COMPARISON OF TRADITIONAL AND ACTIVITY THEORY BASED ANALYSIS METHODS FOR VERBAL PROTOCOL DATA

by

Yogesh D Bhatkhande

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Industrial and Systems Engineering

Dr. Tonya Smith-Jackson, Chair
Dr. Brian Kleiner
Dr. Maury Nussbaum

May, 2006
Blacksburg, VA, USA

Keywords: Activity Theory, Critical Incidents, Think Aloud Method

COMPARISON OF TRADITIONAL AND ACTIVITY THEORY BASED ANALYSIS METHODS FOR VERBAL PROTOCOL DATA

by

Yogesh D Bhatkhande

ABSTRACT

The think aloud method has been used in this research to generate data that reveals the thoughts of participants of a study while they are performing tasks. The pioneers of this method, Simon and Ericsson, have provided a method to analyze the data so as to obtain meaningful results. However, this analysis method is complicated and time consuming. Most researchers use some form of categorization to perform their analysis. Critical incidents were used to categorize the data gathered in the tests conducted as part of this research. This research proposed the use of tenets of Activity Theory while performing data analysis so that the cultural and environmental aspects that influence task performance are identified and addressed as part of the analysis. A data analysis template was created that directs the analyst to follow activity theory while performing the analysis. Sample data was gathered using the Think Aloud Method. The results obtained after analyzing this data using the proposed Activity Theory Based method were compared with those obtained when the same data was analyzed using a representative traditional method of analysis. The research included positive critical incidents, negative critical incidents and level of severity of negative critical incidents as the dependent measures. No significant differences were found between the two methods based on these dependent measures. Task type had a significant effect on the number of positive and negative critical incidents identified.

ACKNOWLEDGEMENT

I would like to take this opportunity to recognize and thank all the people who were instrumental in this research. I would like to thank my advisor Dr. Tonya Smith-Jackson. This work would not have been possible without her guidance, support and confidence in my abilities. I would like to thank my committee members Dr. Maury Nussbaum and Dr. Brian Kleiner for their help and input towards improving this work. I would like to thank all my friends from Blacksburg and San Diego who helped where they could and lastly I would like to thank my family for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Background

As part of usability testing, there exists a need to understand users and get access to what goes on in their minds while using a product. One of the most widely used methods in usability engineering for elicitation of this information is the Think Aloud technique (Ericsson and Simon, 1984). However, there is often no consistency in the application of this method (explained later) by usability practitioners (Boren and Ramey, 2000). The think aloud method can have serious shortfalls as it is possible that the participant might not be telling all that is known, or might be saying more than what is known (Nisbett and Wilson, 1977). As far as analyzing the data collected is concerned, there is either insufficient or no information provided. There are also several different approaches followed. In light of these circumstances, there has been a call for alternative approaches for dealing with the think aloud method. This thesis explores one such alternative, Activity Theory (Bannon, 1997).

1.2. Problem statement

Broadly speaking, the process of gleaning an insight into a person's thought process could be divided into three parts. The first part is generating data that contains the user's thoughts. The second part is observing this data to identify where the required thoughts of interest might lie. The third part is to analyze the observations or make sense of the thoughts, in order to achieve some tangible results. Data analysis is performed to generate information based upon an informed decision of reviewing certain identifiable signs. But for the correct signs to be visible, we need the correct data

and the correct observation method. This thesis proposed an alternate unit of analysis for the observations and analysis, this unit is an activity. An activity may be defined as the process of transforming an object into a desired outcome using some process (This is explained in more detail in the sections to follow). The intent was to investigate an activity theory based analysis method for verbal protocol data while comparing effectiveness with a traditional method of analysis. The following diagram (Figure 1) depicts the same.

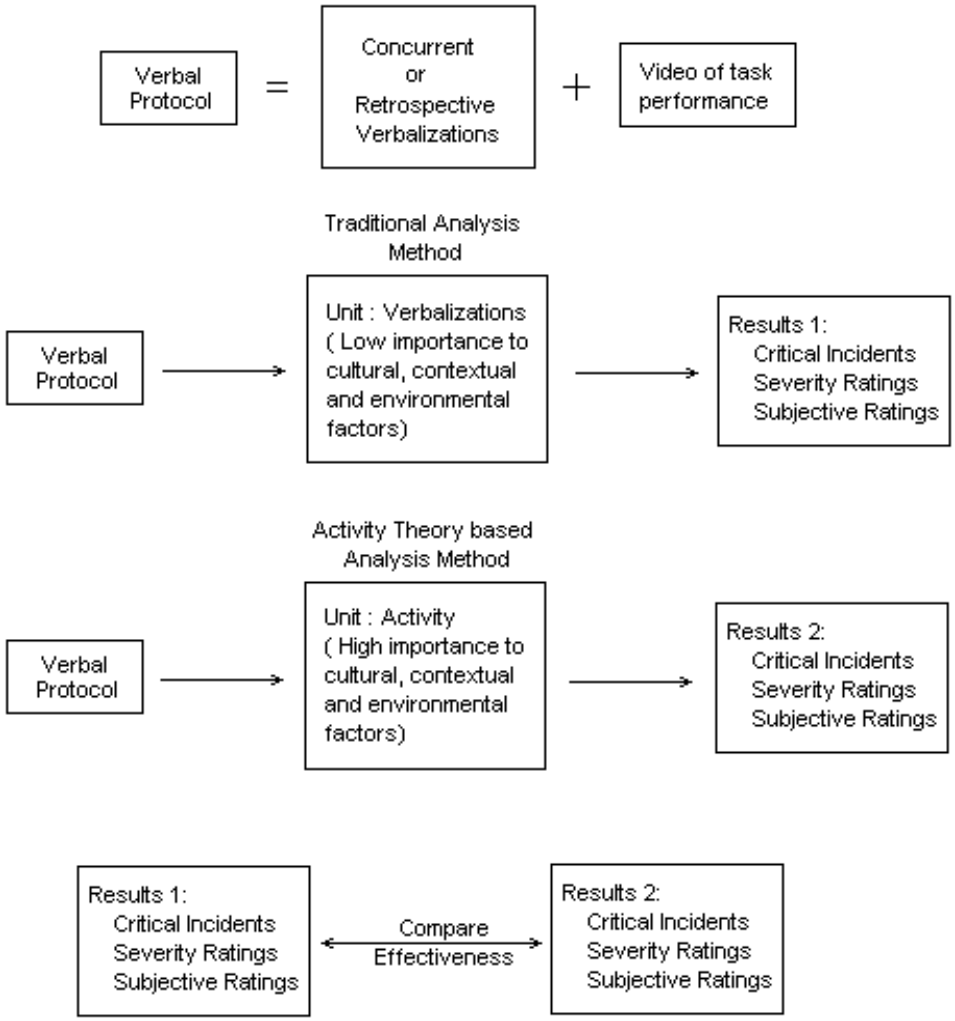


Figure 1 Problem Statement

1.3. Usability and mobile phones

The mobile phone industry is highly dependent upon designing easily usable products. A mobile phone has several attributes (small size and compact form factor, understandable user interfaces, limited interaction capabilities, etc.). The special needs of this industry require usability professionals to work differently. The test bed that was used for this study is a mobile phone user interface.

The mobile phone user interface is what the user uses to interact with the product. It includes the software that is seen on the screen (where present) as well as the hardware (various keys on the device). There is no specific set of users for this product. Young and old, educated and uneducated, able bodied and disabled people use it. As a result it is imperative that when a usability test is conducted on this device, a large amount of data be gathered quickly and efficiently without severely discomforting the participant. The think aloud method lends itself to be a natural choice. Since the majority of the gathered data is in the form of verbalizations, the participants can be asked to perform short tasks that will not consume too much of their time while yielding lots of good data regarding the usability of the product.

2. LITERATURE REVIEW

2.1. Observation

Rosson and Carrol (2002) define a usability evaluation as “any analysis or empirical study of the usability of a prototype or a system” (pp. 227). Thus, there are primarily two types of usability evaluation methods, analytical evaluation and empirical evaluation. Researchers have been trying to address the needs of evaluating the usability of products for a long time. A multitude of methods have already been developed for both analytical evaluation (claims analysis, usability inspection, user models, etc.) and empirical evaluation (controlled experiment, think-aloud experiment, field study, etc.). One form of empirical evaluation is formative evaluation. The importance of formative evaluation (or evaluation that is conducted during the design or creation of a system or product) has been well stated in literature (Scriven, 1967). Observation is a method of formative empirical evaluation.

“Observation is one of the three principal research methods – the other two being experimentation and survey” states Michael Baker (Baker, 2002 pp. 167). In fact, observation is used in almost all research in some form or another. The task of identifying the occurrence of an incident of interest to the research at hand is indeed an observation. There are basically two forms of observation, namely participant observation and direct observation (considering observations that are made first hand and not deductions made on a set of observations).

2.1.1. Types of Observation

Participant Observation usually has the observer immersed (to varying extents) within the system or context that is being observed. This process usually spans over a long period of time. As the name suggests, the observer makes observations while playing the role of a participant. Participant Observation is most commonly used in the social sciences (Spradley, 1980).

Discount Usability testing methods usually employ the Direct Observation concept. In the case of Direct Observation, the observer does not normally try to become an integral part of the system or context being observed. Thus the observer dons the role of a spectator of sorts. There are two forms of direct observation that are used in the case of Usability Testing namely, unobtrusive observation and obtrusive observation.

2.1.2 Direct Observation

2.1.2.1. Unobtrusive Observation

Unobtrusive Observation is sometimes also referred to as covert observation. This is because it might not be evident (to those being observed) that an observation is being made. While performing an unobtrusive observation, the observer avoids interacting with the person or context being observed. No questions are asked and answering questions that are asked to the observer is avoided.

2.1.2.2. Obtrusive Observation

On the other hand we have obtrusive observation (also sometimes referred to as overt observation) where the observer interacts with the person being observed. Questions related to the purpose of the observation are asked and answered. The presence of the observer is made evident to the person being observed.

2.1.2.3. Structure of Observation

Direct observation can be structured based upon the purpose of observation. In case the observation is being made to identify the occurrence of known events, a highly structured checklist can be used. However, there is an absence of structured methods for the use of observation in Usability Testing where the observer might not know what to expect or the possibilities of events occurring are very large.

Selltiz et al. (1959, pp.200) have stated 4 criteria that must be satisfied for observation to be considered a scientific technique.

> "Observation becomes a scientific technique to the extent that it (1) serves a formulated research purpose, (2) is planned systematically, (3) is recorded systematically and related to more general propositions, and (4) is subjected to checks and controls on validity and reliability."

2.1.2.4. Benefits and Shortfalls

The direct observation method can allow the observer to gather untarnished data first hand. However there are several disadvantages of this method as well. Table 1 lists some of the benefits and shortfalls of direct observation.

Table 1 Benefits and Shortfalls of Direct Observation

Benefits	Shortfalls
The expert observer can reduce the effect of distortion of facts as compared to situations where they are self-reported.	The observer has to be present at all times (to a certain extent, automatically recording events, using a video camera for example, can reduce the need for constant attendance).
Tasks that are performed very often tend to become habitual in nature (an experienced driver shifts gears in a manual transmission car as if it were a natural task), the external observer is able to identify them although the participant being observed might fail to report them.	The subject of the observation might exhibit a resistance to being observed (this could be for a multitude of reasons ranging from personal privacy to inabilities of the observer).
The observer can also pick up on occurrences that a participant might refrain from stating.	While observation-recoding methods like audio and videotaping have helped to alleviate these problems, they add the additional burden of bias. The observer now sees and hears only what the tape has recorded.
	There is also a large doubt cast upon the validity of the data gathered since the observer himself/herself can be biased while reporting data (see Hertzum and Jacobsen, 1999; Hertzum and Jacobsen, 2001 and Nielsen, 1992).

Validation of the observational data is extremely important. Without it the research community will not treat the information that is published as legitimate. As a result, the traditional form of the direct observation method is often substituted by more modern methods that invariably make use of some sort of statistical analysis techniques to provide for reliability.

2.2. Think Aloud

2.2.1. Introduction

Think aloud is a method of observation. It was originally described in work related to experimental psychology. Since introduction, the think aloud method has been used in other fields like Cognitive Psychology and Human Computer Interaction. The most commonly cited reference to Thinking-Aloud is Ericsson and Simon's (1984) work where they have described the use of the Thinking-Aloud method based upon the Information-Processing model. 'Think-Aloud' involves a person verbalizing thoughts that were generated while performing a task. It is believed that the verbalizations will allow access to the person's cognitive processes involved in performing the task, his/her perceptions of different events, even the mental models that might be relied upon.

2.2.2. Think-Aloud in use

This research primarily considered the use of the Think-Aloud (TA) method in Usability Engineering a part of the field of Human Computer Interaction. Nielsen (1993, p. 195) states, "Thinking aloud may be the single most valuable usability engineering method". The Think-Aloud technique has been used in various experiments from those related to the evaluation of searching behavior on the Internet (Waes, 1998) to using electronic ballot machines for casting votes (Bederson et al., 2003), for studying and understanding industrial inspection related tasks (Kleiner and Drury, 1998) and navigation capabilities of mobile applications (Kaikkonen and Roto, 2003). It is also heavily used in industry (Burr and Bagger, 1999; Denning et al., 1990; Lewis, 1982).

Lewis (1982) was one of the first to write about the Think-Aloud technique being applied in usability testing. He provides a good insight into the need for the kind of information that the think aloud method is unique in revealing, as mentioned in the quote below:

> "In designing the physical interface, of a computing system the problem is to match the characteristics of a keyboard or display to the characteristics of fingers and eyes. In designing the cognitive interface the problem is to fit the informational aspects of the system, including menus, messages, training manuals and much more, to the mental characteristics of users." (p. 1)

The TA technique stands out for its ability to elicit information related to what the user is thinking or providing insights into the mental characteristics of the user. The TA technique has been used as a data generation method in several cases presented in the research literature, Bowers and Snyder (1990) used TA to study window usability, Bederson et. al. (2003) used it to study usability issues with electronic voting systems, to name a few. Several other instances of the use of the TA method have been mentioned throughout this literature review. There are also several works that implore on the validity of the method, the way it is applied, its effectiveness as a usability testing tool (Boren and Ramey, 2000; Deffner, 1990; Nielsen et. al., 2002 to name a few). Nisbett and Wilson (1977) provide a psychological perspective on the concept of verbalizing thoughts where they question the truth behind what is being told by the participant.

2.2.3. Benefits and Shortfalls

Literature shows that TA is one of the most widely used methods for this purpose. As compared to other usability testing methods there are several benefits to using TA. These along with the shortfalls of the technique are listed in Table 2 (below).

Table 2 Benefits and Shortfalls of the Think Aloud method

Benefits	Shortfalls
Since a participant is actually speaking out aloud, the method allows the analyst to easily locate a problem.	Having to speak at all times in an uncommon surrounding in the presence of unknown people (viz. the experimenters) is awkward (however, Lewis, 1982, reports that the participants do get used to it).
The data collected can provide an idea of why a particular issue is troublesome to a participant.	It is possible that the added burden of TA might actually affect the cognitive process itself (Preece, 1994).
As compared to other observation methods, interviews and questionnaires, the TA technique captures a problem when it happens (the other methods might require the researcher to repeatedly evaluate a data log, or build context for the participant to aid memory).	Due to the dependence on observer intrusion, one experimenter cannot run multiple participants simultaneously (the constructive interaction technique, explained later, could be considered an exception).
Participants do not usually distinguish between the levels of importance of a problem; they will verbalize a minor problem if they can recognize its existence.	A TA generates large volumes of data, analyzing this is very time consuming. There is also no definite method that is followed for the purpose of analysis.
A researcher can understand how a participant feels about the system or product being tested (if they like or dislike it) from what is being said.	Not all that is said by a participant may be accurate (Participants might not be able to say all that they want to due to lack of time and later forget, or that they might say something to satisfy their own intentions, etc.).
If time is not an important factor (the speed of performing a task is not considered), then low fidelity prototypes can be used to perform TA at very early stages of design.	Literature seems to indicate that time should not be used as a metric when concurrent TA (explained later) is used since it slows down task completion. However, there are reports that both support (Rhenius and Deffner, 1990) and refute (Bowers and Snyder, 1990) this issue.

Research (Nielsen, 1994; Virzi, 1990) also suggests that the technique is beneficial (considering the quality and quantity of the issues revealed) when performed using fewer participants (as low as 3 – 6 participants).	There is also some evidence of method bias resulting in an improvement of participant performance when verbalizing concurrently (Wright and Converse, 1992). Rhenius and Deffner (1990) have concluded however that as far as a final solution is considered, thinking-aloud was very similar to a silent group.
	It is also believed that the method cannot be used effectively if a speech-oriented task is being performed (Karsenty, 2001, used retrospective TA in the evaluation of a voice based directory system).

2.2.4. Types of Think Aloud

Verbalizations can either be concurrent or retrospective (there is also another form of TA that is in popular use it is called constructive interaction). Each method has its specialties and each provides a way to work around the problems that might exist in the others.

2.2.4.1. Concurrent TA

Concurrent TA is considered traditional TA (emphasized by Ericsson and Simon, 1984). Concurrent TA requires a participant to continuously vocalize what he/she is thinking while simultaneously performing a task. Participants are usually unclear about what they should do when they are asked to think aloud. The act of speaking continuously is also difficult to maintain, resulting in participants often falling silent, which in turn causes intrusive questioning or prodding on the part of the examiner to get them talking again. Kurniwan et al (2003) used the concurrent think aloud method in evaluating joystick operated full screen magnifiers for visually impaired users, Kaikkonen and Roto (2003) used concurrent think aloud to study the usability of navigating XHTML websites using a mobile phone.

2.2.4.2. Retrospective TA

In the case of retrospective TA, participants are allowed to perform the task first. Their task performance is often recorded (usually on video). Once task performance is completed, they are asked to think aloud. Most commonly, they are shown the recordings of their task performance while thinking aloud so that they may find it easier to remember what had happened during task performance. A study by Bowers and Snyder (1990) involved comparing the results of a concurrent and retrospective TA. They reported that there was a significant difference in the quality of data collected when the different methods were used. The data collected from the retrospective TA was richer in content (the verbalizations were more explanatory in nature) as compared to the concurrent TA. Participants also seem to readily recollect task related information like what they were supposed to do, what their strategies for doing them were, where they faced problems, etc. Nielsen and Christiansen (2000) performed a different kind of retrospective TA where their participants were only shown portions of a video collected from an event that did not actually have the participants performing any specific experimenter designed tasks, the participants were able to remember what had happened in great detail.

2.2.4.3. Constructive interaction

This technique is also referred to as the “co-discovery” technique (Kennedy, 1989) or “paired-user testing” or “co-participation”. The specialty of the constructive interaction lies in the fact that two participants are allowed to work together on a common task. The idea is that a natural dialog would ensue between them (for reasons of explanation, argument, etc. about why a particular task was performed or how it

should have been performed, etc.). The method helps to reduce the unnatural feeling brought about by having to think aloud (participants find it more acceptable to speak to a colleague rather than to themselves or to an inanimate object like a camera or tape recorder). Thus constructive interaction provides information about how people would work together to solve problems or complete tasks. The coaching method, where the test participant is allowed to ask questions to an expert coach, could also be considered as a variant of the constructive interaction method. Denning et. al. (1990) mention the use of the co-discovery for usability testing in their case study. Kennedy (1989) describes the use of the co-discovery method to test the usability of telephones and other communications products from Northern Telecom.

2.2.5. Method of performance

Having elaborated on the different kinds of TA, it should be noted that the greatest problem that TA has as a method is that although it is popular, there is no consistency in the way in which it is applied (Boren and Ramey, 2000, Nielsen et al., 2002). Usability professionals are known to change the method of application; often very less information is provided about the method of analyzing the data. Books on Usability Engineering and Human Computer Interaction do mention the use of TA as a popular Usability technique (Nielsen, 1993) but very little is said about the right way of doing it. Authors seem to implore about their application of the technique rather than building a common path for working TA that others can follow (Dumas and Redish, 1993; Preece, 1994; Rubin, 1994). An analysis of the literature indicated that it is important to prepare the participant for TA. Table 3 lists out basic steps that can be followed to provide instruction to a potential participant of TA.

Table 3 Things to do Before Letting the Participant Think Aloud

Order of event	Step
1	Instructions with example
2	Warm-ups with example
3	Description of Prompt (to make the participant start talking again)
4	Determination of when to prompt and why
5	Debriefing after the trial

The steps involve explaining to the participant what the experimenter expects from them, and showing by example. A description of the application method (or how the participant would perform the TA; commonly, authors merely state that a TA protocol was followed) should be provided along with an explanation of the method of analysis (commonly, only the results are provided).

As far as reporting of the method is concerned, particular attention has to be given to mentioning the specifics of the different parts. These pieces of information would play the role of stepping-stones towards reducing inconsistency in method application.

2.2.5.1. Instruction

Many papers that mention the use of the TA method do not provide detailed information regarding the provision of instruction to the participants on what TA is and how it has to be performed (Mehlenbacher, 1993, Rowley, 1994, Waes, 1998). In the case of others, there is a mention of instruction being provided but not of what the instructions actually were (Rhenius and Deffner, 1990). Below are quotations from a few of the works that do list instructions

Lewis, C. (1982, pp. 4)

> "Tell me what you are thinking about as you work"…."we are not interested in their secret thoughts, but only what they are thinking about the task under study. We make clear that we have no stake in the system they are using, so they need not be afraid of hurting our feelings if they make a criticism. We stress that it is the system that we are evaluating, and not they themselves. This is extremely important since we have found participants tend to blame themselves for any problems they encounter. This creates bad morale, and can interfere with the study as well as being unpleasant for the participant."

Ericsson and Simon (1984) have elaborated on the importance of the kinds of instructions that are provided to the participant by different experimenters (pp. 80 – 82) and have also provided an example of what they themselves do when asking participants to "Talk Aloud" and "Think Aloud" (pp.376, pp.378; quotations below).

> "In this experiment, we are interested in what you say to yourself as you perform some tasks that we give you. In order to do this, we will ask you to TALK ALOUD as you work on the problems. What I mean by talk aloud is that I want you to say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time, I will remind you to keep talking aloud. Do you understand what I want you to do?"

> "In this experiment, we are interested in what you think about when you find answers to some questions that I am going to ask you to answer. In order to do this I am going to ask you to THINK ALOUD as you work on the problem given. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you have given your final answer to the question. I don't want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will ask you to talk. Do you understand what I want you to do?"

2.2.5.2. Practice or Warm-ups

Since the concept of TA is very different, the participants of a test might find it easier to understand what is expected of them if the "experimenter demonstrates the method to them" (Ebling and John, 2000 pp. 290). It might also be useful to have the participants have a few trial runs of the method. Participants get accustomed to thinking aloud after a while as the awkwardness of the method wears off.

2.2.5.3. Prodding

Participants often stop verbalizing while performing a task. This might be due to several reasons ranging from getting engrossed in the task, by being overwhelmed by the demands placed upon them, to simply not having anything to say. The common tactic in such cases is for the experimenter to remind the participant to speak. Non-directive statements ("Please Keep Talking", Ebling and John, 2000 pp. 290; "Keep Talking", "What are you thinking about?" Ericsson and Simon, 1984 pp. 83; "What are you thinking?", "What is that telling you?", Lewis, 1982 pp. 4) are usually effective to restart verbalizations. However, in practice, experimenters often ask probing questions to get the participant to speak of the issues that are of interest to the examiner. Often experimenters take this prodding too far leading to a participant getting annoyed. Boren and Ramey (2000) point out that there is again no consistency in the reason for reminding or method used. They suggest that the reminders should come after a predetermined time (15 second to 1 minute pauses, Ericsson and Simon, 1984, pp. 83) has passed without the participant verbalizing and the reminder should be short, non-directive (also see Rhenius and Deffner, 1990) and neutral in nature.

2.2.6. Data

2.2.6.1. Data Collection

The data produced by the think aloud method, obviously, is in the form of spoken discourse. Due to this form, generally qualitative data is produced. During task performance, at least an audio recorder is used to record the verbalizations so that they can be reviewed later. Often video is used to record the visual on goings along with the verbalizations, this richer medium (Daft and Lengel, 1986) allows the experimenter to build context while analyzing the data later (this can be especially useful when the person analyzing the data is not present at the time of experimentation). Of the literature that was reviewed, “audio only” as the format of recording was very rarely used (see Karsenty, 2001). Most often the favored method of data collection and recording was videotaping that captures both audio and video (see Bederson et al., 2003; Denning et al., 1990; Ebling and John, 2000; Kaikkonen and Roto, 2003; Koenemann-Belliveau et al., 1994; Nielsen and Christiansen, 2000 and Bowers, 1990 for examples). In certain cases, the experimenters have used a different form of visual data capture to either make the participant feel more comfortable (Rowley, 1994) or to help the participant remember (Triggs et al., 1990) or to collect additional form of data more efficiently (Waes, 1998, used an online camcorder to collect data about searching behavior).

2.2.6.2. Data Analysis

Experimenters have also followed several different styles of analyzing the data that is collected. Very often, notes are made while the actual experiment is being performed and these notes are heavily consulted during analysis (see Denning et al., 1990; Koenemann-Belliveau et al., 1994; Rowley, 1994 for examples). Notes have been used to index events with the time of occurrence of the recording on the media used (audio or video tape) or as a quick reference for the most important happenings. Some experimenters rely on transcripts (Ebling and John, 2000; Karsenty, 2001) of the verbalizations (either as a record or for use in a method of analysis like Conversation Analysis, Markee, 2000; Monk and Gilbert, 1995, etc.). Transcription, however, is a very tedious procedure (Denning et al., 1990). Due to this reason, several researchers follow different policies where in they code or categorize information based upon their own experience (this thesis followed a similar approach using the Critical Incident method, explained later) in some cases an explanation is provided for the manner of categorization (Bowers and Snyder, 1990; Ebling and John, 2000; Triggs et al., 1990). This categorization also allows quantitative analysis of the data in certain cases.

2.2.7. Alternative Approach

As stated earlier, the aim of this thesis was to propose an alternative approach to the traditional methods of analyzing verbal protocol. The intent was to create a structured data-reporting template. Experimenters do not always explain how the verbalized data is analyzed. In certain cases, experimenters have categorized the spoken statements based upon the type of content (Bowers and Snyder, 1990); interface components (Ebling and John, 2000). While these classifications indicate the

relevance of the statements made to the “type of comment” and “specific location”, they might not be very indicative of what the participant’s goals might have been. This thesis developed a data analysis method oriented towards identifying the goals of the participant from the activities performed and verbalizations made based on Activity Theory (explained below). Evaluators used the Critical Incident method (explained later) to identify and log issues.

2.3. Activity theory

The concept of Activity theory originated in Russia. The implications of Activity Theory for Human Computer Interaction are summed up well by Kaptelinin (1996) in a set of principles shown in Table 4. It is not a predictive theory but more like a conceptual system.

Table 4 Activity Theory Principles (Kaptelinin, 1996)

Principle	Explanation
Unity of consciousness and activity	The human self image and interactions with the environment are to be treated as one.
Object-orientedness	Environmental properties (social and cultural properties) coexist with other properties of objects (physical properties, etc.).
Hierarchical structure of activity	Differentiation between the levels of process operation with significance to the intended objects.
Internalization – Externalization	Using the imagination (for example) to simulate an activity and carrying it out in the real world.
Mediation	Tools mediate activity. Differences in the types of tools exist due to socio-cultural differences.
Development	Understanding the change of a complex system to its current form.

In the case of Activity Theory, the basic unit is an activity (Kuutti, 1991). This activity is performed in context to the overall task. An activity has several properties. There is an actor or a subject who performs the activity. The subject knows the purpose for performing the activity. This purpose component is called an object. The presence of the activity is to get a desired outcome. Engestrom's (1987) structural model for activity provides a clear picture. Figure 2 shows the simple model and Figure 3 shows the basic structure of an activity in the context of a mobile phone related task.

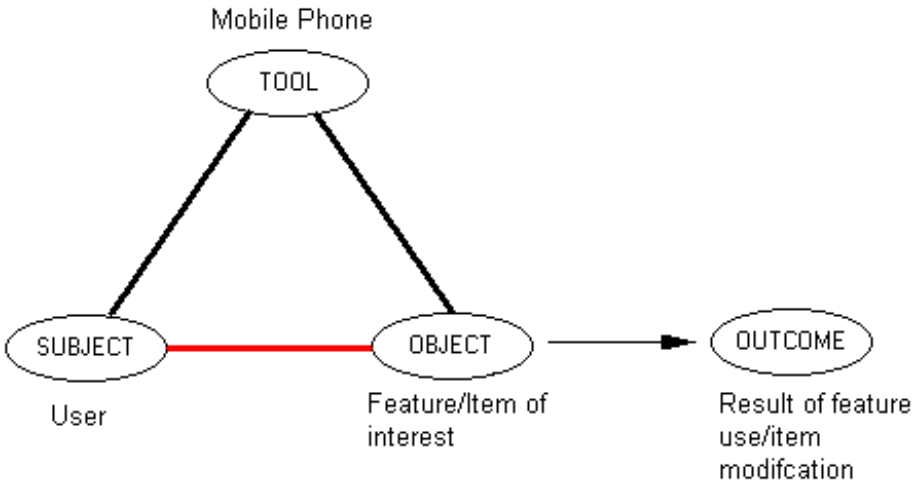


Figure 2 Simple Engestrom Model (adapted from Engestrom, 1987)

The simple model involves a mediating item between the subject and the object that comprises of the various tools that might be used. For example, a person (subject) uses a mobile phone (tool) to save phone numbers (object). The activity is always performed within the community and hence the object of the activity is shared with the community. The relationship between the subject and the community is mediated by the rules component and the division of labor (DOL) component mediates the relationship between the community and the object. Following the example mentioned above, the person will have to identify the correct location for saving the information based upon

the software design, the software design will also attempt to work as per existing metaphors within the community (rules). The phone designers (involved community) will attempt to design such that the person does not have to remember the phone number while finding the location to save (division of labor).

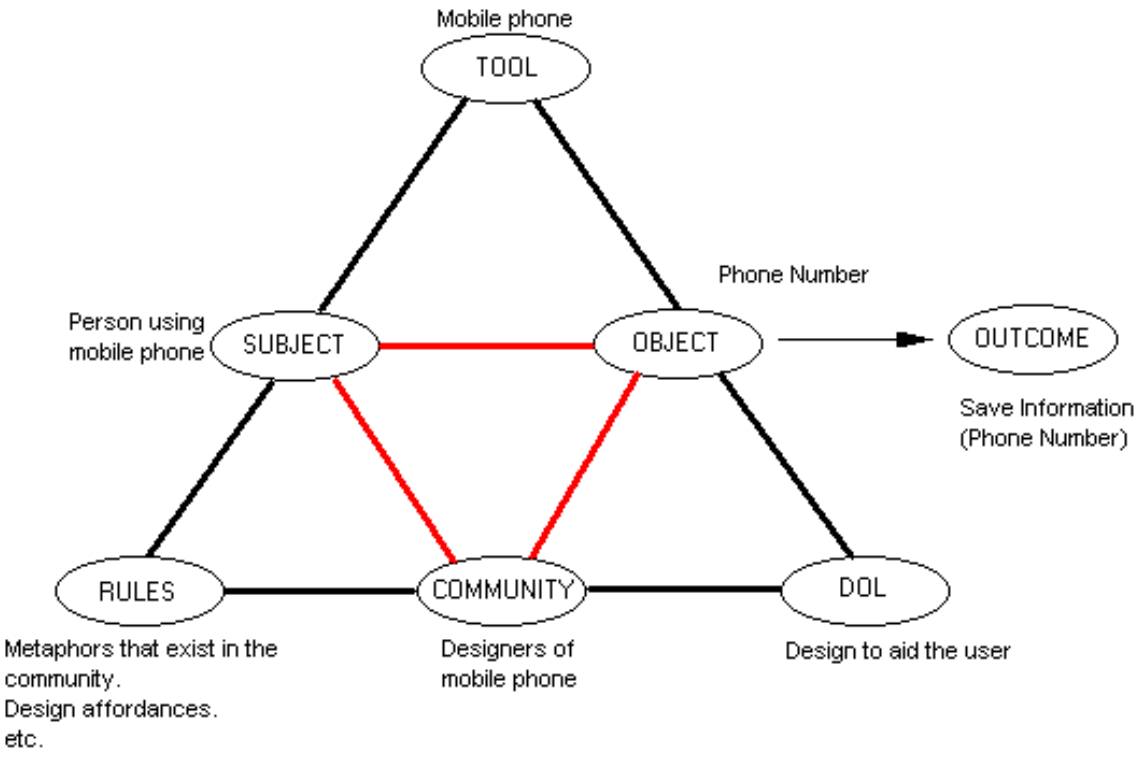


Figure 3 Basic Structure of an Activity (adapted from Engestrom, 1987)

Kuutti and Arvonen (1992) have used activity theory to identify potential Computer Supported Collaborative Work applications. The authors have indicated that the activity theory unit of analysis (activity), is more manageable and meaningful for identifying individual actions as compared arbitrarily selected concepts. Kaptelinin, Nardi and Macaulay (1999) developed an activity checklist based on the concept of activity theory that would allow clarification and elucidation of contextual factors to aid in the design of HCI based systems. Honold (2000) used activity theory to understand the use of a product. The investigators were able to identify the effect of the existing culture

and traditions (Indian households finish washing by 1:00pm) by conducting the interviews in the homes of the users (typically housewives). They were able to identify shortcomings in the product design with respect to the environmental situation that exists (higher frequency of use and in mornings).

2.4. Critical Incidents Technique

The Critical Incidents technique (CIT) developed by Flanagan (1954) has been widely used by researchers. Flanagan defines Critical Incidents as follows:

> "By an incident is meant any observable human activity that is sufficiently complete in itself to permit inferences and predictions to be made about the person performing the act. To be critical, an incident must occur in a situation where the purpose or intent of the act seems fairly clear to the observer and where its consequences are sufficiently definite to leave little doubt concerning its effects." (p. 327)

From the perspective of activity theory, critical incidents can be considered as events significant towards achieving the purpose of an activity. Observers (subject matter experts) who are knowledgeable in the field of the ongoing activity identify these critical incidents. The critical incident itself can be both supportive as well as disruptive towards the activity and can thus be classified as a positive or a negative critical incident (respectively). The word "critical" does not intend to indicate an emergency of any sort (though in certain medical cases this is the intended meaning). It merely points out that the incident has a serious effect on the outcome of the ongoing activity.

There is a vast availability of literature on the many instances of the use of the CIT and work continues on improving the method. Castillo, Hartson and Hix (1998) studied the use of the user reported critical incident technique where participants identify and note critical incidents on their own. Starr-Schneidkraut, Cooper and Wilson used the CIT to evaluate the impact of MEDLINE (1989). Carrol et al. (1993) have

elaborated on the use of critical threads to capture information when an incident occurs at a point that is remote from its context of use during activity performance.

The CIT finds special use in the field of usability evaluation. It has several advantages over other methods. The CIT does not require immediate observation, but provides information about users performing tasks in real environments. Thus, tasks can be recorded while participants perform them and reviewed by experts at some future time. This saves precious resources of time and money. Combined with the think aloud method, the critical incident technique can be used to provide an insight into user thought processes as well. The most serious disadvantage of the CIT is the reliance on subject matter experts for identifying the existence or occurrence of the incident. The CIT is very susceptible to what is called the “Evaluator Effect”. Hertzum and Jacobsen (2001) define the Evaluator effect as “the differences in evaluators’ problem detection and severity ratings”. Nielsen (1992) found the existence of this effect when novice, expert and double expert observers were used for heuristic evaluations.

2.5. Summary

The Think Aloud method is a versatile technique that can be used to generate useful and insightful information in the actual words of experimental participants. However, there are several shortfalls that are associated with it. This research has provided an alternative solution to one of those shortfalls. A way to analyze the vast quantities of data collected using a non-traditional approach based on Activity Theory has been proposed and compared with a traditional approach. The literature review also yielded information that will help other practitioners to apply the Think Aloud technique.

3. METHODOLOGY

This chapter outlines the details of the research that was conducted. There were two parts to the research. In the first, verbal protocols were elicited and recorded. Participants were asked to complete tasks using a proprietary user interface. Verbal protocols generated as part of these tasks were recorded using a video camera. In the second part, this verbal protocol was analyzed using data analyses templates created to emulate the proposed Activity Theory method and a traditional approach. The results of these methods of analysis were compared in a 2x2 factor between subjects design experiment which is explained later.

3.1. Research Purpose

The purpose of this research was to create an activity theory based structural model for analyzing verbal protocol data. A template for logging data and analyzing the data using activity theory was developed. The intention of the template was to direct the attention of the analyst to the perceived goals (identified from the activity being performed) of the person in performing a task and how the test system helped or hindered their progress in achieving these goals. The identification of issues was based upon the critical incident technique. Usability practitioners can use the template as a tool for data gathering/analysis purposes. The research also compared the proposed approach with a traditional method of analyzing data generated from a think aloud session.

3.2. Data Generation

3.2.1. Purpose

The first part of the research involved gathering data for the purpose of analysis in the actual experiment. In order to do this, a usability test was conducted. The focus of the test was the software user interface of a mobile phone. Only the verbal protocol was of importance to this research. The tasks that were performed by the users in the usability study (the selection of these users was decided by the mobile phone manufacturing company based upon their requirements). The users were asked to think aloud to generate the verbal protocols. Each user performed either a concurrent or retrospective think aloud.

3.2.2. Users participating in usability study

Eight employees (4 male and 4 female) of the company that manufactured the product participated in the usability test. The mean age of these users was 34.5 years with a standard deviation of 6.23. The users satisfied the selection criteria of not having any experience with the think aloud method and owning and using a mobile phone for at least one year. The company selected current employees to be users as there was concern regarding the confidentiality of the product which was not publicly released at the time of the study. Due to the fact that all users worked for the company, it is possible that they had at least seen or heard of the product. Care was taken to recruit users who would have had only minimal exposure to the product.

3.2.3. Equipment

3.2.3.1. Video camera

Data were recorded using a video camera. The video image captured the LCD display screen of the mobile phone. The video camera also recorded conversations or verbalizations made by the users for both forms of verbal protocol collected. In the case of Retrospective Think Aloud, the camera captured the images (of the mobile phone LCD screen) off a TV monitor.

3.2.3.2. Mobile phone

The participants performed tasks using a Kyocera, KX2 mobile phone. The phone sported a swivel form factor and had an inbuilt camera. This was the company's first phone with a camera feature and the company was interested in identifying usability issues that might exist with the user interface of this feature in particular.

3.2.3.3. Television and VCR

As part of the retrospective think aloud, the task performance images captured were displayed (with sound muted) to the user using a television and VCR. The user was asked to think aloud while viewing a replay of their task performance; this was done in order to aid the user to recollect their thoughts.

3.2.4. Tasks

The users were asked to perform four tasks that involved use of the camera feature or associated applications. The tasks were selected by the company that manufactured the product so that they could obtain data that was of use to them for studying the usability of the product. The users were familiarized with the think aloud method before task performance. As mentioned in the literature review, a little practice makes participants feel more comfortable about thinking aloud. The familiarization

process was completed in about 10 – 15 minutes (more details of the familiarization are provided in the discussion section). The users were then asked to perform the tasks (using the mobile phone) in a randomized order. The tasks were scenario based and replicated normal use situations. Appendix A lists out the four tasks and appendix B contains a sample sheet that was provided to the users instructing them of the task they had to perform.

3.2.5. Order of presentation

Each user performed four tasks (Appendix A). Four users were asked to provide a concurrent verbal protocol and the other four were asked to provide a retrospective verbal protocol. The orders of presentation were randomly selected from the table below and each user was randomly assigned to one of them (each order was used once for each type of protocol).

Table 5 Order of Presentation

1st	2nd	3rd	4th
A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A, B, C and D →tasks to be performed

3.2.6. Procedure

The users completed a demographics sheet (Appendix D) and were briefed about the purpose of the study (in this case a two fold purpose existed, one was the actual usability test required by the manufacturer and the second one was that of this research). The users were then familiarized with thinking aloud. Ericsson and Simon's commonly followed guidelines for application of the think aloud method were followed

for this process. Steps included providing instructions along with an example, participant warm up along with an example, description of prompts to continue thinking aloud and an indication of when the experimenter will prompt and participant debrief after the trial.

After the familiarization process was completed, the actual experiment was conducted where users were asked to perform tasks while thinking aloud concurrently or retrospectively according to one of the order of presentations mentioned in Table 6. A video camera captured activities performed by the user (seen on the mobile phone LCD screen and keypad) and the verbalizations made. In the case of the users performing the retrospective TA, visuals captured during silent task performance were shown to them immediately after performance with audio muted. This aided them to remember their thoughts during the actual task performance.

3.3. Experiment

3.3.1. Purpose

The experiment compared the effectiveness of the proposed activity theory based data analysis method to a representative traditional method. As stated earlier several different methods of analysis of Think Aloud data can be found in literature, there is no single structured method available that can be said to be the traditional method. A representative method was used for the purpose of this study (described later). Similar to several studies that were conducted, transcripts of the verbal protocol were not created, a method of categorization was used instead (Bowers and Snyder, 1990; Ebling and John, 2000; Triggs et al., 1990 to name a few).

3.3.2. Research Hypotheses

The null hypotheses were as follows:

The method of analysis and type of verbal protocol will not have an effect on the number of positive critical incidents identified.

The method of analysis and type of verbal protocol will not have an effect on the number of negative critical incidents identified.

The method of analysis and type of verbal protocol will not have an effect on the number of negative critical incidents of the same severity ratings.

The method of analysis and type of verbal protocol will not have an effect on the subjective ratings for any of the performance measures.

3.3.3. Experiment Design

A fixed 2x2 between factors design was used as the experimental design. The two between-subject factors were “Analysis Method” and “Verbal Protocol”. The two levels of “Analysis Method” were Activity Theory Based and Traditional (described later). The two levels of “Verbal Protocol” were Concurrent and Retrospective. The analysts were treated as random-effects variables. Table 6 shows the overview of the experimental conditions and factor levels.

Table 6 Overview of Experimental Conditions and Factor Levels

		Analysis Method	
		Activity Theory Based	**Traditional**
Verbal Protocol	Concurrent	C1 (S_1 and S_2)	C2 (S_3 and S_4)
	Retrospective	C3 (S_5 and S_6)	C4 (S_7 and S_8)

C1 - C4 → Possible treatment conditions Analysts → S_1 - S_8

3.3.4. Dependent Measures

The dependent measures gathered during this experiment were number of critical incidents (positive and negative), the number of negative critical incidents of each type of severity and subjective ratings provided by the participants at the end of the task.

3.3.4.1. Positive Critical Incidents

A critical incident is classified as a positive critical incident if it is supportive towards the successful completion of the task being performed. The sum total of such critical incidents was considered. The participants, based upon their expert judgment, classified an incident as a positive critical incident.

3.3.4.2. Negative Critical Incidents

A critical incident is classified as a negative critical incident if it is disruptive towards the successful completion of the task being performed. The sum total of such critical incidents was considered. The participants, based upon their expert judgment, classified an incident as a negative critical incident.

3.3.4.3. Number of problems of each type of severity found

The need for a severity rating of the usability problems that are identified in a study is obvious when the time comes to resolve these issues. Based upon the level of impact that a particular issue has on the usability of a system, one can decide which problems have to be solved and then which of these problems have to be solved before others. Effectively, the identified problems are first rated with respect to each other. The method for rating that was followed is similar to the one described by Nielsen and Mack (1994).

For each individual Critical Incident that was identified, the participant provided a rating based upon three attributes. The attributes are listed and described in Table 7 below.

Table 7 Attributes Considered for Severity Rating (Nielsen and Mack, 1994)

Attribute	Explanation
Frequency	The frequency of occurrence. Is it common or rare?
Impact	The effect of problem on task performance.
Persistence	Will users be able to overcome this problem with instruction, or will it trouble users repeatedly?

The ranking involved the participant identifying the level of importance of each Critical Incident from the perspective of usability. The ranking scale is provided below in Table 8.

Table 8 Explanation of Ratings (Nielsen & Mack, 1994)

Rank Level	Name	Explanation
0	Not a usability problem	The problem might be a coding related error.
1	Cosmetic	Problem is very insignificant and need not be fixed unless time is available on a project.
2	Minor	This problem exists but has low priority in getting fixed
3	Major	It is important to fix this problem, it is given high priority
4	Catastrophe	This problem has to be fixed immediately.

The total number of problems that lie in each individual severity rank level or category served as a dependent measure.

3.3.4.4. Subjective Ratings

After the participants performed the entire analysis, they were asked to provide subjective ratings for three performance measures (with respect to a usability evaluation method), viz. Thoroughness, Validity and Downstream Utility (Hartson, Andre & Williges, 2003). These subjective ratings provided information about what novice usability practitioners felt about the method of analysis that they used. They are explained in Table 9 below:

Table 9 Explanation of subjective rating performance measures

Rating Type	Explanation
a. Thoroughness	The level of completeness. In this case, the participant's opinion of how complete an analysis was performed based upon the method that was used.
b. Validity	The level of correctness of results. In this case, a subjective rating of the method in identifying the right results as compared to false alarms.
c. Downstream Utility	The participants were explained that they should keep the perspective of the analysis in an industrial setting. In this case a subjective rating of the usefulness of the data generated and the method itself.

3.3.4.5. Qualitative Data

In addition to the quantitative data mentioned above, the participants were interviewed after the task performance was completed. They were all asked the same series of questions. The interview questions are listed in Appendix G. The purpose of the interview was to understand what the participants felt were the merits or demerits of the analysis method that they used and to understand their overall experience with the method.

3.3.5. Analysts

Eight Virginia Polytechnic Institute and State University (Virginia Tech) graduate students agreed to participate in the study as analysts of the collected data. Two of these students later opted out of participation. As a result, two additional analysts were recruited. They were graduate students from the University of Southern California. Discounting those that withdrew, there were 6 male and 2 female participants. All analysts were either from the Human Factors or Computer Science disciplines. Their mean age was 27.87 years with a standard deviation of 3.04. The analysts were considered experts in the field of usability. The proposed analysis method was designed for the novice to intermediate human factors practitioner and as a result, the recruited students were beginners.

3.3.6. Equipment

3.3.6.1. Videos

Data reduction was performed on the data gathered from the usability tests. This was done to randomize the task performance order when viewed by the analysts. The concurrent and retrospective think aloud were kept on separate tapes and two sets of tapes were created for each such that the task performance on one was in reverse order to the other. This was done to reduce the effect of bias.

3.3.6.2. Analysis Templates

Two analysis templates were created (see Appendix F). One was structured so that the analyst would follow the traditional approach to analyzing the data with verbalizations as the unit of analysis and the other modeled an activity theory based approach for analysis with the unit of analysis being an activity (the template is described in the procedure section below).

The templates would allow the analysts to direct their analysis such that meaningful recommendations for design changes with intent to aid task performance would result. The templates also ensured that the analysts followed the method of analysis that they were assigned.

3.3.7. Procedure

The analysts were briefed on the purpose of the study (to compare the effectiveness of the activity theory based method of analyzing data with a traditional approach for data analysis). After that they were asked to sign an informed consent form (Appendix C) and fill out a demographics sheet (Appendix E). Based upon the experimental condition to which they were assigned, the analysts were familiarized (see appendix H for more information about the familiarization method that was used for both types of analysis methods) in either the activity theory based method or the traditional method of analysis.

As part of the familiarization process, the analysts were provided with an instructional packet. The packet included factual material about the experiment itself and some background information about the method of analysis that would be used (See Appendix I for a sample, analysts were provided information that was applicable for the treatment condition to which they were assigned). In addition, two sample task videos along with their analysis were included. The videos were pre-analyzed using both the methods of analysis; however each analyst was only shown the one that applied to them. The experimenter demonstrated the method using the first sample video. The experimenter clarified questions after which the analysts were required to perform the analysis on the second sample video. The results of this analysis were

compared with those generated from the pre-analysis. The analysts had to successfully complete at least 75% of the steps required as part of the analysis procedure that they were required to follow. If the analysts did not meet this criterion, the errors in their task performance were pointed out to them and they were asked to repeat the analysis. Once the analysts demonstrated the required proficiency in performing the analysis, they were provided with a videocassette that contained the videos to be analyzed based upon the assigned experimental condition. The analysts analyzed the data using the assigned method and returned the results of the analysis (completed data analysis templates) to the experimenter. The methods are elaborated below.

3.3.7.1. Traditional Method

Traditionally there is no one universally accepted method for analyzing data gathered from a think aloud method. The methods used in Ericsson and Simon's model for analyzing verbal protocols are detailed and complicated. They suggest the need for transcription, encoding and then segmentation of the data before the actual analysis is performed. Ericsson and Simon implored on the problems related to segmentation of the protocol (p.205). In addition, Ericsson and Simon were primarily interested only in verbalizations with content from 'Short Term Memory'. They were thus focused only on concurrent TA. They used retrospective TA only as a means of verification of the participant being truthful. For the purpose of this research, the traditional method that was considered followed the approach of categorization of information as mentioned in several cases in literature.

Verbal protocol data has been categorized in several different ways; by parts that are specific to the system being tested (Denning et al., 1990), by the lighting up of indicator lights (Ebling and John, 2000), by information processing classifications

(Triggs et al., 1990) and type of verbalization (Bowers and Snyder, 1990), to name a few. For this experiment, a data analysis template was created that allowed the analyst to categorize the verbalizations from the task performance into positive or negative critical incidents (see appendix F). Once the categorization was completed, the analysts reviewed the results, identifying usability issues and suggesting recommendations for improvement. Once that was done, they assigned a severity rating to each item.

3.3.7.2. Activity Theory based method

This research has created a tool based upon activity theory for analyzing verbal protocols. The proposed method makes use of this tool, a data analysis template, to have the analyst follow activity theory while performing the analysis. As mentioned earlier, in the section on alternative methods, the purpose of this method was to identify the perceived goals of the person performing the task based upon the ongoing activity (see figure 4 below).

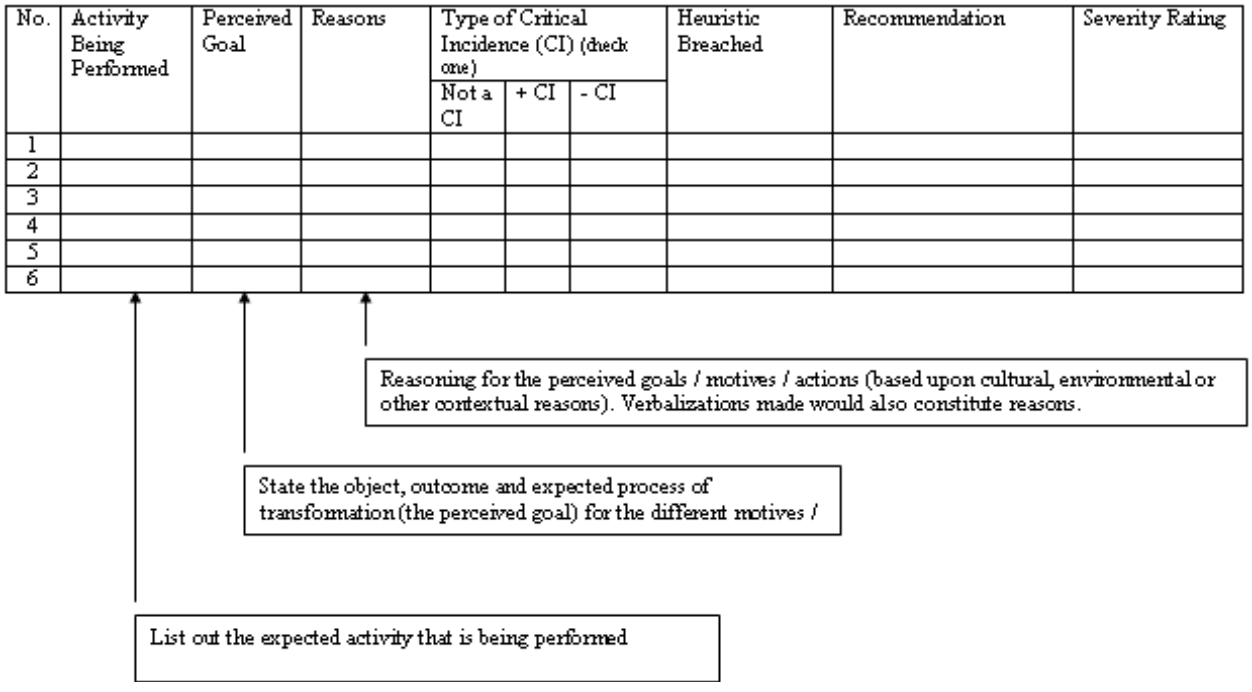


Figure 4 Activity Theory based analysis template

The template uses an activity as the unit of analysis. During the analysis, the analyst listed out each activity in the first column of the template. In the second column, the analyst indicated what they perceived as being the goal of this activity. In doing so, they identified the object, the outcome and the assumed process of transformation of the object to outcome. In column three, the analyst listed out the reasoning employed to make this assumption. The reasoning included the cultural, contextual or environmental issues or considerations made to reach the conclusion mentioned in column two.

Thus, the differences between the two methods include the unit of analysis (verbalization v/s activities) and the incorporation of additional contextual factors in the activity theory based method. The verbalizations aided the analysts in confirming the possible goals of the user while performing the activity. Then these activities were categorized based upon the occurrence of critical incidents. The analysts then provided a severity rating to the usability issues identified.

3.3.8. Data Analysis

The analysis consisted of two independent variables, Analysis Method and Verbal Protocol, they were analyzed with an analysis of variance (ANOVA) for the following dependent measures:

- Positive Critical Incidents
- Negative Critical Incidents
- Number of Negative Critical Incidents of the same Severity Rating
- Subjective ratings for three performance measures of usability evaluation methods

The original experimental design was a two factor, between subjects design (Table 7). The structural model is $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{k(ij)} + \alpha\beta_{(ij)} + \varepsilon_{l(ijk)}$. Table 10 shows the between-subject design ANOVA.

Table 10 Between Subject Design Used

Sources of Variation	Degrees of Freedom
Method of Analysis (M)	1
Verbal Protocol (D)	1
S/MD	4
M X D	1
Total	7

4. RESULTS

The actual data analysis considered task type as an independent variable in addition to the method of analysis and type of verbal protocol for the analysis of the positive and negative critical incidents. In the case of the severity levels, task type was not considered as the rating was provided with respect to the impact of the negative critical incident under consideration which was not dependent upon the task. In the case of the subjective ratings, task type was not considered since the rating was provided for the method overall. As mentioned earlier, one of the purposes of this research was to compare the proposed activity theory based method of analysis with a traditional method of analysis. In order to perform this comparison, four different hypotheses were tested. The data were analyzed by running a series of ANOVAS. Two ANOVAs were run on each set of dependent measures. A full linear model was used to identify the significance for all involved interactions. Then a reduced linear model was used which included only those interactions that were significant at the 20% level (for the full linear model) or if even one of the components might be accounting for significance (Bozivich, Bancroft & Hartley, 1956). SAS (Alpha = 0.05) was used to perform the analysis. The ANOVA model considered the four types of tasks as a within subjects factor.

4.1. Hypothesis 1

Hypothesis 1 was that the method of analysis and type of verbal protocol will not have an effect on the number of positive critical incidents identified.

The linear models for the ANOVAs are listed below along with descriptions of the terms used.

Full Model:

$$y = \mu + M_i + D_j + M*D_{ij} + S(M*D)_{(ij)k} + T_m + M*T_{im} + D*T_{jm} + M*D*T_{ijm} + \epsilon_{ijklm}$$

M – Method D – Data Type

S – Participants T – Task

Table 11 Test of fixed effects for Positive CI (Full Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	4	0.44	0.54
Data_type	1	4	0.59	0.48
Method*Data_type	1	4	0.00	0.97
Task	3	108	5.32	0.002*
Method*Task	3	108	1.42	0.24
Data_type*Task	3	108	0.92	0.43
Method*Data_typ*Task	3	108	1.12	0.34

Reduced Model:

$$y = \mu + M_i + D_j + S(M*D)_{(ij)k} + T_m + \epsilon_{ijklm}$$

M – Method D – Data Type

S – Analysts T – Task

Table 12 Test of fixed effects for Positive CI (Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	5	0.54	0.49
Data_type	1	5	0.74	0.43
Task	3	117	5.26	0.002*

No significant differences were found in the number of positive critical incidents identified based upon the method of analysis used (Activity Theory based or Traditional).

No significant differences were found in the number of positive critical incidents identified based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol).

The task type was found to have a significant difference on the number of positive critical incidents that were identified. The tasks themselves were selected by the company that manufactured the product. This research was not focused on studying the relevance of task type to the number of critical incidents identified, however a post hoc Tukey adjustment was performed on the task types, which revealed a significant difference between the number of positive critical incidents identified by task A and C and task A and D. Figure 5 below displays the mean positive critical incidents identified for the various task types.

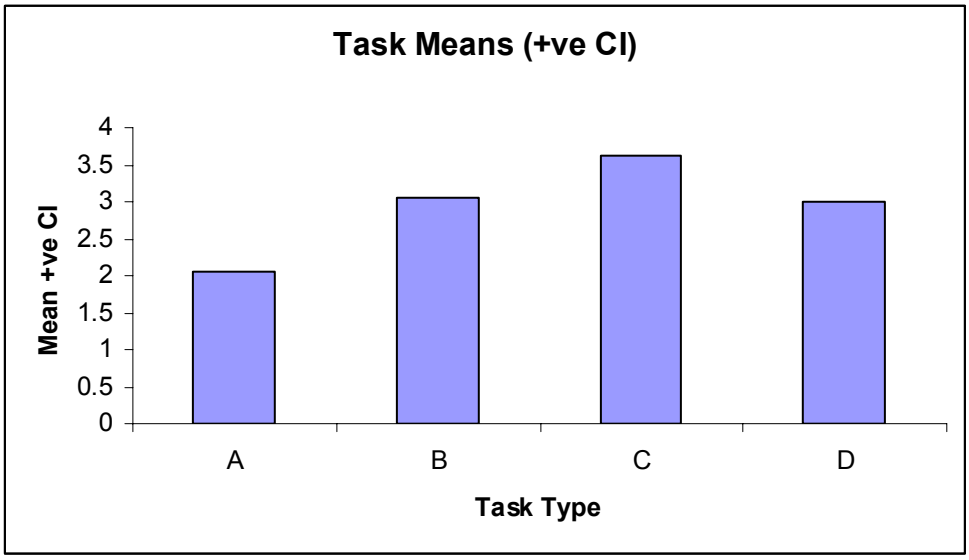


Figure 5 Graph of the mean positive critical incidents identified for various tasks

The M*T interaction displayed a trend, a plot of the mean positive critical incidents identified by method and task is shown below along with the trend lines. The activity theory based method appears to be able to identify more positive critical incidents as compared to the traditional method.

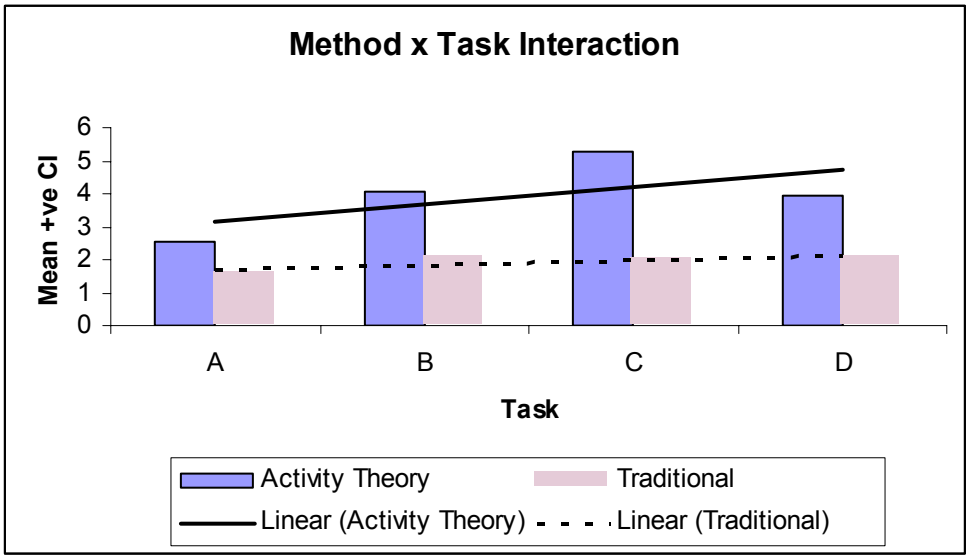


Figure 6 Graph of method x task interaction

4.2. Hypothesis 2

Hypothesis 2 was that the method of analysis and type of verbal protocol will not have an effect on the number of negative critical incidents identified.

The linear models for the ANOVAs are listed below along with descriptions of the terms used.

Full model and reduced model:

$$y = \mu + M_i + D_j + M*D_{ij} + S(M*D)_{(ij)k} + T_m + M*T_{im} + D*T_{jm} + M*D*T_{ijm} + \epsilon_{ijklm}$$

M – Method D – Data Type

S – Analysts T – Task

Table 13 Test of fixed effects for Negative CI (Full and Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	4	0.04	0.84
Data_type	1	4	0.00	0.96
Method*Data_type	1	4	0.44	0.54
Task	3	108	11.63	<.0001*
Method*Task	3	108	1.49	0.22
Data_type*Task	3	108	0.52	0.67
Method*Data_typ*Task	3	108	2.69	0.049*

Since the M*D*T interaction was found to be significant, the M*T, M*D and D*T interactions were not discarded for the reduced model as these might act as mediating variables for the M*D*T interaction.

No significant differences were found in the number of negative critical incidents identified based upon the method of analysis used (Activity Theory based or Traditional).

No significant differences were found in the number of negative critical incidents identified based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol)

The task type was found to have a significant difference on the number of negative critical incidents that were identified. The M*D*T interactions displayed a significant difference in the number of negative critical incidents identified. The tasks themselves were selected by the company that manufactured the product. This research was not focused on studying the relevance of task type to the number of critical incidents identified, however a post hoc Tukey adjustment was performed. In the case of the task effect, it revealed a significant difference between the number of negative critical incidents identified by task groups A and B; A and D; B and C and C and D. A graph of the mean negative critical incidents identified by task type is shown below.

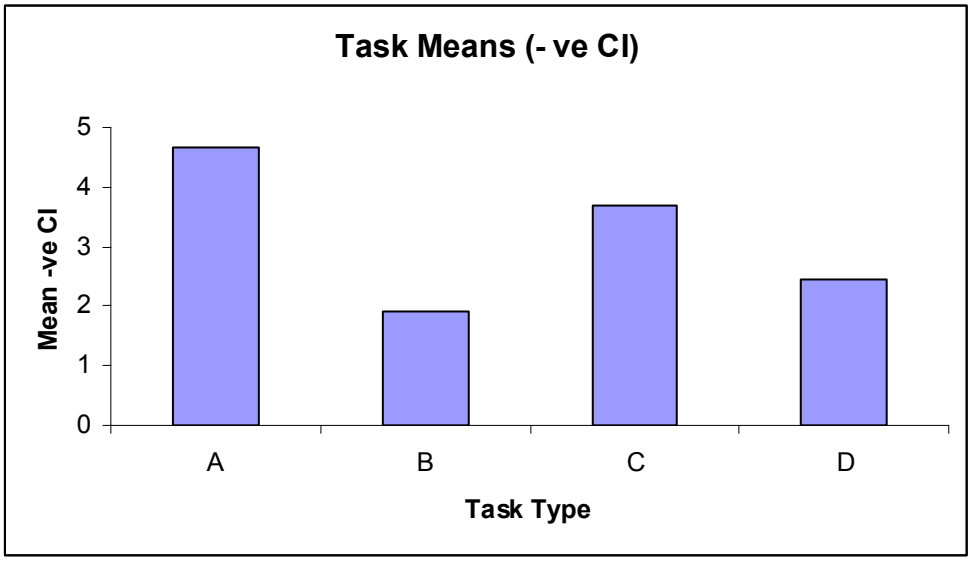


Figure 7 Graph of the mean negative critical incidents identified for various tasks

In the case of the M*D*T interaction, the post hoc analysis revealed that a significant difference exists between the number of negative critical incidents identified by task A and D when the type of method used was activity theory and type of verbal protocol was concurrent. A plot of the mean negative critical incidents identified for this treatment condition is shown below.

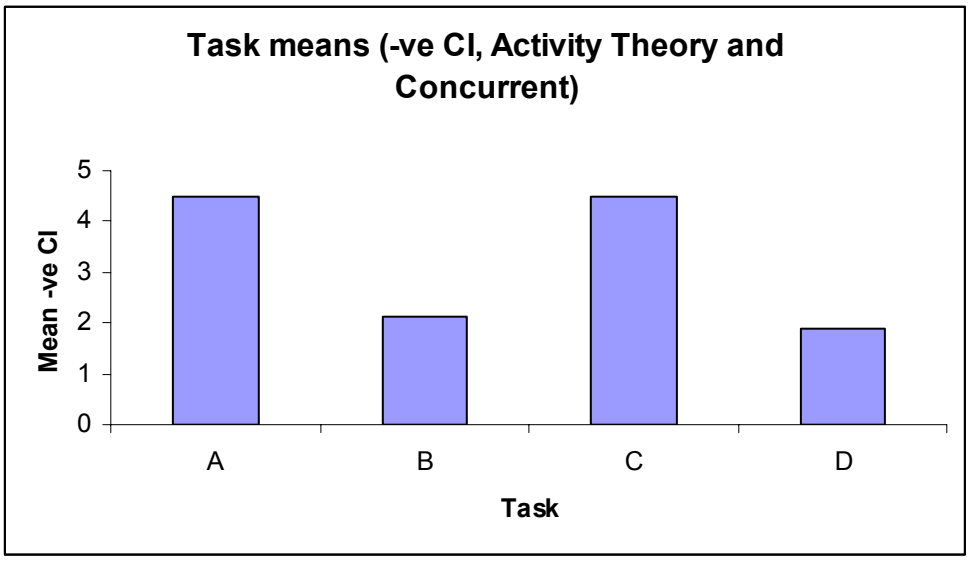


Figure 8 Graph of mean negative critical incidents identified using activity theory based method of analysis on concurrent verbal protocol

A significant difference was also found between the number of negative critical incidents identified by task A and B and task A and D when the type of method used was traditional and type of verbal protocol was retrospective. A plot of the mean negative critical incidents identified for this treatment condition is shown below.

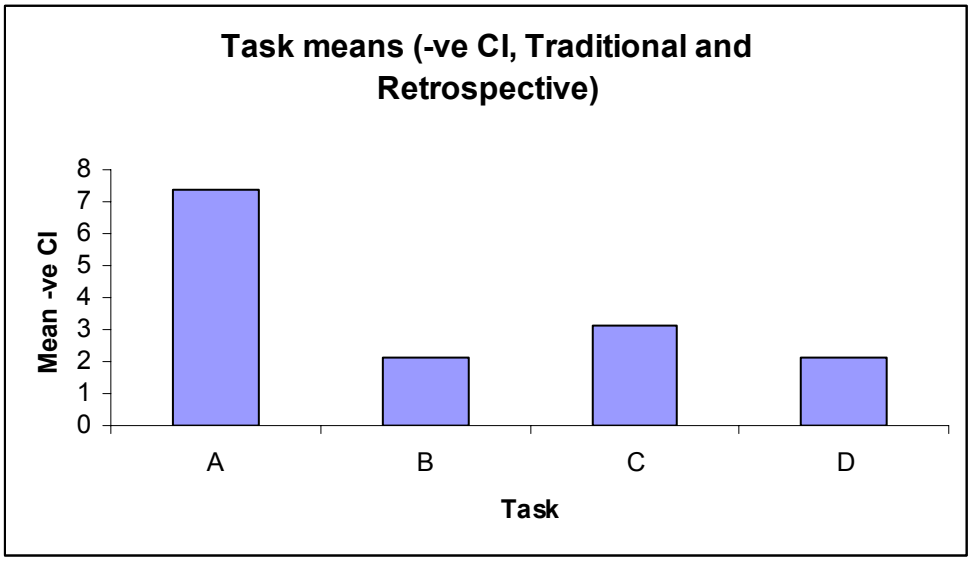


Figure 9 Graph of mean negative critical incidents identified using activity theory based method of analysis on concurrent verbal protocol

4.3. Hypothesis 3

Hypothesis 3 was that the method of analysis and type of verbal protocol will not have an effect on the number of negative critical incidents of the same severity ratings. In order to test this, the severity ratings were classified into two groups, Low Severity and High Severity. Tasks were not considered as an independent variable in the ANOVAs run for this hypothesis since the severity rating provided was dependent on the impact of the incident itself.

4.3.1 Low Severity

The total numbers of low severity (severity rating 1 or 2) negative critical incidents identified were considered for this analysis. After reviewing the data, it was seen that a total of only 5 negative critical incidents were classified as level '0' by all the analysts together and were hence discarded from the analysis. Due to the high level of complexity that arose while analyzing these results, ANOVAS were run by considering

the averages for the level over the persons and the tasks. The linear model for the ANOVA is listed below along with descriptions of the terms used.

Full model:

$y = \mu + M_i + D_j + MD_{ij} + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 14 Test of fixed effects for Low Severity Level Negative CI (Full Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	4	0.97	0.38
Data_type	1	4	0.05	0.84
Method*Data_type	1	4	0.01	0.93

Reduced model:

$y = \mu + M_i + D_j + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 15 Test of fixed effects for Low Severity Level Negative CI (Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	5	1.21	0.32
Data_type	1	5	0.06	0.82

No significant differences were found in the number of low severity negative critical incidents identified based upon the method of analysis used (Activity Theory based or Traditional).

No significant differences were identified in the number of low severity level negative critical incidents that were identified based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol).

4.3.2. High Severity

The total numbers of high severity (severity rating 4 or 5) negative critical incidents identified were considered for this analysis. Due to a high level of complexity that arose while analyzing these results, ANOVAS were run by considering the averages for the level over the persons and the tasks. The linear model for the ANOVA is listed below along with descriptions of the terms used.

Full model:

$y = \mu + M_i + D_j + MD_{ij} + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 16 Test of fixed effects for High Severity Level Negative CI (Full Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	4	1.38	0.31
Data_type	1	4	0.00	0.98
Method*Data_type	1	4	0.62	0.47

Reduced model:

$y = \mu + M_i + D_j + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 17 Test of fixed effects for High Severity Level Negative CI (Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	5	1.49	0.28
Data_type	1	5	0.00	0.98

No significant differences were found in the number of high severity negative critical incidents identified based upon the method of analysis used (Activity Theory based or Traditional).

No significant differences were identified in the number of high severity level negative critical incidents that were identified based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol).

4.4. Hypothesis 4

Hypothesis 4 was that the method of analysis and type of verbal protocol will not have a significant effect on the subjective ratings for any of the performance measures.

4.4.1. Thoroughness

The linear model for the ANOVA is listed below along with descriptions of the terms used.

Full Model:

$y = \mu + M_i + D_j + M*D_{ij} + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 18 Test of fixed effects for Subjective Ratings of Thoroughness (Full Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	4	1.29	0.32
Data_type	1	4	0.14	0.72
Method*Data_type	1	4	1.29	0.32

Reduced model:

$y = \mu + M_i + D_j + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 19 Test of fixed effects for Subjective ratings of Thoroughness (Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	5	1.22	0.32
Data_type	1	5	0.14	0.73

No significant differences were found in the subjective ratings given for thoroughness based upon the method of analysis used (Activity Theory based or Traditional).

No significant differences were identified in the subjective ratings given for thoroughness based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol).

4.4.2. Validity

The linear model for the ANOVA is listed below along with descriptions of the terms used.

Full Model:

$y = \mu + M_i + D_j + M*D_{ij} + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 20 Test of fixed effects for Subjective Ratings of Validity (Full Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	4	0.09	0.78
Data_type	1	4	0.09	0.78
Method*Data_type	1	4	0.09	0.78

Reduced model:

$y = \mu + M_i + D_j + S(M*D)_{(ij)k} + Є_{ijk}$

M – Method D – Data Type

S – Analysts

Table 21 Test of fixed effects for Subjective ratings of Validity (Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	5	0.11	0.75
Data_type	1	5	0.11	0.75

No significant differences were found in the subjective ratings given for validity based upon the method of analysis used (Activity Theory based or Traditional).

No significant differences were identified in the subjective ratings given for validity based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol).

4.4.3. Downstream Utility

The linear model for the ANOVA is listed below along with descriptions of the terms used.

Full Model:

$y = \mu + M_i + D_j + M*D_{ij} + S(M*D)_{(ij)k} + Є_{ijk}$

M – Method D – Data Type

S – Analysts

Table 22 Test of fixed effects for Subjective ratings of Downstream Utility (Full Model)

Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Method	1	4	2.67	0.18
Data_type	1	4	0.67	0.46
Method*Data_type	1	4	0.67	0.46

Reduced model:

$y = \mu + M_i + D_j + S(M*D)_{(ij)k} + \epsilon_{ijk}$

M – Method D – Data Type

S – Analysts

Table 23 Test of fixed effects for Subjective ratings of Downstream Utility (Reduced Model)

Effect	Num DF	Den DF	F Value	Pr > F
Method	1	5	2.86	0.15
Data_type	1	5	0.71	0.44

No significant differences were found in the subjective ratings given for downstream utility based upon the method of analysis used (Activity Theory based or Traditional). No significant differences were identified in the subjective ratings given for downstream utility based upon the data type used (Concurrent Verbal Protocol or Retrospective Verbal Protocol).

4.4.4 Agreement

The data collected for the subjective ratings was also analyzed for intra-class correlation to test for level of analyst agreement using Pearson's Test (assumes H_0: **ρ** = 0, where **ρ** is the correlation co-efficient) at a significance of 0.05. The null hypothesis cannot be rejected for the first treatment condition (Method: Activity Theory; Verbal Protocol: Concurrent), r(2)=-0.87, p = 0.33. A strong negative correlation for treatment condition two (Method: Traditional; Verbal Protocol: Concurrent), r(2)=-1.00, p < 0.001, two-tailed was found. Thus the null hypothesis is rejected at a 95% confidence level. This may be interpreted as a significant inverse agreement between the analysts for treatment condition 2.

In the case of treatment condition 3 (Method: Activity Theory; Verbal Protocol: Retrospective) and condition 4 (Method: Traditional; Verbal Protocol: Retrospective), the Pearson's correlation could not be computed as the estimated standard deviation was zero. Hence no further results could be drawn for these conditions. The estimation of p values for the null hypothesis might not be robust due to the low number of data points.

4.5. Qualitative Data

After the participants completed the analysis, they were interviewed over telephone to collect qualitative data about their experience with the method of analysis that they had to use. The participants were all asked the same set of questions (see Appendix G for the list of interview questions). The experimenter took notes during the interview. These responses were verified with the participants by email exchange. Some interesting comments from these interviews are provided below (they are listed out as advantages and disadvantages).

Table 24 Activity Theory based method Interview Comments

Advantages
“It provides a detailed view of how the user performs the task. This allows you to know the action sequence.”
“Instead of filling out a questionnaire, I could provide my own detailed opinions about the analysis.”
“It was kind of difficult to understand in the beginning, but after some practice my speed improved.”
Disadvantages
“The video is good, but incase I loose context while watching the video, I do not have access to someone who can clarify my concern and I have to make an assumption (still it is better than a transcript).”
“Someone who might not have a usability background might find this method difficult to use”
“The overall analysis took too long for me to complete”

Table 25 Traditional method interview comments

Advantages
“Liked the idea of associating the verbalizations with heuristics since it helped me to understand what was happening overall.”
“You are able to get a lot of details from the verbalizations”
“Liked the fact that it was useful to point to the actual usability issues and not just focus on their verbalizations.”
“Liked the concept of paying more attention to the verbalizations as compared to the actual actions as I have followed the similar approach in my work as well.”
Disadvantages
“I would have added a column to indicate actions that might have been associated with the verbalizations. I feel that this would make the analysis more comprehensive and better help to understand how to fix the problem.”
“Time consuming if you have to be thorough.”

5. DISCUSSION

This research had two major goals. To study the information available in existing literature regarding the think aloud method and help to improve its application and create an alternative approach to analyze the data that is collected. This chapter discusses the various findings that were made.

5.1. Applying Think Aloud

The literature review that was conducted as part of this research identified the importance of getting participants familiarized with Think Aloud. The Think Aloud method often places the participant in the most unusual of circumstances, they are asked to talk to themselves usually in the presence of another person and, very often, while being videotaped. In order for a think aloud to be effective, it is important that the participant feels a minimum amount of discomfort and uneasiness with the concept. At the same time, it is important that the process be as natural to the participant as possible. Providing information to the participant about what the method is and showing by example might help alleviate their anxiety and produce good data from the experiment. The approach mentioned below will provide a step wise approach to familiarize a participant with the Think Aloud method (before they provide the actual verbalizations) based upon documented information from literature.

1. Instructions and example

The participant is at first instructed about the Think Aloud method. The instructions need not get into the details of the method itself. The participant should however be told that the reason they need to try to talk continuously is so that the

experimenter can understand what it is that they are thinking when they perform the tasks needed as part of the study.

The instructions should mention all the events that would occur during the Think Aloud performance. By letting the participant know about the events that can occur (related to Thinking Aloud) before hand, the participant will not have to face any unexpected surprises. A sample instruction is provided below:

"I would like you to continuously say what you are thinking while you are performing the task that is given. Try to speak to yourself as if I am not present in the room. Do not make a judgment of what to say and what not to. Say everything that you think about while performing the task. Incase you fall silent; I will prompt you to restart verbalizing. I do not intend to disturb you while you are performing the task."

The participant might be able to understand better if the experimenter demonstrates by example. For the purposes of demonstration, the experimenter can Think Aloud while performing a simple task using things that might be present in the room or purely by imagination.

2. Warm-up with example

Once the instructions have been given, the participant should be asked to perform some warm ups. They can be asked to think aloud while performing a sample task and the experimenter can prompt them to continue talking incase they stop.

The warm-up will allow the participant to get their first feel of Thinking Aloud. They might feel awkward and stop talking at times. The experimenter can thus put them through the routine of being asked to "Keep Talking". If the participant is particularly inept, the experimenter can demonstrate the method once again and have the

participant perform another warm-up. These warm-ups and demonstrations can be provided for short tasks.

3. Description of Prompt (to make the participant start talking again)

The participant should also be told about the fact that the experimenter will prompt him/her to continue speaking incase they stop. By stating this before hand, the participant might not be surprised or offended during the actual task performance. The prompt itself should be non-directive, concise and non-intimidating in nature, the intent is to coax the participant to continue talking. An example of a prompt could be “Keep Talking”, or “What are you thinking?”

4. Determination of when to prompt and why

The participant should also be told when the experimenter would prompt. This is usually a certain number of seconds (15 – 60 sec) after the participant has stopped speaking. The participant should be informed that the intent of the examiner is only to get them to continue thinking aloud.

5. Debriefing after the trial is completed

Once the trial is completed, the participant should be debriefed. The experimenter can make use of this opportunity to let the participant know about situations where they were silent, enquire why that was the case and demonstrate what was expected. The participant should be encouraged through out the familiarization process. The increased confidence in the participant might help to assuage the awkward feeling that they might have.

5.2. Making sense of the data

The Think Aloud method yields a very large amount of data. This research intended to provide an alternative to the existing approaches for logging and analyzing these vast quantities of information. A template was created to be used as a tool by analysts to catalog the data that was collected. This template was designed so that the analyst used the activity theory approach to perform the analysis. The results of this analysis were then compared with results obtained by the analysis of the same data using a traditional method.

5.2.1. Activity Theory based data logging template

This research used Activity Theory as an alternate approach for analyzing Think Aloud data. A data logging and analysis template (see Appendix F) was created that used an activity as the basic unit of analysis. The template allowed the analyst to focus their analysis to understand what it was that the person performing the task was attempting to accomplish and why they used a particular way of going about it. The template was structured so that the analyst documented the possible perceived goals of the persons performing the tasks and the reasons why certain steps were performed. In doing so the analyst followed the tenets of activity theory while performing their analysis. This template can be used as a tool by future researchers to analyze data collected in the form of verbalizations.

The post task interviews of analysts using the traditional approach indicated that they felt the method was deficient in its capabilities for focusing on actions and recommendations were made to allow for affordances in the data analysis templates to account for actions as well. Analysts who used the activity theory based method

indicated that they were able to focus on actual issues since the task of identifying the perceived goal required them to pay additional attention to the users expectations.

5.2.2. Comparison of proposed method with traditional

The results from the quantitative data analysis suggest that the null hypotheses could not be rejected. There was no significant difference in the effectiveness of the proposed activity theory based method of analysis as compared to the representative traditional method of analysis. The analysis of the data gathered did not show that either method was significantly superior to the other with respect to the number of critical incidents either positive or negative that were identified. The analysis of the subjective ratings gathered did not show a significant difference in what the participants perceived to be the thoroughness, validity or downstream utility of either method.

Although there was a difference between the numbers of high and low severity critical incidents that were identified when the traditional method of analysis was used, the activity theory based method was found to be equally effective at identifying both high and low severity critical incidents. This indicates that the activity theory based method of analysis could be used to identify critical incidents at all severity levels; it is also possible that the activity theory method identifies more false positives. Both the methods identified a very low number of negative critical incidents at the ‘0’-severity level. This could indicate that the methods are not sensitive enough towards identifying such types of critical incidents.

The post hoc analyses revealed that the task types had a significant effect on the number of positive and negative critical incidents identified. These differences can be a result of the dissimilarities in the task attributes. The tasks required the use of different

features of the product. Some of the task attributes that might have played a role include the level of awareness of the feature being tested, the level of complexity of the task (with respect to the number of steps that would be required to be performed in order to complete the task), the fact that some tasks were comprised of multiple parts, etc.

The analysis of the dependent measures indicated that the type of verbal protocol had no effect on the results. This is in contrast to the findings of Bowers and Snyder (1990) that the retrospective protocol yielded richer data (richer data can help to identify more critical incidents). This implies the possibility that the methods of analysis were not sensitive enough to allow the analysts to make use of the richer data that was available from the retrospective protocol.

Pearson's correlation results for the level of agreement between the analysts for the subjective ratings indicated an inverse agreement when the traditional method of analysis was used to analyze concurrent verbal protocol. One of the analysts in that treatment condition had a computer science background while the other had a human factors background. This difference might have resulted in the negative agreement. A more stringent participation criteria and a rigid training design might help to improve the level of agreement between analysts.

It should be noted that the two methods were compared strictly on the basis of how information was categorized by the analysts. The volume of data collected was very large allow each individual usability concern or design benefit to be considered separately. Thus, the actual nature of issues identified might be different and the method of analysis could be a reason for this difference.

5.3. Limitations and lessons learned

5.3.1. Low Reliability

Due to time and cost constraints, only eight participants could be recruited to participate as analysts. Each treatment condition, therefore, had only two participants. The results of the analysis can have low reliability. If the volume of data to be analyzed were lower, it could have been possible to recruit a larger number of participants to analyze the data for the funds available. This would have resulted in a greater number of participants in each of the treatment conditions. Thus the number of data points collected would have increased allowing for results with higher reliability.

5.3.2. Number of incidents identified

This research might have provided different results if participants with different levels of experience with usability data analysis were available to participate. Those that were selected were considered to be beginners and this might have resulted in lower number of issues being identified (Nielsen, 1992). It is possible that participants who had more experience might have found one of the methods of analysis to be more effective than the other.

5.3.3. Analyst Fatigue

The data that was collected during think aloud was very rich in content. The analysis template that was created required the analysts to be very detailed in documenting their findings. This resulted in the analysis being very time consuming. This could have led to fatigue on the part of the analyst and might have resulted in erroneous analysis or incomplete analysis. Having the evaluators analyze the entire data set over several sessions where small portions of the data could be analyzed might help to mitigate these issues. Also, digitizing the data set can allow the analyst to use an extensive set of software that is available to perform the analysis.

5.3.4. Remote testing

The analysts were responsible for performing the analysis unsupervised. As a result it was not possible for the experimenter to observe the participants' (analysts of the data gathered) experience with the analysis. Observation of this behavior might have lead to additional realizations related to the different methods of analysis. Participants were instructed not to refer to any information other than that provided by the experimenter during the initial familiarization session and to consult only the experimenter via email or phone about questions, however there is no method to verify this. As a result, participants might have been influenced by external sources.

5.4. Future research

This research provides several avenues for future research. Some possibilities are listed below:

5.4.1. Dependent measures

This research compared two different methods of analysis based upon how data was categorized during analysis. It is possible that the dependent measures that were chosen for this categorization (critical incidents) might not have the required amount of sensitivity to allow for an effective comparison. Alternate dependent measures can be considered. One possible measure could be the amount of time taken to complete the analysis. This will require for the analysis to be supervised in order to ensure that the analysts do not doctor their results. Criterion based evaluations could also be performed where the critical incidents that are identified as part of the analysis are compared with those identified by a group of experts. Alternate measures of effectiveness can be considered like Skills/Rules/Knowledge based evaluation as mentioned by Rasmussen (1983).

5.4.2. Training

The template design is such that it will ensure that the analyst follows activity theory. However, the analyst has to be adequately trained to use this tool for it to be effective. A suitable criterion set should be identified to which the analysts can be trained such that the methods are followed rigorously. A stringent training protocol should be developed to achieve the same.

5.4.3. Tool usability

This research resulted in the creation of a tool that can be used by usability practitioners to analyze data using activity theory. The usability of the tool itself should be tested so that its effectiveness and efficiency can be improved.

5.5. Conclusions

Some conclusions made from this research are listed below:

- Both methods of analysis were equally effective in identifying positive critical incidents
- Both methods of analysis were equally effective in identifying negative critical incidents
- There was a significant difference in the number of Low and High severity critical incidents identified when the traditional method of analysis was used.
- Task attributes had a significant effect on the results produced by the two methods
- Analyst recruitment needs to be performed using a stringent set of criteria and in depth training is required before actual analysis is performed.

REFERENCES

Baker, M. J. (2002), Research Methods. The Marketing Review, vol. 3, no. 2. Westburn, Helensburgh, Scotland. pp. 167 – 193. (ISSN 1469-347X/2002/2/00167 + 26).

Bannon, L. [updated: ver 2.0 30 September 1997]. Activity Theory [online] http://www-sv.cict.fr/cotcos/pjs/TheoreticalApproaches/Actvity/ActivitypaperBannon.htm, [Accessed 14 September 2005].

Bederson, B. B., Lee, B., Sherman, R. M, Herrnson, P. S and Niemi, R. G. (2003), Electronic voting system usability issues. Proceedings of the conference on Human factors in computing systems, Ft. Lauderdale, FL, USA. pp. 145 – 152.

Boren, M. T and Ramey, J. (2000), Thinking Aloud: Reconciling Theory and Practice, IEEE Transactions on Professional Communication, vol. 43, no.3, pp. 261 – 278.

Bowers, V. and Snyder, H. (1990), Concurrent versus retrospective verbal protocols for comparing window usability. Proceedings of the Human Factors Society 34th annual meeting, pp. 1270 – 1274.

Bozivich, H,. T.A. Bancroft, and H.O. Hartley. 1956. Power Analysis of Variance Test Procedures for Certain Incompletely Specified Models, I. Ann. Math. Statist. 27:1017-1043.

Burr, J and Bagger, K. (1999), Replacing usability testing with user dialogue – How a Danish manufacturing company enhanced its product design process by supporting user participation, Communications of the ACM, vol. 42, no. 5, pp. 63 – 66.

Carrol, J.M., Koenemann-Belliveau, J., Rosson, M.B., & Singley, M.K. (1993). Critical incidents and critical threads in empirical usability evaluation. In Alty, J., Diaper, S., & Guest, S.P. (Eds.), People and Computers VIII: Proceedings of the HCI-93 Conference. Cambridge, U.K.: Cambridge University Press, pp. 47-55.

Castillo, J. C., Hartson, R. H., & Hix D. (1998). Remote usability evaluation: Can users report their own critical incidents? Summary of CHI'98 Human Factors in Computing Systems, 253 – 254.

Deffner, G. (1990), Verbal Protocols as a research tool in human factors. Proceedings of the Human Factors Society 34th Annual Meeting, October 8 – 12, Orlando, FL. pp. 1263 – 1264.

Denning, S., Hoiem, D., Simpson, M. and Sullivan, K. (1990), The value of thinking-aloud in industry: A case study at Microsoft Corporation. Proceedings of the Human Factors Society 34th Annual Meeting, October 8 – 12, Orlando, FL. pp. 1285 – 1289.

Dumas, J. S. and Redish, J. C. (1993), A practical guide to Usability Testing. Ablex, Norwood, NJ.

Daft, R.L., and Lengel, R.H. (1986). Organizational information requirements, media richness and structural design, Management Science, 32:5, 554-571.

Ebling, M. R. and John, B. E. (2000), On the contributions of different empirical data in usability testing. Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques. ACM Press, New York, NY, USA. pp.289 – 296.

Engestrom, Y. (1987). Learning by expanding. Helsinki: Orientakonsultit.

Ericsson, K. A. and Simon, H. A. (1984), Protocol Analysis: Verbal Reports as Data. Cambridge, MA: MIT Press.

Flanagan, J.C. (1954). The critical incident technique. Psychological Bulletin (July), 51(4), 327-358.

Hartson, H. R., Andre, T. S., and Williges, R. C. (2003). Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction, 15(1), 145 – 181.

Hertzum, M. and Jacobsen, N.E. (1999). The evaluator effect during first-time use of the cognitive walkthrough technique. In Bullinger, H.J. and Ziegler, J. (Eds.), Human-Computer Interaction: Ergonomics and User Interfaces. Proceedings of the HCI International '99 (Vol. I, pp.1063-1067). London: Lawrence Erlbaum.

Hertzum, M., and Jacobsen, N.E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction, 13(4), 421-443.

Honold, P. (2000). Culture and context: An empirical study for the development of a framework for the elicitation of cultural influence in product usage. International Journal of Human-Computer Interaction, 12, 327-345.

Kaikkonen, A. and Roto, V (2003), Navigating in a Mobile XHTML Application. Proceedings of the conference on Human factors in computing systems, Ft. Lauderdale, FL, USA. pp. 329 – 336.

Karsenty, L. (2001), Adapting verbal protocol methods to investigate speech systems use, Applied Ergonomics, vol. 32, no. 1, pp. 15 – 22.

Kaptelinin, V. (1996). Activity Theory: Implications for human-computer interaction. In Nardi, B. (Ed.), Context and Consciousness: Activity theory and human-computer interaction. Cambridge, MA: MIT Press.

Kaptelinin, V., Nardi, B. A.,& Macaulay, C. (1999). The activity checklist: A tool for representing the space of context. Interactions, July-August, 27-39.

Kennedy, S. (1989), Using video in the BNR usability lab. ACM SIGCHI Bulletin 21, 2 (October), 92 – 95.

Kleiner, B. M. and Drury, C. G. (1998), The use of verbal protocols to understand and design skill-based tasks, in Human Factors and Ergonomics in Manufacturing, Vol. 8 (1) pp. 23 – 39. John Wiley and Sons, Inc.

Koenemann-Belliveau, J., Carroll, J. M., Rosson, M. B. and Singley, M. K. (1994), Comparative usability evaluation: critical incidents and critical threads, in Human factors in computing systems: "Celebrating interdependence", ACM Press, New York, NY, USA, Boston, Massachusetts, United States, pp. 245 – 251.

Kuutti, K. (1991). The concept of activity as a basic unit of analysis for CSCW research. Proceedings of the Second European Conference on Computer Supported Collaborative work (pp. 249-264). Amsterdam.

Kuutti, K. & Arvonen, T. (1992). Identifying potential CSCW applications by means of Activity Theory concepts: A case example. Proceedings of CSCW '92 (pp. 233-240). New York: ACM.

Kurniwan, S., King, A, Evans, D. G. and Blenkhorn, P, Design and user evaluation of a joystick –operated full-screen magnifier. Proceedings of CHI '03 (pp. 25-32). ACM press.

Lewis, C. (1982), Using the 'thinking-aloud method' in cognitive interface design. Research Report RC-9265, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA.

Markee, N. (2000), Conversation Analysis. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Mehlenbacher, B. (1993), Software usability: choosing appropriate methods for evaluating online systems and documentation, in Proceedings of the 11th annual international conference on Systems documentation, Waterloo, Ontario, Canada, pp. 209 – 222.

Monk, A. F. and Gilbert, N. (1995), Perspectives on HCI: Diverse Approaches. Academic Press, San Diego, CA, USA.

Nielsen, J. (1992), Finding usability problems through heuristic evaluation, in Proceedings of CHI'92 (Monterey, CA, May 1992), ACM Press, 373-380.

Nielsen, J. (1993), Usability Engineering. Academic Press, San Diego, CA, USA.

Nielsen, J. (1994), Estimating the Number of Subjects Needed for a Thinking Aloud Test, International Journal of Human-Computer Studies, vol. 41, no. 3, pp. 385 – 397.

Nielsen, J. and Mack, R. L. (1994), Usability Inspection Methods. John Wiley and Sons, NY. Chapter 2: Heuristic Evaluation (pp. 25 – 62)

Nielsen, Janni & Christiansen, N. (2000), Mindtape: A Tool for Reflection in Participatory Design, in Participatory Design, New York, pp. 303-313. Database, Esprit 3066, AMODEUS RP7/WP14.

Nielsen, Janni., Clemmensen, T. & Yssing, C., (2002), Getting access to what goes on in people's heads. Proceedings of the second Nordic conference on Human-computer interaction, Aarhus, Denmark, October 19-23, 2002. pp. 101 – 110.

Nisbett, R. E. and Wilson, T. D. (1977), Telling More Than We Can Know: Verbal Reports on Mental Processes, Psychological Review, vol. 84, no. 3, pp. 231 – 259.

Preece, J. (1994), Human-Computer Interaction, Addison-Wesley, England.

Rhenius, D. and Deffner, G. (1990), Evaluation of concurrent thinking aloud using eye-tracking data. Proceedings of the Human Factors Society 34th Annual Meeting, October 8 – 12, Orlando, FL. pp. 1265 – 1269.

Rosson, M. B. & Carrol, J. M. (2002), Usability Engineering: Scenario-based development of Human Computer Interaction. San Francisco: Morgan Kaufmann.

Rowley, D. E. (1994), Usability testing in the field: bringing the laboratory to the user, in Human factors in computing systems: “Celebrating interdependence”. ACM Press, New York, NY, USA, Boston, Massachusetts, United States, pp. 253 – 257.

Rubin, J. (1994), Handbook of Usability Testing: How to plan, design and conduct effective tests. John Wiley and Sons.

Scriven, M. (1967), The methodology of evaluation. In Tyler, R., Gagne, R. & Scriven, M. (Eds.), Perspectives of Curriculum Evaluation, pp. 39-83. Chicago: Rand McNally.

Selltiz, C., Jahoda, M., Deusch, M. and Cook, S. W. (1959), Research Methods. Social Relations, 2nd edition, London: Methuen

Spradley, J P (1980), Participant observation. New York: Holt, Rinehart and Winston.

Starr-Schneidkraut N, Cooper MD, Wilson SR. [updated: 2 May 2001]. 'Use of the Critical Incident Technique to Evaluate the Impact of MEDLINE', United States National Library of Medicine, [online] http://www.nlm.nih.gov/od/ope/cit.html (1989), [Accessed 27 January 2004].

Triggs, T. J., Kantowitz, B. H., Terrill, B. S., Bittner Jr., A. C. and Fleming, T. F. (1990), The playback method of protocol analysis applied to a rapid aiming task. Proceedings of the Human Factors Society 34th Annual Meeting, pp. 1275 – 1279.

Waes, L. V. (1998), Evaluating on-line and off-line searching behavior using thinking-aloud protocols to detect navigation barriers, in Proceedings of the sixteenth annual conference on Computer documentation, ACM Press, New York, NY, USA, Quebec, Canada, pp. 180 – 183.

Wright, R. B. and Converse, S. A. (1992), Method bias and concurrent verbal protocol in software usability testing. Proceedings of the Human Factors Society 36th Annual Meeting, pp. 1220 – 1224.

Virzi, R. A. (1990), Streamlining the design process: Running fewer subjects. Proceedings of the Human Factors Society 34th Annual Meeting. Pp. 291 – 294, Orlando, FL, 8 – 12 October.

Appendix A

Task List

Task List

The tasks listed below will be supplied to participants in accordance with one of the orders of presentation mentioned in table 5.

Task - A

A friend has been looking to purchase a desk. You feel that he/she might be interested in the desk that is in this room. Please perform the following tasks. After performing the first part, please press the 'End' key and close the phone.

Part 1: ***Adjust the resolution of the phone to 640 x 480.***

Please make sure that the phone is closed before you start performing the second part of the task.

Part 2: ***Open the phone and take as large a picture of the desk as possible. Use the 'Zoom' feature for this purpose.***

After you have taken the picture, please send this picture to your friends cell phone @ number 858-555-5555. After completing the task, please press the end key and close the phone.

(Note: **This note will not be included in the task sheet**. The task will be termed as completed once the participant presses the 'Send' key itself. The experiment will signal them to stop at this point. They would not be required to actually enter any information).

Task - B

You had taken a picture of a car earlier that was not the right side up. ***Please find that picture and rotate it by 180°.***

Once you complete the task, please press the end key and close the phone.

(Note: **This note will not be included in the task sheet**. There will be a single picture of a car in the pictures list and this picture will be buried in the Camera Pictures section of the media gallery so that the participant does not see it as soon as that feature is opened. As part of the experimental setup, the picture will be pre-modifed to be up side down).

Task - C

The experimenter's contact information already exists in the phone under 'Yogesh'. As a reminder of how you got to know him, ***Please take a picture of the TV set and assign it as his picture caller ID.***

Once you complete the task, please press the end key and close the phone.

(Note: **This note will not be included in the task sheet**. Since the experimenter will be filming, the participants cannot be asked to take a picture of the experimenter.).

Task - D

You had missed a call from the following phone number:
858-882-2705

Please save this number under the name ‘Mark’ as a work phone number.

Once you complete the task, please press the end key and close the phone.

Appendix B

Sample Task Sheet

Task Sheet

Task No. **A**

You had taken a picture of a car earlier that was not the right side up. ***Please find that picture and rotate it by 180°.***

Once you complete the task, please press the end key and close the phone.

Appendix C
Study 2 Informed Consent form

INFORMED CONSENT FORM

Title of Project: Comparison of an Activity Theory Based method with a traditional method when used to analyze verbal protocol data.

Principal Investigator: Yogesh D. Bhatkhande
Other Investigators: Tonya L. Smith-Jackson, PhD.

PURPOSE OF PROJECT
This study will compare two methods of data analysis (based on concepts of Activity Theory and Conversation Analysis respectively). The study will also add to the available literature related to participant opinions of being participants of the Think Aloud method.

INFORMATION
As part of this study you will be required to analyze data. You will be provided with data in the form of a videotape of people performing a set of tasks using a cellular phone. You are to investigate the data and identify Critical Incidents. These critical Incidents will have to be logged using a Data Analysis Template that will be provided to you.

RISKS
Participation in this project does not place you at more than minimal risk of harm.

BENEFITS
You will be compensated for your participation, and you will be given information to contact the principal investigator to get information about the outcomes of the study. You will also benefit from knowing that you have participated in worthwhile research that has immediate and positive applications.

CONFIDENTIALITY
The information gained in this research project will be kept strictly confidential. At no time will the researchers release the results of the study to anyone other than individuals working on the project without your written consent.

You will be identified only by a 3-digit study code. Data will be stored securely and will be made available only in the context of research publications and discussion. No reference will be made in oral or written reports that could link you to the data nor will you ever be identified as a participant in the project.

COMPENSATION
You will be compensated at the rate of $7.5 per hour (Seven Dollars and Fifty cents per hour) for participation in this research

FREEDOM TO WITHDRAW
You are free to withdraw from this study at any time without penalty

APPROVAL
This research project has been approved, as required, by the Institutional Review Board for Research Involving Human Participants at Virginia Polytechnic Institute and State University and by the Department of Industrial and Systems Engineering.

PARTICIPANT'S RESPONSIBILITIES
It is very important that you keep the activities and information discussed confidential, since others will be participating in this research.

QUESTIONS
If you have questions, or do not understand information on this form, please feel free to ask them now.

PARTICIPANT'S PERMISSION
I have read and understand the Informed Consent and conditions of this project. I have had all questions answered. I hereby acknowledge the above and give my voluntary consent for participation in this project.

If I participate, I may withdraw at any time without penalty.

Signature _____ *Date* _____

CONTACT
If you have questions at any time about the project or the procedures, you may contact the principal investigator, Yogesh D. Bhatkhande at 540-818-5298 or yogeshdb@vt.edu.

If you feel you have not been treated according to the descriptions in this form, or your rights as a participant have been violated during the course of this project, you may contact Dr. David Moore, Chair of the Institutional Review Board Research Division at 540-231-4991.

Appendix D

Study 1 Demographics Sheet

Demographics Sheet (Study 1)

Participant Number: _____

Age: _____ Gender: _____

Current Occupation: _____

Experience using a cell phone _____ (years)

Highest Education: □ **Some school**

(Please check) □ **High school**

□ **Some college**

□ **Undergraduate**

□ **Graduate**

□ **Postgraduate**

Have you heard about the "Think Aloud" method?

(Please check) □ **No**

□ **Yes**

If yes, have you ever used or been asked to perform the "Think Aloud" method?

(Please check)

□ **No**

□ **Yes**

Appendix E

Study 2 Demographics Sheet

Demographics Sheet (Study 2)

Participant Number: _____

Age: _____ **Gender: _____**

Current Occupation: _____

Highest Education: □ **Some school**

(Please check) □ **High school**

□ **Some college**

□ **Undergraduate**

□ **Graduate**

□ **Postgraduate**

Have you heard about the "Think Aloud" method?

(Please check)

□ **No**

□ **Yes**

If yes, have you ever used or been asked to perform the "Think Aloud" method?

(Please check)

□ **No**

□ **Yes**

Have you heard about "Activity Theory"?

(Please check)

- □ **No**
- □ **Yes**

If yes, please provide a brief explanation for your experience with "Activity Theory".

_____

_____

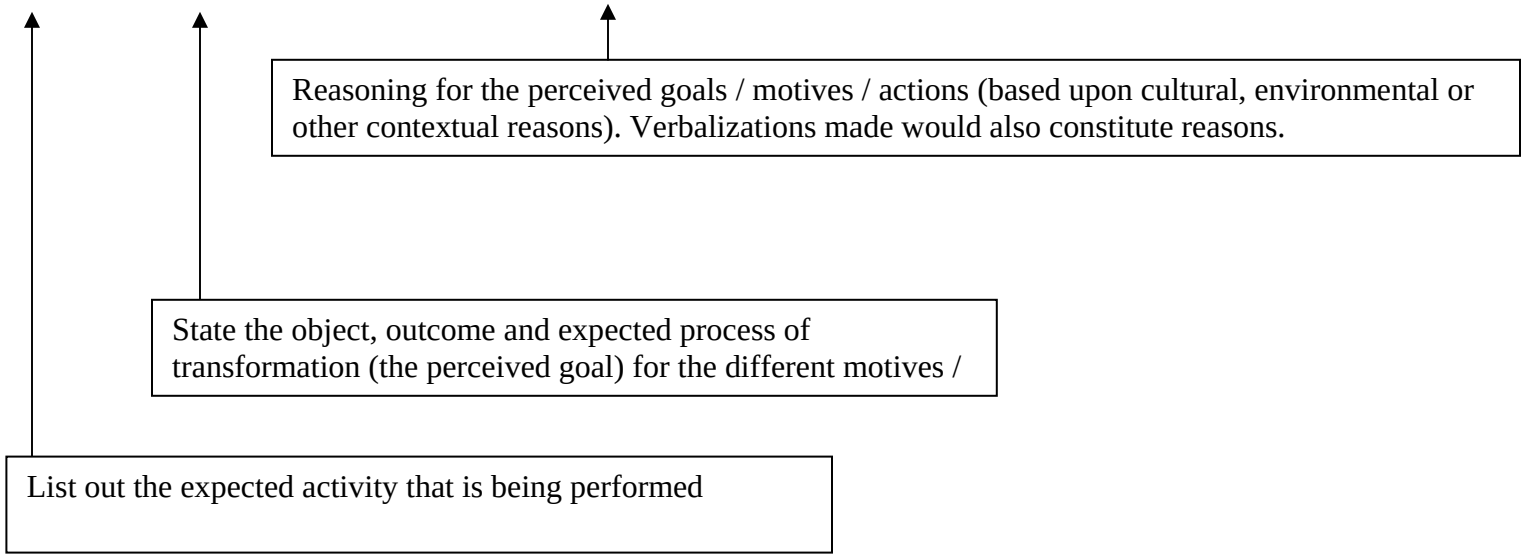
_____

_____

Appendix F

Analysis Data Templates

No.	Activity	Perceived Goals (PG) of Actions	Reason	Type of CI			Heuristic Breached	Recommendation	Severity Rating
				Not CI	+ CI	- CI			



Sample template for activity theory based method of Data Collection

Sample template for traditional method of Data Collection:

No.	Verbalization (spoken sentence)	Type of Critical Incidence (CI) (check one)			Heuristic Breached	Recommendation	Severity Rating
		Not a CI	+ CI	- CI			
1							
2							
3							
4							
5							
6							

Appendix G

Interview Questions

Interview Questions

Each participant will be asked the same set of questions. These are mentioned below:

a. Name 3 things that you liked about the method of analysis that you followed.

b. Name 3 things that you disliked about the method of analysis that you followed.

c. Do you feel that the method was easy to learn/follow/use?

d. During the performance of your analysis, did you employ any special technique (in addition to the instructions provided to you) which you feel aided you in some way (faster analysis, more detailed analysis)?

e. If you were asked to redesign this method, what part of the current design would you keep and why?

f. If you were asked to redesign this method, what part of the current design would you change and why?

Appendix H

Familiarization for Data Analysis

Familiarization for data analysis

The familiarization process will be divided into four portions. It is intended to be short in duration.

1. Introduction and Explanation

The participants should first be told about the purpose of the study and what is expected of them. The method of analysis is then explained to the participants based upon the treatment condition that they fall under.

2. Demonstration

The experimenter will then demonstrate the performance of the method using an example. This sample case can be unrelated to the actual data to be analyzed (the nature of the example should however be similar to the actual data; the theme and content can be different).

3. Warm-up

The participants will then be asked to perform a similar analysis using a different example. In case all participants are undergoing the familiarization together, the sample task should be appropriately sectioned so that everyone gets a chance.

4. Debrief

After the warm-up is over, the experimenter can identify shortcomings in the analysis made by the participant (if any). The experimenter should answer questions and issues that the participants might have. This will help the participants understand what is expected of them more clearly.

Appendix I

Analyst Information Sheet

Information Sheet (Concurrent/Activity Theory)

As a part of this experiment, you will be required to analyze video taped data of participants performing certain tasks. This document provides information regarding some concepts that will be used as part of the analysis and a detailed explanation of how to use a tool for performing the analysis.

Concepts:

- *Think Aloud Method*

The think aloud method is a popularly used data generation method. Subjects are directed to continuously speak their thoughts regarding given task or job. The spoken dialogue is often transcribed or recorded in some fashion (In the current case, transcription has not been made. Instead, a video recording of the task performance along with the participant's verbalization of thought has been made.). The collected information is called 'Verbal Protocol'.

There are several methods for the elicitation of Verbal Protocol. The data that will be presented to you has been collected using the Concurrent think aloud method. In Concurrent Think Aloud, the participant is asked to verbalize concurrently while performing a task.

The information collected from the think aloud method is often raw data and requires further analysis to be of use. For this experiment, you will be treated as a usability expert and will be required to review videotapes of task performances and the associated verbal protocol. The purpose of your analysis is to identify issues related to the usability of the product being tested. Some concepts that are relevant to the analysis are explained below followed by the method of analysis itself.

- *Critical Incidents*

A Critical Incident (CI) is an event that is significant to the purpose of analysis. In this case, the purpose of analysis is the usability of a product. Any event (this includes actions and verbalizations) that you identify in your analysis will have to be classified as either:
1. Not a CI → Not significant from a usability perspective
2. Positive CI → Significant from a usability perspective in a helpful manner (e.g.: aids task completion, matches users mental model, etc. In short some thing that is in line with known heuristics for a usable design)
3. Negative CI → Significant from a usability perspective in an unhelpful manner (e.g.: unnecessary or avoidable steps in task completion, does not provide sufficient information or too much information is provided, etc. In short some thing that breaches or breaks known heuristics for a usable design)

- *Heuristics for a usable design*

These are a collection of rules of thumb that have been suggested in literature to be considered while designing a product. These are seen as being conducive to the development of a more usable product. A few sample heuristics for a user interface are mentioned below (from http://www.useit.com/papers/heuristic/heuristic_list.html):

Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

User control and freedom

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Consistency and standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place.

Recognition rather than recall

Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Flexibility and efficiency of use

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Aesthetic and minimalist design

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

- *Severity Ratings*

This analysis will identify several usability related issues. In the industrial environment, it will not be possible to address all the issues identified within reasonable time to meet pre-existing commitments. As a result, it is important to segregate the more serious issues that exist from ones that may be less disastrous. This will be done by having the analyst provide a rating for the severity of a problem. In order to assign a severity rating, you will consider the following attributes:

Attribute	Explanation
Frequency	The frequency of occurrence. Is it common or rare?
Impact	The effect of problem on task performance.
Persistence	Will users be able to overcome this problem with instruction, or will it trouble users repeatedly?

The rating to be assigned will be in the form of a rank level ranging from 0 to 4. The table below explains the meaning of each level:

Rank Level	Name	Explanation
0	Not a usability problem	The problem might be a coding related error.
1	Cosmetic	Problem is very insignificant and need not be fixed unless time is available on a project.
2	Minor	This problem exists but has low priority in getting fixed
3	Major	It is important to fix this problem, it is given high priority
4	Catastrophe	This problem has to be fixed immediately.

Analysis of Think Aloud Data:

- *Activity Theory based method of analysis*

The method of analysis that you will be asked to follow is modeled after the Scandinavian 'Activity Theory' concept. Explicit knowledge of activity theory is not required to perform the analysis. However some basic information is provided here.
An activity can be considered to be an act of transforming an object in order to achieve a particular outcome using some process. Analysis of the activity is basically an attempt to recognize the perceived goal of the subject and identifying reasons or explanations (based upon preexisting, cultural, environmental, contextual, etc. factors) as to why the individual actions that constitute the activity were performed. You will be provided with a data collection template that you will have to follow in order to perform the analysis.

- *Data Collection Template (DTC)*

The activity theory based method of analysis uses an activity as the unit of analysis. One activity might contain more than one verbalizations made and actions performed by the participant. The DTC is structured around the analysis of each activity. The activity itself will often encompass more than one action. An explanation of how to use the template is mentioned below:

Column 1

In order to use the template, you will begin by identifying an activity that is being performed. Write out the activity in the first column of the template in the form of a statement that will be situated within the overall context of the task.

Column 2

a> After that, you will address each action performed as part of that activity. List out the different actions that subject performs as part of the overall activity.

b> Once you have listed out the individual actions, you will be able to identify the perceived goal of the subject in performing the individual actions. The perceived goal is your opinion of the purpose of the user's actions.

c> Lastly you will list out the object, outcome and process used to achieve this perceived goal.

Column 3

You will provide a reason for why each action that was listed in column two was an attempt towards meeting the perceived goal. The reasons that you will list will take into account various cultural, environmental, contextual, etc. factors that might play a role in the participant's behavior. This reason can also be a verbalization made by the participant.

Column 4

You will classify each action that was identified as some type of CI.

Column 5

Here you will identify why the action was classified as a particular type of CI by indicating the relationship (followed/violated) to a known heuristic.

Column 6

You will then provide a recommendation to indicate how the CI can be used to improve the usability of the product.

Column 7

Finally, you will provide a severity rating to the issue identified.

Videos:

You have been provided with a video cassette containing the training tasks that we will go over today and 16 tasks that you will have to analyze. Following are a few things to keep in mind while performing the analysis:

1. These tasks have been recorded on the cassette in the order in which you are expected to perform the analysis. Please perform your analysis in this specific order.
2. You are only expected to provide analysis for the actions/verbalizations of the participant, ignore all comments made by the experimenter.
3. In case you feel that the participant is stuck (for e.g. repeatedly goes to the same spot to look for a feature) you can indicate the occurrence once in the analysis and mention that the participant does the same thing x number of times.
4. Each task is separated from the next by a placeholder that displays the upcoming task number (there are a total of 16 tasks).
5. While analyzing, please refer to each task by the task number and start a new task on a new sheet.
6. If you have any questions while performing the analysis, you can contact the experimenter by email at yogeshdb@vt.edu or yogeshb@kyocera-wireless.com or by phone at 540-818-5298 or 858-882-2705