Spearman’s rank correlation coefficient on Per-
formance and Cognitive Load . . . . . . . . . . . . . . . . . . . . . . . . 64
4.9 Correlation tests using Spearman’s rank correlation coefficient on Back-
ground Awareness, Performance and Cognitive Load . . . . . . . . . 66
List of Figures

2.1 A screenshot of the ratings scales from the Nasa TLX application ... 32

3.1 An example screenshot of the application using the temperature data that was evaluated .................................................. 43
3.2 An example screenshot of the application using the population data that was evaluated .................................................. 44
3.3 An example screenshot of the application highlighting animation ... 45
3.4 An example screenshot of the application highlighting functionality . 46
3.5 A screenshot of the timing application developed to record the duration of each task .................................................. 50
Usability is an important area in Computer Science. It is the practice of designing and evaluating an interface where the focus is the end-user. How a user interacts with an application or website is critical. Poor designs can affect both the end-user as well as the company. Confusing or cluttered website designs, for example, may cause users to leave a company’s website resulting in lost revenue (Visser & Weideman, 2011). Badly designed interfaces may also waste a user’s time by not being able to efficiently move them through an interface. Usability evaluations and practices help ensure that the interface that a user interacts with is intuitive, efficient and easy to use.

Usability in a hand held environment is just as important as in a desktop environment. A developer of an interface of a hand held device, however, must take into account extra considerations and limitations when compared to a desktop computer. These developers need to account for different device capabilities, such as: varying screen sizes and ratios, colour palettes, audio input and output systems, hardware restraints and capabilities, and forms of user input. For example, a touch-screen device interface might have buttons on the screen appear larger than devices with physical controls so that a user to accurately navigate the interface. One technique
developers use to provide hand held device users with rich interfaces on small screen sizes is information visualization.

Information visualization is the visual representation of data. Any amount or type of data can be visualized. For example, a visualization can be as simple as a calendar, or as complex as a movie database. Graphics and animations tend to be utilized by developers to help users find information in the data. For example, animating the amount of sales of a product over time can give a person a better idea of any trends that are occurring than simply reading raw data. Evaluating and improving the designs of visualized information can help users efficiently discover previously unknown and important information. Providing the tools to assist users in quickly finding the information they are searching for is one of the goals of information visualization.

Usability practices for a hand held environment should consider variables that exist in the real world. For example, to save screen space an interface might require a user to mentally keep track or process some of the information. The evaluation might not show any usability issues with this when performed in a laboratory environment. However, usability issues may crop up when using it in a real-world environment. A user might be using the device in a crowded area where he or she is distracted by their surroundings and suddenly becomes unable to keep track of the information. Testing and evaluating hand held devices in the field can help identify real world usability issues with interfaces. However, field testing can cost extra time and make data collection difficult (C. M. Nielsen, Overgaard, Pedersen, Stage, & Steenild, 2006). An alternative to field testing, while still allowing for the discovery of
real world usability issues, is the simulation of real world conditions in a laboratory setting.

The goal of this thesis is important to the usability field as it may bring about a gradual change to more realistic testing situations. Current testing methods of hand held interfaces, conducted in a laboratory environment, do not take into account real world scenarios and distractions (external from the hand held device being evaluated) that may arise when using a hand held device. Evaluating an interface in a laboratory environment may lead the development team in one direction but actually using the device in the real world may show that the team should have gone in a different direction; this may force the team to go back and re-evaluate the design at a later stage. Evaluating an interface early in the development life cycle can save money and development time. It is more costly to revisit the design of an interface in the latter stages of application development than to discover usability issues early on (UXResearch, 2012).

Section 1.1 overviews the objective of this document. It includes the thesis statement and its importance. Next, section 1.2, outlines the remainder of the document.

1.1 Objective

This thesis aims to show that there is a difference between the outcome of a usability evaluation in a quiet (also referred to as a laboratory environment) and a noisy environment. Differing results between these two environments could
mean that: a) noise-based distractions have an effect on a user’s performance, b) the distractions in the noisy environment reflect real-world conditions and c) the noisy environment better identifies usability issues than the quiet environment. Showing the difference between the two testing environments will also illustrate the need for additional research to be conducted in other areas such as simulating real-world scenarios in a laboratory environment and identifying how much attention an interface requires. The other potential result from the findings of this thesis is that current evaluation methods for hand held devices in a laboratory environment may need to be reconsidered as they may not accurately detect issues with an interface since users are able to focus most or all of their attention on the device in the quiet environment than in the real world.

The thesis statement of this document is the performance of users during a hand held usability evaluation will be better when the evaluation takes place in a quiet environment as opposed to a noisy environment. Better performance in a quiet environment means that during an evaluation each user performed his or her tasks more accurately, in less time and with less cognitive load than in a noisy environment.

1.2 Outline

Chapter 2 of this document is a review of the current and past literature from environmental psychology and human-computer interaction. The first part of this chapter is an overview of usability practices relating to hand held devices. It examines the usability factors of visualizing information on a hand held device, the
constraints developers need to keep in mind when developing software to be used on a hand held device and how the environment can affect a user’s ability to use a hand held device. Next, there is an investigation of environmental effects on performance. It shows how various environmental distractions can affect a person’s performance. The last section of the literature review looks at cognitive load. It examines how a user’s attention is divided between his or her focus (such as using a hand held device) and surrounding environment.

Chapter 3 contains the experiment planning and design. It includes the rationale and design of the experiment. The section about rationale contains the hypothesis that was tested as well as the background of the experiment. It includes decisions and justifications regarding each component. The experimental design section identifies the procedures and the design elements of the experiment.

Chapter 4 provides the results and analysis of the experiment. The first section in this chapter details the results of the experiment. The second section is the analysis of the data collected from the experiment.

Chapter 5 contains the discussion, conclusion and future work. The discussion section considers the results and analysis from chapter 4 as they relate to the hypothesis. The conclusion shows what the thesis set out to accomplish and the final results of the work conducted. This chapter is brought to a close with examples of future work that could derive from the findings from this document.
Chapter 2

Literature Review

The goal of this literature review is to show that the environment in which a hand held device is used should influence not only the design of an interface but the usability evaluation as well. More specifically, it looks deeper into what affects a user’s performance using a hand held device out in a real-world environment (also referred to as in the field), in an attempt to simulate, in a laboratory environment (also referred to as a quiet environment), the factors that affect performance.

2.1 Usability Testing for Hand Held Devices

Usability is the study of human interaction with electronic devices, such as computers and hand held devices. User interactions are influenced by how well an interface is designed. Poorly designed interfaces can lead to users wasting their time because a user could be stuck looking for information. A user-centric interface allows users to intuitively navigate the interface and not make errors or waste time. Good designs can also affect visualized information in the same way. Well designed visualizations portray the information the user is searching for in an easy-to-understand manner.
The environment in which a hand held device usability study is conducted can be an important factor as to the type of issues discovered in an interface. Evaluating a hand held device in a laboratory environment may not bring to light all of the usability issues. An evaluation in the field however, forces real-world variables, such as distraction, upon participants that may affect the results of a study. Each environment has its own costs; this section examines the use of field testing to determine if it is worth the cost, time and effort.

There has been some controversy as to whether it is better to perform a usability evaluation in a laboratory or in the field. Some evaluations have shown that some usability evaluations performed in the field unearth the same issues with an interface as a laboratory environment, showing that it is a waste of resources to perform an evaluation in the field (Kjeldskov and Stage (2004), and Kaikkonen et al. (2005) as cited by C. M. Nielsen et al., 2006). These studies may have suffered from a small sample size of participants, varying data collection methods between field and laboratory environment testing, and varying task assignments between the two environments (C. M. Nielsen et al., 2006). Other studies have confirmed that an evaluation in the field will reveal usability issues that did not exist in the laboratory environment evaluation (C. M. Nielsen et al., 2006). Certain locations in the field may not yield the best results. Evaluating an interface in one location may also yield different results than another location. Golledge (1992) as cited by Reilly and Inkpen (2007) stated that one location may mean something to one person and something completely different to another. If a usability study is performed in the field, it has additional costs, including a significant time investment, difficult means
of accurately collecting data, and unforeseen and unknown variables appear (Baillie (2003), Johnson (1998), Kjeldskov and Stage (2004), and Kjeldskov et al. (2004) as cited by C. M. Nielsen et al., 2006; Duh, Tan, & Chen, 2006); for example, recording screen information and participant actions of the device can be difficult and data may be missed (Esbjørnsson et al. (2003), Johnson (1998), and Kjeldskov and Stage (2004) as cited by C. M. Nielsen et al., 2006). Missing data from recording in the field could be a result of background noise. The extra effort and cost of field testing can make it not seem worthwhile, especially if there are limited resources or time constraints.

C. M. Nielsen et al. (2006) believe that even though additional resources and time are required to conduct usability evaluations of hand held devices in the field, the end result was worth it since the researchers were able to discover usability issues that did not exist in laboratory environment testing.

Traditionally, a hand held device’s usability evaluation is performed in a laboratory environment where scenarios attempt to simulate the real-world (Duh et al., 2006). These scenarios try to identify usability issues that can occur when a user is using the device outside of the laboratory environment. Where is a suitable locale to perform a usability evaluation that will accurately reflect real-world conditions? General-use visualizations, such as a calendar, have no set environment that they are used in. Looking at visualized data in a subway station may have different usability implications than at a park, for example. Conversely, if data is visualized for a specific use, then a field evaluation may mirror similar circumstances within which it would normally be used.

A study done by C. M. Nielsen et al. (2006) compared the results of a
laboratory environment versus field testing; the hand held system that the researchers evaluated was a specialized system to be used by workers in the field and was not for general usage. The experiment data was collected visually (by camera) and by timestamped logs, where both settings had identical collection methods (C. M. Nielsen et al., 2006). The evaluations that took place in each setting were conducted in the same manner, to eliminate the possibility that the results would be affected by the setup. The field tests showed more usability problems with the interface when compared to the laboratory environment tests (C. M. Nielsen et al., 2006); the field tests also demonstrated a much higher time to complete tasks than the laboratory environment tasks (C. M. Nielsen et al., 2006). It was also found during the field tests that some usability issues were unique to the field setting; they were not found in the laboratory environment tests (C. M. Nielsen et al., 2006).

Reilly and Inkpen (2007) used two mapping techniques and compared usability tests of the application between laboratory and real-world environments. Participant recall was the dependent variable; it was measured by providing questionnaires to participants after tasks with each map were completed. The one map interface scored higher in the real-world environment whereas the other map interface scored higher in the laboratory environment (Reilly & Inkpen, 2007). Reilly and Inkpen (2007) found that the setting can impact the results of the evaluation of visualization techniques.

Duh et al. (2006) performed a usability evaluation with a hand held device that compared the results between a field test and a laboratory environment test. The number of critical and serious problems found with the device were significantly
higher in the field as opposed to in the laboratory environment (Duh et al., 2006). Duh et al. (2006) also found that the time to complete tasks was considerably longer in the field. They concluded that some issues and behaviours were only found in the field (Duh et al., 2006). Environmental factors, such as noise, movement (the field tests were conducted in a train), privacy, and mental and physical loads, greatly affected each subject’s ability to perform well and thus brought about usability issues with the device (Duh et al., 2006).

Reilly and Inkpen (2007) said that “visual perception and cognition are heavily influenced by contextual factors”. One major factor affecting a subject’s performance is the external setting around the participant (Bell et al. (1990) as cited by Reilly & Inkpen, 2007). Environmental psychology examines and analyzes the effects that the environment has on a person; the “...setting can impact visual perception, concentration, and reasoning in unpredictable ways” (Reilly & Inkpen, 2007). Duh et al. (2006) found that environmental effects, such as sound, for example, had a significant impact during the usability evaluations that they performed. These effects are the essence of field testing: how the environment affects performance.

2.2 Environmental Effects on Performance

The physical environment where a person uses a hand held device can impact the performance of how the device is used. Factors such as quality of lighting, sound cancellation, and the layout of a person’s space can have an influence, affecting productivity, efficiency, and accuracy (Holahan, 1987, p. 123). A wide-open space,
such as an outdoor environment for example, can be detrimental to a person’s ability to learn and perform well (Holahan, 1987, p. 123–124). This section’s focus is to illustrate the factors that can affect human performance and how distraction can play a role in mental activities.

Steele (1973) as cited by Holahan (1987, p. 126) defines three categories of types of performance: physical (external activities outside of the subject), mental (internal activities inside of the subject), and interactional (activities that occur between the subject and one or more people). Steele (1973) as cited by Holahan (1987, p. 126) describes the physical category as containing variables that affect physical conditions in the outside environment. For example, temperature can be defined as a variable in the physical category. Next, Steele (1973) as cited by Holahan (1987, p. 126) claims that the mental variable involves a person’s cognitive abilities. Factors such as noise and lighting can influence a person’s mental capabilities. Steele (1973) as cited by Holahan (1987, p. 126) identifies the last variable, interactional, as how a person is affected by the interactions in their surrounding space. Crowds are one factor that can have a significant effect on a person. The variables that attribute and define the performance categories can be summed up in two environmental variables: ambiance and layout. These two variables, that directly affect and influence a person’s performance, are the primary focus of this section.

The ambient environment is made up of characteristics such as light, sound and temperature (Holahan, 1987, p. 124–125). These features can have a large impact with performance and learning. An example of how the environment affects learning and performance is illustrated when a storage room at a grade school in Texas was
converted to a classroom. The noise, lighting and heat in this environment was shown to affect performance as overall student performance had a sudden significant drop (Holahan, 1987, p. 125). Temperature induced restlessness in the students because of heat in the room due to the lack of sufficient air conditioning (Holahan, 1987, p. 125). Noise from the adjacent room (a gymnasium) heavily distracted the students who, as a result, were unable to focus during class (Holahan, 1987, p. 125). Lastly, the students complained of being unable to read the blackboard. The root of this issue was found to be the lighting; while adequate for storage purposes, it was not sufficient in the classroom environment (Holahan, 1987, p. 125).

The layout of the environment is composed of physical objects such as features or people and their location (Holahan, 1987, p. 125). Objects, their position and orientation in the environment can have a direct influence with human performance. An example of layout affecting performance is when a hospital moved offices to another location with a wide-open design with no walls or partitions; several people complained that they could not work effectively in an open office environment (Holahan, 1987, p. 125–126). A lack of sound dampening, noise from typewriters and chatting could be heard by everyone in the room, causing a significant amount of distraction (Holahan, 1987, p. 125–126). While this example showed that environmental factors can have a negative impact with performance and learning, it is possible that they can also increase performance.
2.2.1 Noise

Noise has a major effect on human performance (Cohen & Weinstein, 1982, p. 45). Noise is any sound that is undesired by the listener that can interfere with his or her focus (Kryter (1970) as cited by Cohen & Weinstein, 1982, p. 46) (Holahan, 1987, p. 134) (Canter & Stringer, 1975, p. 64). One example of noise being detrimental to performance is when Lawrence, Ward and Suedfeld (1973) as cited by Holahan (1987, p. 165) caused student participation to drop significantly by playing very loud, pre-recorded, traffic noises nearby. A second example was when Cohen, Glass and Singer (1973) as cited by Ahrentzen, Jue, Skorphanich, and Evans (1982, p. 233–234) discovered that the noise from trains was causing the students, who were at one end of the school closest to the train tracks, to do worse academically than the students further away from the tracks. The students began to develop issues with filtering out both useful and background noise, which resulted in “...poor auditory discrimination” (Ahrentzen et al., 1982, p. 234). This may be an example of information overload. When a subject is exposed to noise, it may overload his or her cognitive abilities to handle all of the information, including the task(s) at hand, as found by Cohen (1978) as cited by Holahan (1987, p. 184–185) and Cohen and Weinstein (1982, p. 48). Performance is lost due to noise because of the cognitive resources allocated to process the noise; being aware of noise ensures that a person is unable to provide their complete focus on a task.

state that certain noises can affect performance under certain conditions (Cohen & Weinstein, 1982, p. 46). These conditions include: “type of task, noise characteristics, and timing” (Holahan, 1987, p. 135). Complex tasks (Bogges and Simon (1968), and Eschenbrenner (1971) as cited by Holahan, 1987, p. 135) and tasks that require focus and concentration (Broadbent (1957), Eschenbrenner (1971), Sanders (1961), and Theologus, Weaton and Fleishman (1974) as cited by Holahan, 1987, p. 135) are two types of tasks that can be disrupted by noise. The characteristics of noise that cause the greatest degradation of performance is when the noise is random and noncontinuous (Broadbent (1957), Eschenbrenner (1971), Sanders (1961), and Theologus, Weaton, and Fleishman (1974) as cited by Holahan, 1987, p. 135) (Canter & Stringer, 1975, p. 74). Acton (1970), and Obszewski, Rotton and Soler (1976) as cited by Holahan (1987, p. 135) found that intelligent speech was actually the most disruptive noise; human evolution has a major part in this since the frequency of human speech falls into where the ear’s sensitivity is at its best (Canter & Stringer, 1975, p. 58). As the duration of exposure to noise that a subject endures increases, the subject’s performance increasingly gets worse (Hartley (1973), and Hartley and Adams (1974), as cited by Holahan, 1987, p. 135). It can be concluded that long durations of intelligent human speech cause the most significant performance issues with people.

Intelligible speech is the most distracting type of noise. A study done by Brunetti (1972) as cited by Ahrentzen et al. (1982, p. 234) demonstrated this by showing that social noise affected students more than general noise. The probable cause is that social conversations are something that may pique the interest of the
subject and thus turns their attention to it instead of the task at hand. Elder et al. (1979), Goodrich (1979), Nemecek and Grandjean (1973), and Wineman (1981) as cited by Wineman (1982, p. 266) found that phone conversations taking place in the office space were the most distracting types of noise to coworkers; Nemecek and Grandjean (1973) as cited by Wineman (1982, p. 266) furthered this claim and showed that adjacent conversations were the highest distractions because of their informational content. While noise degrades performance, speech seems to affect it the most. Social or interesting speech cues are involuntarily picked up by the brain and the subject is forced to devote some of their cognitive processing ability to understand what is being said.

Noise is a significant stressor and can cause performance degradation in subjects. The most pronounced effects are those when the subject is performing a task while there is intelligible speech; while other forms of noise have an impact with a subject, speech that is interesting to a subject forces him or her to divert some of their cognitive processing to understand what is being said.

2.2.2 Temperature

Temperature ranges that are commonly found in indoor environments do not usually cause issues with human performance (Holahan, 1987, p. 137); on average, a person is most comfortable between 26°C and 28°C (Mackworth (1950) as cited by Canter & Stringer, 1975, p. 47) (Holahan, 1987, p. 137). Human subjects with different characteristics such as age, gender and geographical location showed no difference in the range of comfortable temperatures, which was studied by McNall et
al. (1967), and Fanger (1972) as cited by Canter and Stringer (1975, p. 42) and Fanger (1972) as cited by Holahan (1987, p. 137) (Bell & Greene, 1982, p. 82). A person using a hand held device in an outside environment may not experience temperatures in their comfort range. Tasks requiring complex mental or extraneous physical effort see a greater change in performance when the temperature is 5°C above or below a person’s comfortable temperature (Pepler (1963) as cited by Canter & Stringer, 1975, p. 47). Experiencing environmental temperatures too hot or too cold may either increase or decrease mental and physical performance.

Several contradictory studies have found that heat either has a negative effect, positive effect, or no effect with human performance (Holahan, 1987, p. 137–138). Common findings on performance changes, which are affected by heat, are dependent on the complexity of the task, the amount of heat and the length that a subject is exposed to the heat (Holahan, 1987, p. 138). Some studies found that the more complex the task, the more that performance decreases in heat (Bell & Greene, 1982, p. 84) (Holahan, 1987, p. 138). Another author, Wyon (1974) as cited by Holahan (1987, p. 138), found that heat temperatures above a person’s comfort level predictably hinder performance. However, Griffiths and Boyce (1971) as cited by Canter and Stringer (1975, p. 49–50) “found that there is no linear relationship between temperature and performance” during their experiment, where heat was gradually increased and performance kept shifting between a positive and negative factor. While the results seem to vary from experiment to experiment, performance was affected by temperatures in one way or another.

The effect of cold on a subject’s performance is attributed to physical detri-
ments. Fox (1967) as cited by Bell and Greene (1982, p. 85) found that the sense of touch, dexterity and a person’s strength are all affected negatively by cold. There is much less research performed with how cold affects performance, however (Holahan, 1987, p. 138). Some research has found that physical exertion tasks, such as lifting, are affected drastically by cold (Holahan, 1987, p. 138). Since the research that focuses on cold and performance features physical and not mental effects it is not known how the cold will affect the use of a hand held device.

2.2.3 Illumination

Lighting and illumination have a major influence with performance. Changing the amount of light in an environment can affect a person’s visual acuity (Canter & Stringer, 1975, p. 86). Boyce (1973), and Muck and Bodmann (1961) as cited by Canter and Stringer (1975, p. 103) discovered that increasing the level of light increased a subject’s performance; the increase in performance was dependent on the level of light and the difficulty of the task.

Markus (1965), and Bitter and Van Ierland (1965) as cited by Canter and Stringer (1975, p. 120) performed a study and concluded that artificial light was not preferred over daylight and direct sunlight. While natural light is the preferred light-source, it does not necessarily mean it is the best environment to evaluate a hand held application. Hand held devices are used in both and indoor and outdoor environments.

Light can be a detriment to performance if there is too much glare (Canter & Stringer, 1975, p. 83–85); Bennett (1977) as cited by Holahan (1987, p. 133) states
that glare exists when the light in the surrounding environment is of a higher intensity than that of the object at which the subject is looking. Glare hinders a person’s ability to discern colours because of over-saturation (Canter & Stringer, 1975, p 83–85). Using a hand held device in direct sunlight produces significant glare and can make colours on the screen not visible to the eye.

2.2.4 Crowding

Crowding refers to a subject’s personal restrictions with the space around them, defined by Stokols (1972) as cited by Holahan (1987, p. 198). This means that people, for example, may be densely arranged around a subject, but the subject may not feel uncomfortable; Rapoport (1975) as cited by Holahan (1987, p.198) further explains crowding is when a subject feels that his or her surrounding space is not adequate to feel comfortable. The problem that a person experiences when they feel crowded is that “...the task of managing and coordinating that environment increasingly drains attention ordinarily available for goal attainment” (Epstein, 1982, p. 133–134). As a person begins to feel more and more crowded in an environment, the more they have to devote their attention to their surroundings. Crowding can be considered a form of distraction when using a hand held device, since attention is required in order to manage and coordinate the environment.

Complex problems or the need to process a large amount of information in a social or spatially dense area caused significant performance detriments, as studied by Bray, Kerr and Atkin (1978), Evans (1979), Paulus and Matthews (1980), and Paulus et al. (1976) as cited by Holahan (1987, p. 213). The number of people in
a given subject’s immediate surroundings is referred to as social density (Holahan, 1987, p. 199). Spacial density, on the other hand, is the space available in a subject’s surroundings that are available (Holahan, 1987, p. 199). Crowding is a significant stressor and causes frustration in a subject, especially when he or she is focusing on a task.

2.3 Visualization on Hand Held Computers

Information visualization is the process of transforming data into something that is not only readable to an end-user, but is also usually interactive (Fuchs & Schumann, 2006). Normally, any well-designed visualized data should show data that the user needs. It must be able to lead a user to generate insight from the data and must be efficient as to not waste unnecessary time or resources (Fuchs & Schumann, 2006). For example, displaying extraneous data not relevant to the user’s tasks and goals may delay the subject from discovering what they are looking for in the data or it may lead him or her to incorrect conclusions. Secondly, the creation of insight is the main reason as to why data visualization exists (North, 2006; Yi, Kang, Stasko, & Jacko, 2008). And, lastly, the visualization should improve the efficiency in learning from the data as opposed to just looking at the raw data; if the speed and efficiency that users are able to gain knowledge from visualized data is the same as or slower than from raw data, then the visualization is not adequately performing its function.

Visualizing data with a hand held device is more difficult than with its
non-portable counterpart. The developer must take into account screen-size and aspect ratio, limited hardware constraints, battery life, slow or limited data transfer constraints (Fuchs & Schumann, 2006; Yoo & Cheon, 2006; Karstens, Rosenbaum, & Schumann, 2003; Gutwin & Fedak, 2004). As computer technology advances, most of these problems are becoming insignificant. For example, the processing power and memory in newer hand held devices are able to run and render high-quality graphics (Capin, Pulli, & Akenine-Moller, 2008). These improvements can greatly affect how information can be conveyed to users.

The developer must keep in mind that the capabilities of a hand held device can vary from device to device. For example, utilizing and focusing on colour as a way to convey data may work exceptionally well with a brand new smartphone. However, this method may not work with older devices since these devices tend to have a lower resolution or lower depth of colour. These problems or limitations can be amplified when the environment a hand held device is used is taken into account. Usability problems may arise with an application that relies on vibrant and sharp colours, for example, when the device is used under certain lighting conditions. This usability issue may not be revealed during usability testing due to the environment where the device is evaluated.

2.3.1 Insight

When a user finds a previously unknown pattern or trend in data, this phenomenon is called insight. The term refers to the discovery of a pattern or new information discovered inside a given set of data (Pike, Stasko, Chang, & O’connell,
information is visualized for this reason: to make insight achievable as well as easier and/or faster to attain. Insight is the objective that the developers of an information visualization interface hope that the end-user achieves (Card et al., 1999, p. 6). The information visualization community however, has had trouble finding a proper definition and measurement of insight (Yi et al., 2008; Chang, Ziemkiewicz, Green, & Ribarsky, 2009; North, 2006). The importance of a definition cannot be understated. It would allow tools and techniques to be created and standardized to measure and compare information visualization interfaces. Even though the community cannot come to a consensus on a definition, the basic concept of insight can be employed to determine what the user has learned from the data.

There is no agreed upon definition used by professionals and researchers in this field (Yi et al., 2008). The lack of consensus between professionals on a clear, concise definition has lead to a number of problems. First, there is no way to determine if a testing method to measure insight is valid. One common testing method that North (2006) found is that many researchers asked participants if they liked the visualization; this method does not yield a way to determine if the participants correctly discovered insight in the data or anything quantifiable about the visualization; it is a qualitative metric. The evaluators would have no way of knowing if the insight gained from the data could be complete or correct. Second, a clear definition of insight would enable comparison of different visualizations. Without a definition, there is no way to determine if one visualization yields better insight with data over another. For instance, one developer could state that his or her visualization provides
better insight into a specific dataset; another visualization, however, may use different insight metrics and thus achieve different results. Comparing the results of these two visualizations would lead to incorrect conclusions as to which provided better insight. Defining a clear and concise definition that information visualization designers can use would alleviate both of these issues.

Many authors have published characteristics and definitions in an attempt to identify how to measure and define insight (Yi et al., 2008; North, 2006; Chang et al., 2009). Instead of defining insight, North identifies important characteristics that an insight must fulfill to be valid. He claims that in order for an insight to be meaningful or significant, it must fulfill all five of his characteristics, which include complexity, deepness, quantitativeness, unpredictability, and relevance (North, 2006). Regarding complexity, he states that “[i]nsight is complex, involving all or large amounts of the given data in a synergistic way...” (North, 2006). The visualization, as a whole, leads a person to discover insight from the data. If the visualized data is not aggregated to form a whole in a logical way, then it is unlikely that insight or valid insight will be found. The definition of complexity does not account for a simplistic answer to a question that a user might be seeking. For example, a person may be looking for the time of a meeting in a calendar. Bederson et al. (2003) developed a calendar that utilizes the fisheye view, a form of visualizing information, to condense and expand relevant information to the user. This type of visualization provides simple insight that does not involve all of the data, only a piece of datum. It does not mean that since the insight is not complex that it is not valid.

North’s next characteristic of insight, deep, affirms that “[i]nsight builds
up over time, accumulating and building upon itself to create depth” (North, 2006). When insight is gained, it should lead a person to further questions and insights. While this may hold true for complex insights, simple insights may just be an answer to a single question. Karlson et al. (2005) developed AppLens, a multi-level fisheye visualization to display contextual information of icons using a hand held device. The application does not allow for deep insight; the information is available to the user at a glance, which does not bring about further questions. The third characteristic, qualitative, claims that “[i]nsight is not exact, can be uncertain and subjective, and can have multiple levels of resolution” (North, 2006). The developers of a visualization may not know what kind of insights will be found in the information that is visualized; the insights gained may also vary from user to user as it is subjective. Most insight may very well be subjective, varying from user to user, but some visualizations may only allow for one answer. For example, the Halo technique developed by Baudisch and Rosenholtz (2003) is a map visualization that allows for objects outside of screen space to be displayed, relative to its location. The fourth characteristic, unexpected, describes insight as “... often unpredictable, serendipitous, and creative” (North, 2006). A person does not know what the insight will be nor will they know how or when it will occur. Lastly, the relevance of “[i]nsight is deeply embedded in the data domain, connecting the data to existing domain knowledge and giving it relevant meaning” (North, 2006). All insight must have a larger meaning in its domain; insight is significant. It can simply be a new piece of information that a person has learned. However, insight attained from data that is simple (such as finding an event in a calendar) does not fit into the categories described. The categories expect that all
insight will be significant and meaningful as even a single datum may lead to insight.

Chang et al. (2009) draw similarities and differences between insight as measured in cognitive science and insight in information visualization. Insight definitions in cognitive science relate to the surprise moment, or eureka moment, when a person realizes a piece of information that he or she did not know before (Mai et al. (2004) as cited by Chang et al., 2009). Saraiya, North, and Duca (2005) define insight as “an individual observation about the data by the participant, a unit of discovery”. The cognitive science community sees insight as an experience whereas the visualization community sees insight as a tangible piece of knowledge (Chang et al., 2009). Chang et al. (2009) put forward that any attempt to define insight should take into account both definitions, one from the visualization community and one from the cognitive science community; insight is the eureka moment, where knowledge is furthered or a piece of information is learned. The combination of definitions works well for complex insights, but only the visualization part of the definition applies to simplistic insights. The combination of definitions relies on the idea that the two types of insight lead to one another. The eureka moment leads to knowledge and knowledge leads to the eureka moment. However, there may not be a eureka moment with simplistic data or a simplistic visualization.

Yi et al. (2008) describe insight as a process or product of “...exploration without an initial destination”. Instead of trying to produce another definition, as many other authors have done, they attempt to identify how insight is gained. The four processes that they have found to bring about insight are: provide overview, adjust, detect pattern, and match mental model. The first part of the process is
where a person understands the visualization of the data as a whole (Yi et al., 2008); in essence, it allows a person to understand what they know about the data. The next step in the process is an exploration by changing his or her perspective of the visualized information (Yi et al., 2008); this process allows for a person to find interesting and useful information in the data. The third step in the process, detect pattern, is where a person finds relevant information in the data in the form of patterns or trends (Yi et al., 2008); these discovered patterns or trends can be new or existing information to the subject. The last step in the process, match mental model, is a mental visualization of the data in the user’s head (Yi et al., 2008); this mental image of the data allows a person to understand the data and linking it to a person’s current ideas and knowledge. While this process of understanding insight provides a guideline as to how insight is formed, it does not define insight; instead, this process of understanding insight is a good step forward to defining it.

Defining insight has been a troubling task in the visualization community. Authors have put forward definitions of insight that attempt to draw upon the complexity of the results; many of these authors succeed except for the fact that not all insight is complex or deep. The purpose of any visualized information is for a user to achieve insight (Card et al., 1999, p. 6). This means that insight may be achievable from any visualization. From this, it can be concluded that visualizations that provide users with simple answers to questions are in fact valid; however, these visualizations and the insight achieved from them do not fit with many of the definitions or characteristics of insight. Definitions of insight need to account for simplistic results. The most valid definition of insight is also the simplest one: “insight [is] an individual
observation about the data by the participant, a unit of discovery” (Saraiya et al., 2005). This definition can be applied to simple insight as well as complex insight.

### 2.3.2 Measuring Insight

The measurement of insight is important. Measurement allows for evaluators to determine how well a visualization triggers insight in a person. Methods such as evaluating a visualization with “like” or “dislike” responses yield very little information about how well a person is able to find information. North (2006) found many usability evaluators assess their visualizations with boolean metrics, where participants either “liked” or “disliked” the evaluation; an evaluation like this offers no results that can be compared to other evaluations or determine if it is performing its function adequately.

A controlled experiment with tasks has been a staple usability evaluation method of the information visualization community for a long time (Chen and Czerwinski (2000), and Chen and Yi (2000) as cited by Saraiya et al., 2005; Plaisant, Fekete, & Grinstein, 2008). A controlled experiment consists of using independent and dependent variables to conduct an evaluation; the independent variables are adjusted to see how dependent variables are affected. Several of the dependent variables that can be measured quantitatively are accuracy and timing (Saraiya et al., 2005; Tory & Moller, 2004; Huang, Hong, & Eades, 2006). Time is usually measured on a per-task completion basis, from the start of the task until insight is discovered. Accuracy measures the precision, error rate(s), and the number of correct and incorrect responses (Saraiya et al., 2005). The dependent variables that can be qualitatively
measured typically include the participants’ experience and preference with the interface and visualization (Tory & Moller, 2004; Huang et al., 2006). They can be measured with questionnaires, surveys, interviews or other interactions.

Open-ended insight evaluations with complex tasks given to participants facilitate the discovery of insight better than controlled experiments (North (2006) and Saraiya (2004) as cited by Plaisant et al., 2008; Saraiya et al., 2005). Open-ended insight evaluations however, should not replace controlled experiments. Controlled experiments allow for particular tasks to be tested and evaluated (Saraiya et al., 2005). Open-ended evaluations also suffer a number of drawbacks. First, open-ended evaluations are labour intensive (Saraiya et al., 2005). It requires that facilitators and experimenters devote a lot of time to capturing and identifying insights found in participants. Next, it requires that experimenters be trained in the domain of the data (Saraiya et al., 2005). For example, an experimenter would not be able to conclude whether insight was discovered if he or she does not understand the field that the data applies. Participants must be self-motivated to find insight (Saraiya et al., 2005). The discovery of something in data may take a lot more work than a participant is willing undertake.

2.4 Cognitive Load

Speech, as a distraction, in a hand held user’s surrounding environment is an everyday occurrence. Distraction can effect a hand held user’s performance and focus because of cognitive load. Defining and measuring cognitive load in a hand held
computing environment can identify the mental exertion users experience when using a device.

2.4.1 Defining Cognitive Load in a Hand Held Environment

Cognitive load is the demand on a person’s cognitive resources, or working memory, that is allocated towards performing tasks (Huang et al., 2006; Dittrich & Stahl, 2011). Cognitive load can be broken into three different categories, which include: intrinsic, extraneous, and germane cognitive load (Hollender, Hoffmann, Deneke, & Schmitz, 2010). Intrinsic cognitive load is the innate difficulty of the information presented to a subject (Hollender et al., 2010). For example, simple arithmetic will have a low intrinsic load compared to calculus, which has a much higher intrinsic load. Adjusting to this type of load is difficult, since it is inherent to the data itself. Germane load is the load required to build a schema of the information in the participant’s mind (Hollender et al., 2010). Extraneous cognitive load is caused by the method in which information is delivered to a person (Hollender et al., 2010). For example, requiring a user to multi-task will have significantly higher extraneous load over a single task. The summation of these three parts yields the total cognitive load (Paas et al. (2003) as cited by Hollender et al., 2010). In order to learn new information, the three types of cognitive load cannot exceed a subject’s cognitive abilities. If one of the three types requires a lot of cognitive load, then the other two types would need to be proportionately less.

Of the three types of cognitive load, extraneous cognitive load is the most affected by the design of hand held devices and applications. Since this type of cog-
ognitive load is based on the method that information is delivered to a person, making an interface simpler, for example, can reduce the cognitive requirements required of a user. However, the amount of cognitive load required depends on the system being used (Huang et al., 2006); some applications require a higher cognitive load than others. Surgeons, for example, need to focus most, if not all of their attention on surgery and are not able to devote a significant cognitive load to a surgery planning tool (Tory & Moller, 2004). Devices that require high cognitive load leave less resources available for a person’s external environment. An environmental distraction could disrupt a user and cause cognitive resources that were allocated to the device to instead be focused on his or her environment. The change in focus, or a partially divided focus, could cause errors, performance loss and the loss of information in short-term memory (Huang et al., 2006). Since distractions in a user’s external environment are uncontrollable, the focus should be on designing and testing hand held applications to account for these distractions and to reduce the required extraneous cognitive load.

2.4.2 Evaluating Cognitive Load

The goal of any workload measurement tool is to find the mental exertion of a task on an individual. There are three categories that these tools fall into: performance, subjective, and physiological measures (Meshkati, Hancock, and Rahimi (1992) as cited by Rubio, Diaz, Martin, & Puente, 2004). A performance measure is one where difficulty increases demands and, in-turn, decreases performance (Rubio et al., 2004). Subjective measures are those where the perceived effort of an individual
can be evaluated and measured (Rubio et al., 2004). Lastly, a physiological measure is where mental effort can be measured by the level of physiological activation (Rubio et al., 2004). Subjective measures are becoming increasingly more common than the other metrics because of their non-intrusiveness (Rubio et al., 2004). Some of the most common subjective measurement tools include SWAT (Subjective Workload Assessment Technique) (Reid and Nygren (1988) as cited by Rubio et al., 2004), NASA-TLX (Task Load Index) (Hart and Staveland (1988) as cited by Rubio et al., 2004), and the Workload Profile (Tsang and Velazquez (1996) as cited by Rubio et al., 2004).

The success of any workload assessment tool depends on the level it meets within the following criteria: sensitivity (the ability to ascertain changes in the difficulty or demands in a task), diagnosticity (the identification and justification of changes in workload variation), selectivity/validity (take into account only differences in cognitive demands), intrusiveness (the tool should not interfere with performance), reliability (the tool should be consistent), implementation requirements (the implementation of certain features such as time, instruments, and data collection and analysis), and subject acceptability (is the subject's view and perception of the tool) (Eggemeier (1988) as cited by Cao, Chintamani, Pandya, & Ellis, 2009) (Eggemeier et al. (1991) as cited by Rubio et al., 2004). The requirements listed here ensure that a tool accurately captures the correct workload of the participant and does not interfere with the results, by causing additional load on the subject.

Nasa-TLX is a tool that measures subjective workload (Cao et al., 2009); using six scales, it allows for the measure of mental demands, physical demands, tem-
poral demand, performance, effort, and frustration level (Hart & Staveland, 1988). Mental demand is measured with a low/high scale and it is a requirement of mental and perceptual demand during the task (Hart & Staveland, 1988). The second subscale, physical demand is also measured from low to high, gauging how much physical exertion was required for the task (Hart & Staveland, 1988). Temporal demand, also measured from a scale of low to high, is the time constraint that the participant felt that he or she was under (Hart & Staveland, 1988). Scale is measured from poor to good and is the participants level of success at the task (Hart & Staveland, 1988). The fifth subscale, effort, is measured from low to high and is based on how much mental and physical work was required of the participant during the task (Hart & Staveland, 1988). The last subscale, frustration level, measures the participants level of frustration and stress during the task (Hart & Staveland, 1988). The figure below (Figure 2.1) is a screenshot taken of the rating scales from the Nasa TLX application.
Figure 2.1: A screenshot of the ratings scales from the Nasa TLX application

SWAT (Subjective Workload Assessment Technique) measures the load of time, mental effort, and psychological stress at either low, medium or high values (Reid and Nygren (1988) as cited by Rubio et al., 2004). The time rating is the amount of time the participant felt that he or she had to complete the tasks (Rubio et al., 2004). Mental effort measures the amount of work that the participant felt that he or she spent on the task (Rubio et al., 2004). Psychological stress measures the various mental weights placed upon the subject, such as confusion, risk, frustration, and anxiety (Rubio et al., 2004). The assessment itself is a multiple choice questionnaire. It is done by having the participant select the answer that best describes workload that was experienced. For example, the participant may have a choice between often has spare time, occasionally has spare time, and almost never has spare time when
the time rating is assessed.

The Workload Profile has subjects provide the ratio of attentional resources used on each task, in a random order, after all of the tasks have been completed (Tsang and Velazquez (1996) as cited by Rubio et al., 2004). The various attentional resources used in this tool, provided by Wickens (1987) as cited by Rubio et al. (2004), are: perceptual/central processing, response selection and execution, spatial processing, verbal processing, visual processing, auditory processing, manual output, and speech output (Rubio et al., 2004). Definitions are given to the participant prior to using the tool. After the user completes the document that identifies the resources used, each of the resources are added together to find the workload of the series of tasks.

Rubio et al. (2004) performed an experiment, comparing Nasa-TLX, SWAT, and Workload Profile. In terms of intrusiveness, each of the tools required a similar amount of disruption to the task(s) being evaluated (Rubio et al., 2004). Workload Profile was the tool that showed the most sensitivity to changes of difficulty in the task, with Nasa-TLX performing only slightly better than the SWAT (Rubio et al., 2004). The Workload Profile showed that it had the best accuracy in terms of the difference between attention resources required by tasks (Rubio et al., 2004). Both Workload Profile and SWAT experienced issues in regards to participant acceptance. With the Workload Profile, subjects experienced confusion about the dimensions of the tool (Rubio et al., 2004). They concluded that each of the tools has its preferred situations as to where it will work best. The Workload Profile worked the best when the evaluator is comparing the mental load of more than a single task with varying
levels of difficulty (Rubio et al., 2004). Rubio et al. (2004) recommended the Nasa-TLX tool if the objective is to “predict the performance of a particular individual in a task”. Lastly, they recommended that either Workload Profile or SWAT be used to analyze cognitive demands or attention resources of a task (Rubio et al., 2004).

This literature review has examined the usability on a hand held device, environmental effects on performance, information visualization and cognitive load. The first section, 2.1, provided an comparison between laboratory and field testing. Even though field testing comes with additional costs to that of laboratory testing, the usability issues discovered are well worth it. Section 2.2 covered four of the major types of distraction that a person might experience in his or her external environment. Noise, illumination, temperature and crowding all have a major role in distracting a person from his or her tasks. Intelligible speech or social noise was found to be the most distracting type of noise. Next, section 2.3 provided details on information visualization on hand held devices. Insight was found to be the primary goal of information visualization. The last section, 2.4, dealt with cognitive load. One part of the section looked at devoting resources to a hand held device and diverting those resources to the external environment due to a distraction. The other part examined various methods to measure the cognitive load of a participant; it covered tools such as Nasa TLX, SWAT and Workload Profile. Chapter 3 contains the design of the experiment; it uses information from this literature review to help plan and justify each part of the experiment.
Chapter 3

Experiment Design

This chapter of the thesis details how the experiment was performed. Section 3.1 gives a brief overview of the experiment. It contains a general description of the experiment and the processes undertaken. Section 3.2 contains the reasoning behind the decisions made during the planning of the experiment. The last part of this chapter, section 3.3, thoroughly details the experimental design. It provides information about participant selection and use, the materials used, and the experiment details.

3.1 Overview

The goal of this experiment was to compare quiet and noisy environment testing when evaluating a visualized data interface. Participants’ insight (also referred to as accuracy in this experiment) was measured in each environment to show that the environment, in which an interface was evaluated, affected the outcome of the test. If the environment used to conduct an usability evaluation of a data visualization affected a participant’s ability to gain insight, then testing in a quiet environment lead to more accurate insight, less time taken to achieve insight and less cognitive
load utilized than in a noisy environment.

3.1.1 Experiment Overview

The first part of the experiment involved developing a visualization to be evaluated and planning metrics and tasks to help evaluate the performance of participants in each environment. Other initial steps in the experiment included finding a noise distraction to use in the noisy environment and recruiting participants for the experiment.

The visualized data could not be too simple because an interface that is too simple may not demonstrate the differences of the participants’ performance in each of the two testing environments. However, visualized data that is too complex may prevent participants from successfully being able to discover information from the visualizations in either environment or causing them to lose interest in finishing the experiment that is too difficult.

The metrics used needed to reflect and help identify differences in performance and cognitive load in each of the evaluation environments. The two performance measures were the time it took the participants to complete each task and the accuracy of the answers that each participant provided for each task. Two measures of each participant’s cognitive load were also recorded. These included the mental demands and frustration that each participant experienced during the tasks.

Once the initial planning and development stage was completed and the participants were recruited, the experiment began. The experiment consisted of each participant completing a series of four tasks in the quiet environment and five tasks
in the noisy environment. Two different sets of data were used during the test, each utilizing the same method of visualizing the data. Each participant’s task completion accuracy, time, mental demand and frustration were recorded on a per-task basis. The fifth task that was completed by each participant, after he or she finished the tasks in the noisy environment, asked him or her to describe information they could recall from the background audio. After a participant completed the tasks in each environment, the experiment was concluded.

3.2 Rationale

The rationale behind utilizing sound as a distraction in a laboratory environment (as a noisy environment) to evaluate visualized information on a hand held device was to utilize the positive aspects of each of the environments while removing the negative parts. Hand held devices are used in environments very different from laboratory environments. They are used in distracting public spaces, such as on the subway or in a mall. Usability evaluation techniques that take place in a laboratory environment do not take this into account during testing. Field testing, on the other hand, while highlighting usability issues that might not be discovered in a laboratory environment, is costly and time consuming. Utilizing sound as a distraction in a laboratory environment could potentially help discover usability issues that could only be discovered from field testing while maintaining the time constraints and costs to that of a laboratory environment evaluation.
3.2.1 Information Visualization

Evaluating an interface in repeated trials with the same participant can result in the participant learning the interface, which then leads to better performance in each successive evaluation. While this may be successful in identifying usability issues, it would hinder the identification of performance issues when evaluating a testing environment.

A solution to this issue was to use information visualization. Since the goal of information visualization is to discover information in data, learning the interface would have no effect on successive trials if the data differed in each trial. Using two distinct datasets with the same interface allowed participants to experience each environment and not affect the results of the experiment. Information visualization also added a realistic perspective to the experiment because it is used to explore data with hand held devices.

3.2.2 Evaluation Environment

Each participant took part in two usability evaluations, one in a quiet and one in a noisy environment. The quiet environment was the standard laboratory environment to test information visualization. The distraction in the laboratory environment was utilized to have full control over the distraction variable. Testing in a real public space can lead to several experimental issues such as uncontrollable levels of distraction, reproducibility and any other unforeseen issues. Simulating a single distraction allowed the elimination of external and unrecordable variables that may
have affected the results.

3.2.3 Distraction

There are many forms of distraction that may affect a participant’s performance when attempting to gain insight in data. Crowding, temperature, illumination and noise can positively or negatively affect performance. These four distraction types were examined to determine if any are feasible to use in the noisy environment. Noise is a distraction that is both easy to simulate and commonly encountered in the real-world environment. It can range from random background noise to people talking. As mentioned in the previous chapter, intelligible speech (which means non-random words or phrases) is the most significant type of auditory distraction. It was also found that social noise, or speech that a listener found interesting, also had a significant distraction effect. Taking these findings into account, a good choice for a distraction was intelligible speech that piques a participant’s interest. The audio should be at a comfortable level as to not cause the participant pain (too high) nor negate the participant’s ability to hear it (too low).

TED Talks (TED Conferences, LLC, 2011) are a good choice for noise distraction. These talks are scientific in nature, with a speaker that lecturing about his or her subject. However, in order for each participant to hold interest in the talk, it had to be a subject to which they have an interest. It was unlikely to find a talk that all of the participants commonly found interesting. To solve this problem each participant was given a list of Ted Talks to chose from and selected the TED Talk that he or she listened to during the tasks in the noisy environment.
3.2.4 Variables and Measurement

Accuracy is a variable that can be measured both quantitatively and qualitatively. For quantitative-measurable tasks, the participant was asked to find or estimate values in the data. Qualitative-measurable tasks allowed the participant to identify patterns, trends or other interesting information that he or she found in the visualization. An open-ended fifth question was asked of each participant after the noisy environment tasks to gauge how much information from the background noise the participants retained during the tasks. See section A.2 in appendix A for a list of the tasks.

Measuring time allowed for a distinction in the differences of difficulty between the quiet and noisy environments. A longer elapsed time in either of the two environments means that it took longer for the participant to find the information that he or she is looking for from the task. Time was measured in seconds on a per-task basis to allow for direct comparisons between tasks. See section 3.3.5 for more information about the application used to measure time.

Nasa’s TLX tool was used to measure the mental demands of the participants during the experiment. All six of the scales that are available with Nasa’s TLX tool were not needed for this experiment. The scales physical demand, temporal demand, own performance and effort were not used in this experiment. Instead, mental demand and frustration are the two best scales to help identify mental performance differences between the two environments.
3.3 Experiment Design

The controlled experiment consisted of four metrics, which were used to evaluate each participant’s insight: task completion accuracy, task completion time, mental demand and frustration. Each metric was recorded and calculated on a per-task basis. A list of tasks was given to each participant (see section A.2 in appendix A). Each task in the list requires that the participant find the datum, data or pattern in the interface and record the information. The time required to complete the task was recorded. The participant then filled out a form in Nasa’s TLX tool; it was used to measure the cognitive load of a participant, calculated from self-reports.

3.3.1 Experiment Details

Two visualized datasets were used in the experiment, available on the Canadian government’s website. These two datasets were: Canadian historical average temperatures by city (Environment Canada, 2012) and Canadian population growth (Statistics Canada, 2012). The dataset containing the Canadian historical average temperatures by city data will be referred to as the temperature data. The dataset containing Canadian historical population data will be referred to as the population data. The data were visualized using Google’s Chart Tools (Google, Inc., 2011). To utilize a distraction in a laboratory environment, a Podcast of a TED Talk was played, through speakers, in the background. The participant selected the talk that they were interested in listening to from a list (see section A.3 in appendix A).

Since there are two datasets to evaluate and an independent variable, which
has a boolean state (where the background audio is either on or off), the dataset and the state of the independent variable can be ordered in four possible ways. The following table describes the ordering of the environment and the datasets that each group encountered in the experiment.

<table>
<thead>
<tr>
<th>Group</th>
<th>Environment Ordering</th>
<th>Dataset ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group One</td>
<td>Quiet then Noisy</td>
<td>Temperature then Population Data</td>
</tr>
<tr>
<td>Group Two</td>
<td></td>
<td>Population then Temperature Data</td>
</tr>
<tr>
<td>Group Three</td>
<td>Noisy then Quiet</td>
<td>Temperature then Population Data</td>
</tr>
<tr>
<td>Group Four</td>
<td></td>
<td>Population then Temperature Data</td>
</tr>
</tbody>
</table>

Table 3.1: Dataset and environment ordering for each group of participants

Participants were placed into one of these groups (see table 3.1), based on their knowledge and experience, which was obtained from surveys (see section A.1 in appendix A). Preliminary surveys were completed by each of the participants to identify each person’s knowledge and experience, allowing the participants to be evenly distributed, by their skillset, between groups. Distributing participants evenly between groups, based on their knowledge and experience, allowed each of the groupings (see table 3.1) to perform equally.

The experiment was facilitated by a single person. All participant interactions, recording of variables (such as time) and the facilitation was performed by one person.
3.3.2 Application Design

The application was designed to show users either weather or population data over a timeline. The user had control over the animation, allowing them to progress through the timeline. Figure 3.1 below is a screenshot of the application using the temperature data; figure 3.2 is a screenshot of the application using the population data.

![Canadian Historical Average Temperatures by City](image)

Figure 3.1: An example screenshot of the application using the temperature data that was evaluated
Figure 3.2: An example screenshot of the application using the population data that was evaluated.

Figure 3.3 is a screenshot of the application, which was animated to the seventh time step. It shows all of the data up to and including the current position in time. As it progresses through the animation, more information becomes available to the participant. The animation allowed trends and patterns to emerge to the participant from the data, which fostered insight.
Figure 3.3: An example screenshot of the application highlighting animation

The last screenshot, figure 3.4, shows two functionalities. The first functionality being the hiding of data outside of the current timestep. This allowed the participants to see only the data that existed during the current timestep. The second functionality is highlighting data points. Selecting a datapoint allowed the participants to see the exact values of data on the graph.
3.3.3 Participant Characteristics

This experiment needed a multiple of four participants, since two environments are being compared with two different datasets. To discover around 90% of usability issues with a design, there needs to be at least 9 participants in the usability study (J. Nielsen, 2000). However, this experiment’s goal was to find more than usability issues; it compared both the performance and cognitive capabilities of participants between two evaluation environments. C. M. Nielsen et al. (2006) used 14 participants in an experiment that compared field and laboratory evaluations. A total of 16 participants were recruited, which provided enough data to compare the noisy and quiet environments. Characteristics such as technical background, education and gender were used to evenly distribute participants between the two testing
environments. The only restriction was that participants must be at least 18 years of age in order to participate in the study.

3.3.4 Participant Walkthrough

The following steps outline what each participant went through in order to finish the experiment. The time required of each participant ranged approximately between 15 and 20 minutes.

1. Each participant was asked to fill out a pre-experiment survey (see section A.1 in appendix A). It was sent to them via e-mail. The survey was used to get characteristics of each participant to evenly distribute technical knowledge between the two environments.

2. I then scheduled a location and time with each participant to conduct the experiment.

3. When a participant was met at the specified location (the usability testing lab in the Reynolds building), they were immediately issued a consent form.

4. The experiment procedure was explained to the participant.

5. The participant selected the TED Talk, which was later used as background noise during the noisy environment tasks.

6. Each participant was given a brief 2-minute introduction, explaining how to use the interface.
7. Each participant was given a list of tasks with fill-in-the-blank answers for finding information in the application.

8. The quiet or noisy environment then commenced.

9. The participant then used the hand held device to either perform the specified task or find a pattern in the data and wrote it down.

10. After each single task was completed, the participant then filled out the form in the Nasa-TLX tool. The time to complete the task was recorded. Repeat step 8 until all tasks were completed.

11. If the noisy environment was used, then a final question was asked of the participant (see section A.2 in appendix A) regarding the TED Talk that played.

12. The other environment (see step 7) began, which started the second part of the experiment.

13. Steps 6-10 were then repeated in this new environment.

14. The experiment was then concluded and the participant was thanked for their involvement.

3.3.5 Materials

This section details the materials that were used during the experiment. It consists of all of the hardware and software that either the experimenter or participant used.
Physical Materials

- A laptop was used to help facilitate the experiment. Timing and mental exertion measurement software (see section 3.3.5) were installed on the laptop. The laptop was also used to play background audio.

- Speakers were be attached to the laptop for the purposes of the playing the background audio.

- iPod Touch was the device used by each participant to view the data. It was a 32GB iPod Touch, version 5.0.1 and model MC544C.

- A paper copy of the ethics consent form.

- Paper copies of the tasks (see section A.2 in appendix A).

Digital Materials

- A timing application was developed to record each of the task times in seconds for each of the participants. Figure 3.5 is a screenshot of this application.
Figure 3.5: A screenshot of the timing application developed to record the duration of each task

- Nasa’s TLX tool was used to measure the frustration level and mental demands of the each participant.

- A Web browser, Safari, was used by each participant to view the visualization.
Chapter 4

Experiment Results

4.1 Results

The surveys distributed prior to the experiment allowed for the collection of information from the participants, which included their technical knowledge, hand held device usage, age, and level of education completed. This information allowed groups to be evenly distributed by each person’s technical abilities, age, expertise and knowledge. Balancing participant attributes between groups helped to ensure that performance between the groups was similar.

The following are the results from the experiment performed comparing a quiet and noisy environment. Sixteen individuals were recruited, thirteen of which were male and three female. Four of the individuals were between 50 and 60 years of age. The remainder of the participants were between the ages of 20 and 30. All of the participants had at least some university or college eduction. Five of the participants submitted that they had a high or very high technological comfort level while the rest were either low or medium. Twelve of the participants used a hand held device at least once per day. Four of the participants used a hand held device some of the time (less than once per day).
Table 4.1 shows the breakdown of each group by participant and each participant’s age, education level, technical knowledge and hand held device usage. Tables 4.2, 4.3, 4.4 and 4.5 show the task completion accuracy, task completion time, mental demands and the level of frustration of each participant by task.
### Participants Characteristics and Group Distribution

<table>
<thead>
<tr>
<th>Group</th>
<th>ID</th>
<th>Age</th>
<th>Education</th>
<th>Tech</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0011</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>High</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0016</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>Low</td>
<td>Sometimes</td>
</tr>
<tr>
<td></td>
<td>0022</td>
<td>20–30</td>
<td>College/University in Progress</td>
<td>Medium</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0025</td>
<td>50–60</td>
<td>Completed College/University</td>
<td>Medium</td>
<td>Sometimes</td>
</tr>
<tr>
<td>Group 2</td>
<td>0001</td>
<td>50–60</td>
<td>College/University Incomplete</td>
<td>Medium</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0017</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>High</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0020</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>Medium</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0024</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>Medium</td>
<td>Often</td>
</tr>
<tr>
<td>Group 3</td>
<td>0008</td>
<td>50–60</td>
<td>Completed College/University</td>
<td>Medium</td>
<td>Sometimes</td>
</tr>
<tr>
<td></td>
<td>0018</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>Medium</td>
<td>Sometimes</td>
</tr>
<tr>
<td></td>
<td>0023</td>
<td>20–30</td>
<td>College/University in Progress</td>
<td>High</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0027</td>
<td>20–30</td>
<td>College/University in Progress</td>
<td>High</td>
<td>Often</td>
</tr>
<tr>
<td>Group 4</td>
<td>0015</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>Medium</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0019</td>
<td>20–30</td>
<td>Completed College/University</td>
<td>High</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0021</td>
<td>20–30</td>
<td>College/University in Progress</td>
<td>Medium</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td>0026</td>
<td>50–60</td>
<td>College/University Incomplete</td>
<td>Medium</td>
<td>Often</td>
</tr>
</tbody>
</table>

Table 4.1: The characteristics of each participant and the group setup
## Participant Task Completion Accuracy

<table>
<thead>
<tr>
<th>Group</th>
<th>ID</th>
<th>Quiet Environment</th>
<th>Noisy Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Temperature Data</td>
<td>Population Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T 1   T 2 T 3 T 4</td>
<td>T 1   T 2 T 3 T 4</td>
</tr>
<tr>
<td>Group 1</td>
<td>0011</td>
<td>1 1 1 2</td>
<td>0 1 1 0</td>
</tr>
<tr>
<td></td>
<td>0016</td>
<td>3 3 1 2</td>
<td>1 2 1 1</td>
</tr>
<tr>
<td></td>
<td>0022</td>
<td>3 3 1 1</td>
<td>3 3 1 1</td>
</tr>
<tr>
<td></td>
<td>0025</td>
<td>3 0 1 0</td>
<td>1 1 1 0</td>
</tr>
<tr>
<td>Group 4</td>
<td>0015</td>
<td>3 2 1 1</td>
<td>2 2 1 0</td>
</tr>
<tr>
<td></td>
<td>0019</td>
<td>3 3 1 2</td>
<td>1 3 1 1</td>
</tr>
<tr>
<td></td>
<td>0021</td>
<td>2 0 1 2</td>
<td>0 0 1 2</td>
</tr>
<tr>
<td></td>
<td>0026</td>
<td>3 0 1 1</td>
<td>1 2 1 2</td>
</tr>
<tr>
<td>Maximum</td>
<td>0011</td>
<td>3 3 1 2</td>
<td>3 3 1 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 2</th>
<th>Noisy Environment</th>
<th>Quiet Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Temperature Data</td>
<td>Population Data</td>
</tr>
<tr>
<td></td>
<td>0001</td>
<td>3 1 1 2</td>
</tr>
<tr>
<td></td>
<td>0017</td>
<td>3 1 1 0</td>
</tr>
<tr>
<td></td>
<td>0020</td>
<td>3 1 1 0</td>
</tr>
<tr>
<td></td>
<td>0024</td>
<td>3 3 1 1</td>
</tr>
<tr>
<td>Group 3</td>
<td>0008</td>
<td>2 0 1 1</td>
</tr>
<tr>
<td></td>
<td>0018</td>
<td>1 1 1 2</td>
</tr>
<tr>
<td></td>
<td>0023</td>
<td>3 0 1 2</td>
</tr>
<tr>
<td></td>
<td>0027</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>Maximum</td>
<td>0001</td>
<td>3 3 1 2</td>
</tr>
</tbody>
</table>

Table 4.2: Participant task completion accuracy
## Participant Task Completion Time

<table>
<thead>
<tr>
<th>Group</th>
<th>ID</th>
<th>Quiet Environment</th>
<th>Noisy Environment</th>
<th>Temperature Data</th>
<th>Population Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>T 1</td>
<td>T 2</td>
<td>T 3</td>
<td>T 4</td>
</tr>
<tr>
<td>Group 1</td>
<td>0011</td>
<td>111</td>
<td>10</td>
<td>64</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>0016</td>
<td>157</td>
<td>49</td>
<td>119</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>0022</td>
<td>162</td>
<td>133</td>
<td>64</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>0025</td>
<td>63</td>
<td>162</td>
<td>85</td>
<td>108</td>
</tr>
<tr>
<td>Group 2</td>
<td>0015</td>
<td>136</td>
<td>125</td>
<td>136</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>0019</td>
<td>175</td>
<td>132</td>
<td>214</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>0021</td>
<td>165</td>
<td>132</td>
<td>147</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>0026</td>
<td>272</td>
<td>143</td>
<td>113</td>
<td>232</td>
</tr>
<tr>
<td>Group 3</td>
<td>0008</td>
<td>38</td>
<td>60</td>
<td>40</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>0018</td>
<td>80</td>
<td>76</td>
<td>92</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>0023</td>
<td>95</td>
<td>64</td>
<td>81</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>0027</td>
<td>83</td>
<td>28</td>
<td>137</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 4.3: Participant task completion completion time
### Participant Mental Demands by Task

<table>
<thead>
<tr>
<th>Group</th>
<th>ID</th>
<th>T 1</th>
<th>T 2</th>
<th>T 3</th>
<th>T 4</th>
<th>T 1</th>
<th>T 2</th>
<th>T 3</th>
<th>T 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0011</td>
<td>57</td>
<td>6</td>
<td>19</td>
<td>28</td>
<td>47</td>
<td>13</td>
<td>53</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>0016</td>
<td>19</td>
<td>4</td>
<td>39</td>
<td>20</td>
<td>22</td>
<td>16</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>0022</td>
<td>28</td>
<td>11</td>
<td>9</td>
<td>5</td>
<td>36</td>
<td>53</td>
<td>12</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>0025</td>
<td>45</td>
<td>73</td>
<td>15</td>
<td>13</td>
<td>66</td>
<td>58</td>
<td>63</td>
<td>59</td>
</tr>
<tr>
<td>Group 2</td>
<td>0015</td>
<td>93</td>
<td>72</td>
<td>20</td>
<td>42</td>
<td>49</td>
<td>19</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>0019</td>
<td>19</td>
<td>10</td>
<td>29</td>
<td>7</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>0021</td>
<td>92</td>
<td>86</td>
<td>56</td>
<td>81</td>
<td>79</td>
<td>56</td>
<td>44</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>0026</td>
<td>12</td>
<td>5</td>
<td>27</td>
<td>1</td>
<td>47</td>
<td>8</td>
<td>7</td>
<td>68</td>
</tr>
<tr>
<td>Group 3</td>
<td>0008</td>
<td>3</td>
<td>3</td>
<td>17</td>
<td>18</td>
<td>39</td>
<td>12</td>
<td>38</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>0018</td>
<td>32</td>
<td>31</td>
<td>14</td>
<td>12</td>
<td>69</td>
<td>58</td>
<td>59</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>0023</td>
<td>25</td>
<td>40</td>
<td>33</td>
<td>24</td>
<td>58</td>
<td>48</td>
<td>77</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>0027</td>
<td>29</td>
<td>21</td>
<td>23</td>
<td>16</td>
<td>62</td>
<td>38</td>
<td>51</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 4.4: Participant mental demands by task
### Participant Level of Frustration by Task

<table>
<thead>
<tr>
<th>Group</th>
<th>ID</th>
<th>T 1</th>
<th>T 2</th>
<th>T 3</th>
<th>T 4</th>
<th>T 1</th>
<th>T 2</th>
<th>T 3</th>
<th>T 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0011</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>27</td>
<td>2</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>0016</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>40</td>
<td>19</td>
<td>18</td>
<td>8</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>0022</td>
<td>41</td>
<td>12</td>
<td>13</td>
<td>5</td>
<td>40</td>
<td>45</td>
<td>12</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>0025</td>
<td>42</td>
<td>78</td>
<td>45</td>
<td>29</td>
<td>76</td>
<td>72</td>
<td>44</td>
<td>38</td>
</tr>
<tr>
<td>Group 2</td>
<td>0015</td>
<td>55</td>
<td>28</td>
<td>31</td>
<td>22</td>
<td>18</td>
<td>18</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>0019</td>
<td>1</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>0021</td>
<td>22</td>
<td>19</td>
<td>26</td>
<td>40</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0026</td>
<td>9</td>
<td>7</td>
<td>31</td>
<td>4</td>
<td>17</td>
<td>8</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>Group 3</td>
<td>0008</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>28</td>
<td>12</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>0018</td>
<td>9</td>
<td>15</td>
<td>8</td>
<td>9</td>
<td>24</td>
<td>15</td>
<td>41</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>0023</td>
<td>19</td>
<td>23</td>
<td>26</td>
<td>20</td>
<td>22</td>
<td>48</td>
<td>80</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>0027</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>75</td>
<td>24</td>
<td>23</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 4.5: Participant frustration by task
4.2 Analysis

The goal of the analysis of the experimental data was to show that if the environment used to conduct an insight analysis of a data visualization affects a participant’s insight, then testing in a quiet environment will lead to more accurate insight, less time taken to achieve insight and less cognitive load utilized than in a noisy environment. In order to verify this was the case, the accuracy of answers from the participants, time for each participant to complete his or her tasks, the mental demand of the participant during each of the tasks and the frustration each participant experienced during the tasks from the quiet and noisy environments were analyzed and compared.

Before the analysis took place the overall data was first analyzed. The data needed to be tested for normality. If non-normality exists in the data, then statistic methods that do not assume normality exists need to be used. Next an evaluation was performed to determine if the ordering of either the temperature or population data and either the quiet or noisy environment had an effect. If the ordering of the data and environment did not affect the participants’ performance then the groups could be combined. Combining the groups lead to an analysis of two groups instead of four groups.

After the preliminary analysis had taken place then the analysis was performed that compared the quiet environment and the noisy environment. The results from the tasks from the temperature and population data remained separate during this analysis since one these datasets may have been more difficult to work with than
the other. The first part of the analysis compared quiet and noisy environments. It
looked at how accuracy, time, mental demand and frustration differed between the
two environments. Finding a significant positive or negative correlation means that
the evidence supports the environment being a factor in how hand held devices are
used. Next, each participant’s performance (accuracy and time) was correlated to his
or her cognitive load (mental demand and frustration). Correlating performance and
cognitive load would provide evidence that there is a connection to mental demand
and frustration, which would have affect the outcome of a person’s ability to gain
insight and correctly navigate the data. And lastly, each participant’s awareness and
retention of the TED Talk was correlated to his or her performance during the noisy
environment tasks. It was important to find how much information each person was
able to attain from the TED Talk and how it affected his or her performance. Note
that the confidence level used throughout the analysis was 0.05.

4.2.1 Normality

Before analyzing the data, it was important to find out the type of data that
was collected. If the data is not normal, then the analysis requires statistical methods
that do not assume the data is normally distributed. Two tests were performed on
each of the variables to find evidence of normality. First, histograms were used to
help visually identify normality. While histograms do not provide definitive answers
as to the normality of data, they can give a general sense as to whether data could
be normal or not. The second test was the Anderson-Darling test for normality.

Table 4.6 contains the resulting p-values from performing the Anderson-
Darling test for normality on each of the variables from each of the environments. A resulting p-value of less than 0.05 indicates that evidence exists that the data could be normal.

### P-Values from Anderson-Darling Test for Normality

<table>
<thead>
<tr>
<th></th>
<th>Temperature Data P-Values</th>
<th>Population Data P-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.018</td>
<td>0.271</td>
</tr>
<tr>
<td>Time</td>
<td>0.050</td>
<td>0.053</td>
</tr>
<tr>
<td>Mental Demands</td>
<td>0.032</td>
<td>0.107</td>
</tr>
<tr>
<td>Frustration</td>
<td>0.059</td>
<td>0.891</td>
</tr>
</tbody>
</table>

Table 4.6: The resulting p-values from the Anderson-Darling Test for Normality

While no conclusions can be made on normality based on histograms, they do provide some sense as to whether normality could exist or not. The Anderson-Darling normality test further expands on this and provides a numeric probability of having evidence of normally distributed data. The result of these normality tests was that some data could be normally distributed while other data showed evidence that it was not normal. To remain consistent with all variables, non-parametric statistical methods were used regardless of the variable that was analyzed. Non-parametric statistical methods do not require that the data fits a normal distribution.
4.2.2 Significance of Ordering

The following analysis of the data was done to determine if the order of which the temperature and population data that was shown to the participants mattered during the experiment. If ordering had no effect on the results, then the number of groups can be reduced. Reducing the number of groups allows the analysis to be done on two groups instead of four. Groups one and four both performed the tasks from the temperature data in quiet conditions while the tasks from the population data were completed in the noisy environment. The difference between these two groups is that they each looked at the data in a different order. Groups two and three carried out the tasks from the temperature data in the noisy environment while the population data tasks were completed in the quiet environment. The order of the temperature and population data differed between these two groups.

The variable used to compare the performance of the groups was the accuracy of the answers from each of the participants. Accuracy was used because this variable was the primary measure of task performance. Accuracy, when the TED Talk was in effect, was ignored in this part of the analysis; the purpose of this section was to relate task performance between groups and the noise variable may skew the results. The first comparison was between groups one and four. Groups two and three were compared next.

A Wilcoxon rank sum test was used to compare groups. It is a test to see if one of the two samples tend to have larger values than the other sample. A resulting p-value of less than 0.05 means the two samples are non-identical and should not be
combined.

The combination of the accuracy data from groups one and four was used to calculate the Wilcoxon rank sum test, which yielded a p-value of 0.734. This means that the evidence supports the null hypothesis and the two groups can be combined.

The Wilcoxon rank sum test on the accuracy data from groups two and three resulted in a p-value of 0.169. It is statistically probable that the means of the data are equivalent and the two groups can be combined.

Since the Wilcoxon rank sum tests on the accuracies of the two groupings (groups one and four and groups two and three) resulted in the probability in having identical samples, order is not a factor in the results. The order was the only variation between the groups in each grouping. Therefore, the number of groups were reduced to two; group one was merged with group four and group two was merged with group three.

4.2.3 Quiet Environment versus the Noisy Environment

The goal of this section was to find if the results of the experiment differed between the quiet and noisy environments. If the environment was a factor in a person’s ability to attain insight then task response accuracy, task completion time, mental demand and frustration will differ, depending on the environment.

A Wilcoxon rank sum test was performed using each of the measured variables from the resulting data from the experiment. For example, the task completion accuracy from the quiet environment tasks was compared to the task completion accuracy from the noisy environment tasks. A p-value of less than 0.05 rejects the null
hypothesis that the samples are identical. Non-identical samples mean that the data from the quiet environment and noisy environment differed.

**Quiet and Noisy Environment Comparison**

<table>
<thead>
<tr>
<th></th>
<th>P-Value</th>
<th>No Noise</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Temperature Data</strong></td>
<td>Accuracy</td>
<td>9.556x10^{-7}</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>1.055x10^{-6}</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>Mental Demands</td>
<td>1.05x10^{-6}</td>
<td>19.5</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>1.055x10^{-6}</td>
<td>9</td>
</tr>
<tr>
<td><strong>Population Data</strong></td>
<td>Accuracy</td>
<td>1.014x10^{-6}</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>1.055x10^{-6}</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Mental Demands</td>
<td>1.053x10^{-6}</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>1.055x10^{-6}</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 4.7: A comparison between of medians between the quiet and noisy environments

It was found that, based on the p-values, the samples from the quiet and noisy environments were not identical. The median values show that the task completion accuracy was higher in the quiet environment. The task completion time, mental demands and frustration were all found to be higher in the noisy environment.

### 4.2.4 Performance and Cognitive Load

This section identifies if any correlation exists between each participant’s performance and his or her cognitive load. Accuracy and time are the variables
that were used to identify performance. Mental demand and frustration are the two variables that were used to measure cognitive load. A correlation between either of the two performance variables and either of the two cognitive load variables means that there is evidence that supports that a distraction in the noisy environment has an effect on how well a person is able to gain insight from data.

Spearman’s rank correlation coefficient was the method used to determine if there are any correlations. It is a non-parametric method used to determine if correlations between two of the variables exist. A strong correlation results in a rho value close to either -1 or 1; the closer the resulting value is to 0 the weaker the correlation that exists. The null hypothesis however, must also be rejected (the p-value must be less than 0.05). Finding whether a strong, weak or no correlation exists is important. A strong correlation between accuracy and mental demand, for example, would identify how an increase in mental demand affects accuracy.

### Performance and Cognitive Load Correlation

<table>
<thead>
<tr>
<th></th>
<th>Temperature Data Results</th>
<th>Population Data Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-Value</td>
<td>Correlation</td>
</tr>
<tr>
<td><strong>Accuracy and Mental Demands</strong></td>
<td>0.229</td>
<td>-0.318</td>
</tr>
<tr>
<td><strong>Accuracy and Frustration</strong></td>
<td>0.867</td>
<td>-0.045</td>
</tr>
<tr>
<td><strong>Time and Mental Demands</strong></td>
<td>0.871</td>
<td>0.044</td>
</tr>
<tr>
<td><strong>Time and Frustration</strong></td>
<td>0.173</td>
<td>0.520</td>
</tr>
</tbody>
</table>

Table 4.8: Correlation tests using Spearman’s rank correlation coefficient on Performance and Cognitive Load
Performance and cognitive load showed signs of a weak correlation. Spearman’s rank correlation coefficient indicated that every combination of performance (accuracy and time) and cognitive load (mental demand and frustration) yielded a weak relationship. However, the resulting p-values for each of the correlation tests showed that the null hypothesis was not rejected, meaning that the correlations were not significant.

4.2.5 Awareness of Environment

The purpose of this section was to analyze how each participant’s awareness of the environment affected his or her performance and cognitive load. The analysis consisted of individually correlating accuracy, time, mental demand and frustration with awareness. A correlation between any of these variables and awareness means that evidence exists that supports any person that is aware of his or her surroundings would also have his or her performance and/or cognitive load affected by it; it would affect how he or she is able to attain insight.
Background Awareness Correlation

<table>
<thead>
<tr>
<th></th>
<th>Temperature Data Results</th>
<th>Population Data Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-Value</td>
<td>Correlation</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.472</td>
<td>0.194</td>
</tr>
<tr>
<td>Time</td>
<td>0.783</td>
<td>0.075</td>
</tr>
<tr>
<td>Mental Demands</td>
<td>0.504</td>
<td>-0.180</td>
</tr>
<tr>
<td>Frustration</td>
<td>0.152</td>
<td>-0.375</td>
</tr>
</tbody>
</table>

Table 4.9: Correlation tests using Spearman’s rank correlation coefficient on Background Awareness, Performance and Cognitive Load

Attempting to correlate accuracy, time, mental demand and frustration to awareness resulted in finding a weak relationship between these variables. This means that it is unlikely that a participant’s awareness of the surrounding environment affected his or her performance and cognitive load. The null hypothesis was not rejected in any of the cases and the correlations were found to be not significant.
Chapter 5

Conclusion

5.1 Discussion

The results from the experiment suggest that using a distraction in a laboratory environment during a usability evaluation of an application on a hand held device had an impact on a participant’s performance. Participant performance was found to suffer during the noisy environment tests when compared to the quiet environment tests. Participant mental demands and frustration were also found to be higher in the noisy environment.

The experiment tested the hypothesis that if the environment used to conduct an usability evaluation of a data visualization affected a participant’s ability to gain insight, then testing in a quiet environment lead to more accurate insight, less time taken to achieve insight and less cognitive load utilized than in a noisy environment. The statistical methods performed on the data in chapter 4 provided evidence to support this hypothesis. An analysis was also done to help identify patterns and relationships between the variables.
5.1.1 Quiet and Noisy Environments

The first analysis compared the quiet and noisy environments. The accuracy of task answers was found to be higher in the quiet environment when using the temperature and population datasets. The time it required participants to complete their tasks was found to be greater in the quiet environment when using the temperature data; the task completion times when the population data was used was found to be similar between the two environments. Participants experienced higher mental demands and frustration in the noisy environment when using either the temperature or population datasets.

While the accuracy of the answers the participants recorded were higher in the quiet environment, the time it took them to arrive at these answers was the same or greater in this environment when compared to the noisy environment. One explanation for this is that the participants did not feel rushed in the quiet environment and took longer to ensure their answers were correct. Reilly and Inkpen (2007) claimed that setting had an impact on concentration. The quiet environment provided participants with a setting with no distractions and allowed them to concentrate fully on the tasks. This concentration may have allowed the passage of time to elude them. Participants, while completing tasks in the noisy environment, may have just wanted to arrive at an answer and continue because of how the distraction was affecting their focus.

Another possible reason for the higher accuracy and equal-to or greater-than time in the quiet environment as opposed to the noisy environment is that
participants lost motivation to correctly complete the tasks in the noisy environment because the social noise distraction interrupted their focus. Saraiya et al. (2005) claimed that participants need motivation in order to find insight. The higher levels of frustration or mental demands in the noisy environment could have been due to a lack of focus resulting from the distraction. Participants experiencing a lack of focus with frustration and high mental demands may have wanted to just finish a task and submit the first answer they arrived at, explaining the higher accuracy and greater time results in the quiet environment.

One explanation of the results where the accuracy was found to be higher in the quiet environment than the noisy environment is that a distraction affected the participants’ ability to understand the task or answer correctly. Bringing a participant’s cognitive load to or near capacity may prevent him or her from fully thinking about and understanding the tasks. The participants may have thought they understood what the task was asking them to find or perform, while they were being affected by the distraction, but instead misunderstood the task or arrived at a wrong conclusion.

The complexity of the data may have affected the outcome of participant performance and cognitive load. Data that is easier to work with may have shown little difference in the results between the two environments. However, complex data may also have caused participants to give up or submit incomplete or incorrect answers due to the distraction adding to the difficulty of the data and tasks. Bogges and Simon (1968), and Eschenbrenner (1971) as cited by Holahan (1987, p. 135) claimed that people are more susceptible to disruptions from noise when performing complex tasks.
than simple tasks.

The performance and cognitive load results could also be explained by social noise having an impact on performance. Brunetti (1972) as cited by Ahrentzen et al. (1982, p. 234) demonstrated that social noise affected students more than other types of noise. Social noise is a type of distraction where the listener has to build the string of words together to make sense of what is being heard. Understanding what is being said requires focus, which may have been diverted from the participants’ tasks. This was evident from the results in chapter 4, which showed participant scores of background audio awareness were high.

5.1.2 Correlation Between Performance with Cognitive Load

The next part of the analysis attempted to correlate performance to cognitive load. Identifying a correlation between cognitive load with performance could provide a way to predict cognitive load using performance. For example, if a correlation existed between accuracy and mental demand, then by increasing the amount of mental demands (caused by a distraction) a participant experiences, the resulting accuracy could be predicted. Distractions could be fine-tuned in to increase or decrease a participant’s accuracy depending on what the situation calls for.

There could be several reasons for the lack of correlation between the performance variables and cognitive load variables. It could simply mean there is no correlation. Accuracy and time do no correlate, either positively or negatively, to mental demands and frustration in any combination of these four variables.

Next, the sample size may not have been large enough. Other usability eval-
Evaluations have used a similar sampling size of participants. For example, C. M. Nielsen et al. (2006) used 14 participants in their experiment. However, this experiment’s purpose was not only a usability evaluation but an examination of the environment that usability evaluations are conducted. The lack of a large enough sample size is evident from the wide range of task completion times and task completion values. A larger number of participants could have resulted in a determinable relationship between performance and cognitive load.

Third, the way the answers were scored may not have adequately reflected participant insight into the data. Cognitive load scales were ranked between 0 and 100 while accuracy was scored with a maximum of 1, 2 or 3. Evaluating participant accuracy on a similar scale to that of the cognitive load variables could have shown a correlation between accuracy and cognitive load.

5.1.3 Correlation Between Background Awareness and Accuracy and Cognitive Load

The last part of the analysis attempted to identify if any correlation existed between each participant’s awareness of the background noise he or she heard to his or her performance or cognitive load. Correlating background awareness and performance or cognitive load could identify a trend in participants where if one participant recalled no information from the TED Talk then the resulting accuracy or mental demand, for example, could be estimated.

The first explanation for the lack of correlation between the background noise awareness to the other variables is that there is no correlation between these variables.
Background awareness may not have any relationship, either positive or negative with task response accuracy, task completion time, mental demands or frustration.

The second reason for a lack of correlation could be that the sample size of participants was not large enough to find a correlation. More participants would have provided more data points and relationships may have emerged from the data.

Another explanation is that the measure of background noise awareness did not accurately reflect how much information the participants gained from the audio. Participants may have remembered only bits and pieces from the overall recording and still scored a high awareness result. However, other participants that remembered significantly more information from the talks scored similarly as the participants who only remembered a few items.

5.2 Conclusion

This thesis aimed to show that there is a difference in the performance of users between quiet and noisy environment usability evaluations. The tasks were completed more accurately with less cognitive load in the quiet environment as opposed to the noisy environment. However, the time for the participants to complete their tasks was similar between the two environments. While task completion time showed no significance between the two environments, the essence of what affects usability in the field was successfully emulated in a quiet environment.

The insight participants attained depended on the environment being evaluated. The quiet environment provided participants with a setting that allowed them
to concentrate on the tasks and complete these tasks accurately. However, the noisy environment, which used social noise as a distraction, emulated what users could experience in the real world. User performance in the noisy environment more accurately reflected this experience.

Utilizing sound as a distraction in a laboratory environment is a viable alternative to both laboratory environment and field testing. The noisy environment successfully utilized social noise distractions to emulate a real-world scenario, which showed a decrease in participant performance and an increase in participant cognitive load. Utilizing sound as a distraction in a laboratory environment should be used in hand held usability evaluations to better identify usability issues. However, further research needs to first be performed to explore how the real-world environment affects hand held device usage and how it can be better emulated in a laboratory environment.

5.3 Future Work

One possible area that further research could extend into is interruptions. This thesis focused on how distractions, which is a constant source of sound that diverts a person’s attention away from a task, affects usability. Interruptions, however, may affect users differently. An interruption is an instantaneous distraction that diverts a user’s focus away from a task, but allows the user to regain focus until the next interruption occurs. Just like distractions, interruptions exist in the real-world hand held environment. Figuring out if and how interruptions can affect users could
lead to better design and testing of interfaces.

The work done in this thesis can be extended to compare a noisy environment to actual real-world environments. Comparing the performance and cognitive load results from two environments could identify how distractions can be fine-tuned to better emulate distractions in the field. Accurately utilizing sound as a distraction in a laboratory environment might eliminate the need to perform field testing.

Another possible area to explore is to have multiple simultaneous sources of distractions. For example, in a noisy environment, one audio source could be playing noises (such as traffic) while the other playing human speech. Multiple distractions may push a participant’s cognitive load to the limit, which could help identify usability issues when a participant has very little cognitive resources to devote to a hand held device. Multiple distractions might also affect the usability of some interface designs in different ways. Simulating multiple distractions simultaneously may reveal usability issues unique to one type of distraction.

And lastly, research could be done to further explore how other cognitive resources are influenced by distractions. A temporal measure would have allowed each participant to record how stressed for time they felt during the tasks. This measure may have identified if distractions had an effect on a person’s sense of accomplishing a task within a certain duration. Temporal stress on participants could explain anomalies in accuracy and time. If a participant is feeling rushed while a distraction is occurring, for example, they may try to finish on time to alleviate the temporal stress while not adequately finishing the task.

There is a lot of room for research with utilizing sound as a distraction
in a laboratory environment. Additional research in this area could lead to better usability practices for testing hand held devices by providing the benefits of field-testing without the associated time and costs.
References


Fuchs, G., & Schumann, H. (2006). Visualization in multimodal user interfaces of
mobile applications. *Emerging trends and challenges in information technology management*, 1, 345.


Appendix A

Participant Experiment Documentation

This appendix contains the main of the documents that each participant will use and interact with. It includes the initial survey as well as the various tasks required of the experiment.

A.1 Ethics Survey

The survey questions provided in this document will be sent to potential participants who signed up for the experiment. The questions will be on an on-line Web form. Prior to answering any of the questions on the form, a copy of the consent form (see attached or Ethics Consent.doc) will need to be accepted (clicking "I Agree") before any questions will be shown to the potential participant.

Since this is an online Web form, it is anonymous. Each potential participant’s e-mail address will be requested in order to provide future contact with him or her. Each e-mail address will be associated with a unique ID number.

Prior to participating in the experiment, some data will need to be collected. The answers that you provide will help ensure that the collected data can be correctly analyzed.

1. E-Mail Address: __________________________________________________________
2. Age: ________________________________

3. Gender: ________________________________

4. Level of Education Completed:
   - Some High School completed
   - High School Diploma
   - Some College or University
   - College or University Degree
   - Graduate Degree

5. Technology Comfort Level:
   - None (I have never used a computer)
   - Low (I occasionally use a computer to check the Web and/or E-mail)
   - Medium (I frequently use a computer or other electronic devices)
   - High (I can troubleshoot most computer related issues)
   - Very High (I can write/develop software)

6. I use a Mobile Device (such as an iPhone, for example):
   - Never
   - Sometimes
   - Often
   - Many times per day
A.2 Dataset Tasks

The following are the set of tasks that were completed by the participant.

The following tasks are for the Canadian Historical Average Temperatures by City.

- Task 1: Which city encounters the highest average mean maximum temperature? Around when does this happen? Give an estimate of this temperature.

- Task 2: Which city encounters the highest average mean minimum temperature? Around when does this happen? Give an estimate of this temperature.

- Task 3: Compare the weather data of Toronto and Ottawa throughout the entire timeline. Describe any visible patterns that you noticed in these two cities.

- Task 4: Describe, in your own words, what is happening to the average mean minimum and maximum temperatures as time increases.

- Post-Task Question: Which TED Talk Podcast did you listen to while performing the experiment? Describe what it was about.

The following tasks are for the Canadian Historical Population Data.

- Task 1: Which population growth type (immigration vs. emigration or births vs. deaths) encountered the highest loss (ignoring any gains)? When did it happen? Give an estimate to the amount of loss.

- Task 2: Which population growth type (immigration vs. emigration or births vs. deaths) encountered the highest gain (ignoring any loses)? When did it happen? Give an estimate to the amount of gain.

- Task 3: Please describe any patterns or correlations between the two population growth types, if there are any.
• Task 4: Which population growth type encounters the biggest loss (gain minus loss)? When does it occur?

• Post-Task Question: Which TED Talk Podcast did you listen to while performing the experiment? Describe what it was about.

A.3 Ted Talk Selection List

The following list of Ted Talks were used during the experiment. Each participant selected a talk. The talk that was selected played in the background during the simulated real-world environment tasks.

• Alan Russell - Regenerating our bodies

• Alex Steffen - A sustainable future

• Anthony Atala - Growing new organs

• Henry Markram - Builds a brain in a supercomputer

• Irwin Redlener - Surviving a nuclear attack

• Jane McGonigal - Gaming can make a better world

• Jim Fallon - Exploring the mind of a killer

• Jeremy Jackson - How we wrecked the ocean

• Joshua Klein - The intelligence of crows

• Kary Mullis - Next-gen cure for killer infections

• Ken Robinson - Schools kill creativity
• Mark Roth - Suspended animation is within our grasp

• Misha Glenny - Investigates global crime networks

• Patricia Burchat - Sheds light on dark matter

• Philip K. Howard - Four ways to fix a broken legal system

• Rachel Pike - The science behind a climate headline