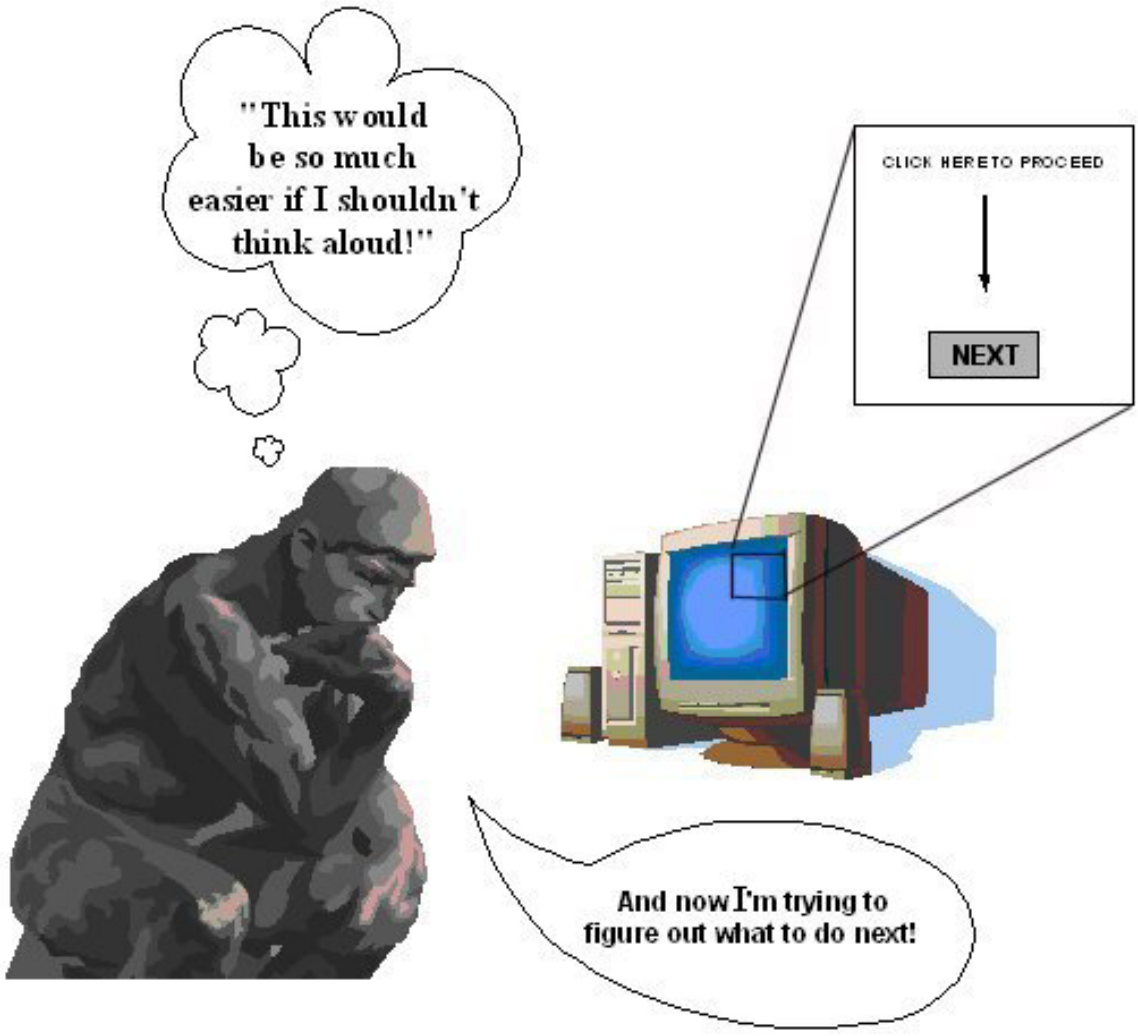
Do words get in the way of (better) usability?

A theoretical-conceptual analysis of the think-aloud method in relation to verbal overshadowing.



René Ginger-Mortensen

Thesis - March 2007
Department of Informatics
Copenhagen Business School (CBS)
Supervisor: Torkil Clemmensen

Preface

This thesis is written as the final paper in the completion of the Master Science program (cand.merc.dat) under the Department of Informatics at Copenhagen Business School, Denmark.

The thesis consists of a total 65 normal pages, and 74 physical pages.

A normal page is in the paper defined as consisting of 2275 strokes pr. page.
Figures are counted as 700 strokes.

The thesis has been supervised by: Torkil Clemmensen

Date:_____ Signature: _____
(Copenhagen) René Ginger-Mortensen

Abstract

The need for good usability is in today's world a parameter that can not be ignored or neglected. Whether it is the operating system on a computer, a web site on the World Wide Web or a custom made application to fulfil the needs of a small group of people, usability has become a parameter that can make or break the companies behind the software developed. One way of achieving good usability is by analysing the data gathered through tests involving the end user. One particular popular way this data is gathered, is by using the think-aloud method to test the usability and thereby getting information on eventual usability defects.

Relying on verbal data to locate usability defects is however not without problems. Advances in cognitive psychology on the phenomenon of verbal overshadowing, have shown how verbalisation in some cases can alter the behaviour and impair the performance of a person thinking aloud while performing other tasks. By the use of a theoretical-conceptual approach this thesis has investigates the possible impairment of usability due to test persons are being told to think-aloud during the usability test.

The result of the analysis, indicate that verbal overshadowing is a potential problem where the validity and reliability of the data produced by the think-aloud method, is depending on the verbal skills of the test subject (i.e. the person's vocabulary) in relation to the domains in which the usability test takes place, and in relation to the test person's perceptual expertise. If those evaluating the data produced by the think-aloud method, in order to locate usability issues, does not take into account that the issues located, in some instances is not caused by defects in the tested application, but instead a caused by the actual test procedure. Impaired usability will inevitably be the result. Based on the analyses it is furthermore hypothesises, that the validity and reliability of the data furthermore can be impaired if an application is relying on the user to solve or perform a task by getting to a solution through insight.

As a further result of the analysis, the hypothesis is being made, that selecting test persons with an impaired ability to think-aloud and perform other task at the same time, could lead to the development of more logic or intuitive application.

Do to lack of research on the subject of verbal overshadowing, within the field of usability, further research is however needed in order to verify or disclaim the different hypotheses postulated in this thesis. Practitioners of the think-aloud method is however encouraged

Resumé

Behovet for god brugervenlighed er i dagens verden en parameter der ikke kan ignoreres eller negligeres. Om det gælder operativ systemet på en computer, en hjemmeside på Internettet eller en fremstillet applikation, til at tilfredsstille behovet for en mindre gruppe mennesker, brugervenlighed er blevet en parameter der kan skabe enten succes eller fiasko for de virksomheder der står bag det udvikler software. En måde at opnå god brugervenlighed er ved analysering af data opsamlet ved test der involvere slutbrugeren. En specielt populær måde at indsamle denne data vedrørende brugervenligheden, er ved brug af tænke-højt metoden for dei igennem at få information om eventuelle ”defekter” i brugervenligheden.

At stole på verbal data til at lokalisere uregemæssighedder i brugervenligheden er dog ikke uden sine problemer. Fremskridt inden for kognitiv psykologi i forbindelse med fænomenet verbal overskygning, har vist at verbalisering i nogle tilfælde kan ændre adfærden og forringe præstationen hos personer der tænker højt under udførelsen af andre opgaver. Denne afhandling har ved brug af en teoretisk konceptuel tilgang undersøgt de mulige forringelser af brugervenlighed, der skyldes at testpersonen bliver bedt om at tænke højt under selve brugervenligheds testen.

Resultatet af analysen indikere at verbal overskygning er et potentielt problem, hvor validiteten og pålideligheden af de data der bliver frembragt ved hjælp af tænke-højt metoden, er afhængig af den enkelte testpersons verbale færdigheder (i form af ordforråd), set i relation til det domæne hvori testen foregår, og set i relation til test personens perceptuelle formåen. Hvis de personer der skal fortage evalueringen af brugervenligheden på baggrund af de data, der er produceret via tænke-højt metoden, ikke tager højde for, at årsagen til lokaliserede problemer, i nogle tilfælde ikke skal findes i den testede applikation, men i stedet skyldes selve test proceduren. Vil dette uundgåeligt føre til forringet brugervenlighed. På baggrund af analysen fremstilles der blandt andet den hypotese at der kan stilles spørgsmål til validiteten og pålideligheden af de data frembragt, hvor løsningen eller resultatet i applikationen, skal findes i form af et øjebliks ”indsigt” hos test personen.

Som et resultat af analysen, omstilles der yderlige den hypotese, at en udvælgelse af test personer med en forringes evne til at tænke højt og udefører andre opgaver på samme tid, ville kunne lede til udviklingen af mere logiske og intuitive applikationer.

Grundet manglende forskning af verbal overskygning i relation til brugervenlighed og brugervenligheds test, er der dog et behov for yderligere forskning, for enten at kunne af eller bekræfte de hypoteser der bliver frembragt i denne afhandling.

Table of contents

1 Introduction

As with all businesses the resources for research and development is not unlimited. The lack of unlimited resources is also the case within software development, where both time and money often is a limited resource, which has to be taken into consideration during the different steps of the development process. It is therefore important that the methods and procedures used give the best value for money, i.e. that the data and material produced is reliable and correct.

The need for accurate and reliable data is also the case within the field of software testing, whether it is functionality or usability that is being tested. If the test method used is insufficient, defects will stay undetected until later in the development face, where the cost of removing them goes up by a considerable margin. In the worst case they will stay undetected until the product is released, and although it is somewhat easier to update software by patches or upgrades, then it is to recall every car model due to defective brakes. It still takes up resources that could have been spend more wisely and in the end reduces profit. Much emphasis has therefore been put on finding methods and procedures that finds as many defect in the development process as early in the process as possible. Especially when it comes to defect in the source code that will prevent the software to work as it was intended.

In the ever ongoing pursued to produce new and better software, one of the key aspects is to produce software with a high emphasis on usability and for the interface to be “user friendly”. Although usability defects in user interfaces in comparison to code defect does not hinder the software from doing the task it was designed to do – it can very well hinder or diminish the user in executing the intended task. Testing usability is however not something that is as easy as testing whether or not a application runs or functions as intended.

One of the most widely used methods for developers to get usability data about there product, is to use the so called think-aloud method (Boren and Ramey 2000). In this test method, test person is encouraged to talk aloud during the test process and express there thoughts, as they use the software and its interface in a given context.

The method itself was first used in the field of psychology, in psychologist’s quest to understand the different cognitive processes in the human brain. But since software developers adapted the think-aloud method to usability testing purposes, its usage has deviated from its original form and background theory (Boren and Ramey 2000). There exists therefore no standard manual for how to use the method within the field of usability testing, and no uniform consensus that describes in which situations the method should be used and can be effective to use, and how the test session itself should be conducted in order to get the most reliable and valid data.

One essential point made by Boren and Ramey is that when the method is used by psychologist, the method is used to get information on the cognitive processes that are going on in the human brain. The different tasks a given test person is subjected to is known, there is nothing unknown about them, and they are not under scrutiny in any way. There is therefore only one unknown factor when the method is used to understand the human brain – the human brain itself!

This scenario is however reversed when it comes to the situation where the think-aloud method is used for usability testing purposes. When the think-aloud method is used in a usability context the task or the tools in the shape of an interface and application is considered the unknown factor. For a mathematician this leaves one big problem! Essentially what software developers is doing when applying the think-aloud method is that they are trying to find the result of one equation with two unknowns. Something that is rather difficult unless you, as software developers has done, presume that the human brain is in fact a known quantity, which does not have an effect on the output and data obtained during the test.

Is collected data reliable when using the Think-aloud method?

The question is therefore whether or not the presumption that the human brain has no influence on the data collected during the think-aloud test is correct?

Since software developers incorporated the think-aloud method into the development process, in order to get data concerning usability, its use and form has changed from what was originally its purpose within psychology. But what is just as important, software developers have neglected to take into account the latest research made within cognitive psychology, concerning the use of think-aloud and the different processes of verbalisation that goes on in the human brain.

Research made from the early 90's and on, within the field of cognitive psychology, shows that the very act of verbalisation can have an altering effect on people, this phenomenon is also known as ***verbal overshadowing***.

The term "verbal overshadowing" covers that the very act of verbalising, changes the way the human brain works, which in some instances can have an altering influence on the outcome of the mental process. Such as for example in the way things are being remembered, or peoples ability to solve certain kind of problems. The basis for assuming that there is no other variables that can affect the outcome of a usability test other then the system that are being tested, has therefore become very questionable.

In order to get reliable data using the think-aloud method and there by getting interfaces and software with better usability, it is therefore necessary to understand what verbal overshadowing is, what the

effect of verbal overshadowing is on the human mind and more importantly how this relates to the field of usability testing and how the think-aloud method is being used at present.

It has in other words become necessary to develop a new equation on what is going on in the human brain, so we end up with two equations with two unknowns that can be solved. The only problem is that very little research if nothing at all has been done on verbal overshadowing in relation to usability testing.

There is no question that the widespread use of the think-aloud method indicates that the method is useful and that system-developers feel that the think-aloud method gives the developers more data that is useful in the development process then where they just observing the test persons. But the new findings concerning the phenomenon of verbal overshadowing raises the question whether or not the data collected through the think-aloud method is reliable and gives the correct bases for defining usability defect in software or its interfaces.

A potential aspect of test persons increased difficulty to solve problems, because they are forced to think-aloud, could therefore be that a lower score for usability was found on a given user interface, then had the interface been tested by test persons not talking aloud. Are developers not aware of this potential problem in specific test scenarios, there are a potential risk of using recourses on improving (i.e. changing) an interface, which during normal use where no verbalisation is involved would otherwise be fine. The low usability score would instead have been the result of having test persons speaking aloud during the test.

Own approach

There is quite a lot of research made within the field of usability testing describing how the usability test should be carried out using the Think- aloud method, in terms of how the test person should be encourage to speak during the test, on the relationship between the test person and the person in charge of the test, and on the technical aspects of the recording equipment used, just to name a few. The common factor for this line of research is that it looks on the form of the think-aloud method. The introduction of verbal overshadowing brings a need to look on how the method is used, in the sense of being able to provide reliable data as a method. Not on the basis of how the method is used (i.e. how the test is carried out), but based on if it should be used at all, in relation to get liable usability data in a given test scenario.

One key element is therefore not so much the think-aloud test itself but the actual person performing the test, since it is the person performing the test that has the potential to be influenced by verbal overshadowing during the actual think-aloud test.

One way to approach the subject of a potential effect of verbal overshadowing when using the think-aloud method to test usability, would be to define one specific potential problem where verbal overshadowing might have an effect on the test result, and then run a series of test scenarios with a silent control group in order to see whether or not the actual verbalisation had an effect on the test results concerning usability. Following this approach would however have the effect, that only the one specific potential problem would be investigated. It has so fare not been possible to find research on verbal overshadowing within usability testing and in respect to the think-aloud method. There is therefore no indication of research trying to provide an overall picture of the potential problems, from a theoretical side, and there is therefore no existing "big picture" that could give clarification to which problems that could potentially exist when verbalising during the test of an application / interface.

Researching this problem using a theoretical-conceptual approach, as has been chosen in this thesis, will in contrast have the potential to cover as many aspects and angles as possible and thereby eliminating, that resources put in one specific empirical experiment, should obscure the "painting" of a "bigger picture" then is seen at present.

Without a "big picture" of how verbal overshadowing could have an influence on the data produced by the think-aloud method, it would not be possible to know if the one specific problem researched in a large empirical study, was the only problem or just one of many. The emphasis will therefore be on providing an analysis on known experiments on verbal overshadowing and the background theory used in these, in order to clarify the potential problem if one exists and specify the arguments on what experiments or knowledge that is needed, in order to eliminate or prove the hypotheses made in this thesis.

Problem

Because of resent research made within cognitive psychology about the effect on verbalisation of thoughts, and the negative effect hereof known as verbal overshadowing. The overall problem under study in this thesis will therefore be the following:

1. Is verbal overshadowing a problem when using the think-aloud method within usability testing?

This problem subsequent raises some subsequent questions which answers could clarify the overall question:

1.1 How does literature prescribe that the think-aloud method is being used and is verbal overshadowing a problem, in the way the think-aloud method is being prescribe used at present?

1.2 Should verbal overshadowing be considered when choosing users for think-aloud testing?

1.3 Can the phenomenon of verbal overshadowing be used in a way that would make the think-aloud method more effective?

Results

As a result of the analysis of the formulated problem above, I will argue that there is a great need for further research on how verbalisation during a think-aloud session affects the test person in relation to the persons ability to test the interface and application. There are as I will argue later a great potential for the generation of flawed data and subsequently misinterpretation by the ones evaluating the gathered data. This potential problem are for a large part, the cause of a non-existing uniform description of when the think-aloud method should be used and when the method should be used with caution.

Although verbal overshadowing in some cases will have no influence on a person, I will state the hypothesis that selecting a user, based on the persons verbal skills concerning the domain in which the application or interface operates, will be relevant in order to minimize the risk of verbal overshadowing and a followed impairment of the data gathered in the think-aloud test, which subsequently will lead to impaired usability. But as I will show, research to verify or reject this hypothesis is needed.

The result of the analysis will also hypothesise that individual differences in peoples ability to perform tasks while thinking aloud, should be considered a relevant factor when choosing test persons for a think-aloud session. Regardless of the type of the task the test person is confronted with during the course of the test, choosing the right user will potentially deliver more reliable data and thereby enhance the usability of the interfaces and applications tested.

The structure of the remaining paper

This paper will from here on consist of a section explaining the reasons for the chosen method for the paper. This will be followed by an introduction and description of the different elements of this thesis. There will be an introduction and discussion of what usability is, the think-aloud method and its original purpose in cognitive psychology, along with a brief introduction to verbal overshadowing, ending with a discussion on how the think-aloud method is being prescribed used by practitioners and the advantages and disadvantages involved with the method as they see it.

The more elaborate description of the implication of verbal overshadowing in relation to usability testing will be made in the subsequent analysis chapter. In this the implication and aspects of first vocabulary or verbal skills will be made. This will then be followed by an analysis of insight problems, both on a general level but also by looking at the individual aspects which was also analysed in relation to verbal skills.

The analysis (chapter 4) will be followed by the erection of hypotheses on how verbal overshadowing will affect the validity and reliability of the data gathered by the think-aloud method. This will be followed by a discussion of how these hypotheses and the finding behind them, in relates to the theory behind the think-aloud method at present.

Finally a conclusion will be made in relation to the overall problem raised within this thesis, and the liability of the result will be discussed, along with recommendation to practitioners of the think-aloud method and recommendation to what should be researched as a result of the finding.

2 Methodology

There is a tradition on CBS (Copenhagen Business School) and in particular under the Department of Informatics at the economical faculty, that the larger papers made by students are very business-oriented. This means that they for the most part are based on or involves, a large empirical study, either with the starting point in a hypothetical assignment, or as part as a consult assignment with a Danish company. The focus point for these papers is thereby more on the achieved result, supported by the empirical work done in the process, then the theories involved. The theories used could better be described as tools or framework in which the work is done, rather then it is the theories themselves that is the focus point for these papers.

Because the main focus-point for this thesis, in contrast to an ordinary paper under CBS, is the background theory on which the think-aloud method or procedure is based upon, and not the use of the method to collect empirical data, it was chosen to take a theoretical-conceptual approach in order to cover as many aspects and angles as possible and thereby eliminating, that resources put in one specific empirical experiment, should obscure the big picture.

An alternative way of approaching verbal overshadowing within the field of usability testing, instead of using the theoretical-conceptual approach, would have been by defining one specific possible problem where verbal overshadowing might have an effect on the test result, and then trying to verify or disclaim this potential problem, by running a series of test scenarios along with a silent control group, in order to see whether or not the actual verbalisation had an effect on the test results concerning usability. Following this approach would however as mentioned in the introduction have the effect, that only the specific type potential problem set up would be investigated. And so fare it has not been possible to find any research on verbal overshadowing within usability testing and the think-aloud method.

There is therefore no one, to my knowledge, who has tried to provide an overall picture of the potential problems, or for that matter argued against them, from a theoretical point of view. There is therefore no existing “big picture” that could give means to an overall framework in which issues with verbal overshadowing could be placed. Conducting an experiment on one specific potential problem in relation to the phenomenon of verbal overshadowing without having an overall “frame” to place it in, would possible lead to a specific result, but the result’s implications to usability and other aspects would be unknown since the knowledge or awareness of these other areas would be non-existing.

The emphasis is therefore on providing analysis on found research, there experiments and especially the background theory used in the found research, in order to shed some light to the problem in hand and thereby hopefully be able to specify the areas in which either more research is needed in order to

eliminate or prove the hypotheses made in this thesis, or where it as this point would be possible to answer the questions raised in this thesis.

The focus of this thesis will therefore not be on an end-result produced by applying different methods and theories as tools or frameworks, but on the very tools and framework used within usability testing. More specifically the think-aloud method and the underlying theory, and how the use and theory relates to the phenomenon of verbal overshadowing.

Data collection

Due to the choice of method in the thesis, there is no conventional empirical data collection. The data in this thesis consist purely of the articles found on primarily usability, the think-aloud method and verbal overshadowing. From at starting point I was so lucky to have access to some articles and material in both the domains of thinking aloud (usability testing) and verbal overshadowing (cognitive psychology). From there on other relevant background articles and material was therefore found as a result of them being referenced to in these starting articles or others found later. However newer articles were found using services accessible over the Internet. This involved services such as Google, ISI Web of Knowledge, ACM and e-journals, in order to find newer research on the topics then those obtained at the beginning. And to locate other research that could describe other areas or aspects of how verbal overshadowing could or would be an influence when using the think-aloud method. The literature referenced in this thesis has all been chosen because it brings what I see as relevant knowledge to the analysis and problems discussed in this thesis.

The obvious problem with this approach is that the data found this way is very depended on the detective work performed. Because of the way relevant articles and other material was found, that could give possible clues to the problem in hand, and this approach could be describes as somewhat accidental in what clues that has been found to this “puzzle”. There is logically a potential chance that some of the conclusions or hypotheses made within this thesis would be different, and others would not have been made, was other materiel found in the process of uncovering the different aspects of the problem at hand. There have unfortunately also been instances where literature referenced in found articles as been impossible to obtain. It is however very difficult to suggest another way of finding or locating the material, where the material found would not be dependent of how well the research was performed.

3 Usability, the think-aloud method and Verbal overshadowing in general.

As stated in the introduction the think-aloud method is widely used as a way to test the usability of user interfaces and there underlying systems. In order to get an understanding of the possible implications by verbal overshadowing on the data gathered during a think-aloud test, it is not only necessary to get an understanding for what is know about verbal overshadowing. It is just as important to get an overview of the domain in which the think-aloud method is used. To do this it is important to know what usability is, and more specifically what defines usability and in what term is it measured.

3.1 Usability

When talking about usability we are talking about the interaction between the user and the user interface not about the product it self. According to Barnum (Barnum 2002) usability therefore has nothing to do with:

- Quality assurance
- Zero defects
- Utility of design features
- Intrinsic in products

The reason for this is that these aspects all relates to the application, and not about the usability.

But finding one clear and unanimous description of what usability is, is more then difficult. Some have even compared it to *"like trying to nail a blob of Jell-O to the wall"* (Barnum 2002, page 6). Although there are similarities in the different kind of descriptions of usability they are not quite alike. Barnum herself describes 3 different examples on how usability is being defined (Barnum 2002, page 6):

ISO 9241-11: "The extend to which a product can be used by specified users to achieve specified goals in a specified context of use with *effectiveness, efficiency*, and *satisfaction*"

Dumas and Redish: "Usability means that people who use the product can do so quickly and easily to accomplish their own tasks"

J. Nielsen: "The measure of quality of the user experience when interacting with something – whether a web site, a traditional software application, or any other device the user can operate in some way or another"

Jacob Nielsen extends the description by saying that the following characteristics applies to usability: (Nielsen 1993, page 26)

- *Learnability*. The system should be easy to learn so that the user can rapidly start doing some work.
- *Efficiency*. The system should be efficient to use, so that once it is learned, the user can achieve a high level of productivity
- *Memorability*. The system should be easy to remember, so that the casual user is able to return to the system after a time and not have to learn it all over again.
- *Errors*. The system should have a low error rate, so that users make few errors and can easily recover from them.
- *Satisfaction*. The system should be pleasant to use, so that users are subjectively satisfied when using it; they like it.

But Nielsen is not the only one to specify the different parts of usability. Patrick W. Jordan (Jordan 1998, page 11-16) describes the five component of usability based on the ISO definition of usability as:

- *Guessability*. The effectiveness, efficiency and satisfaction with which specified users can complete specified tasks with a particular product for the first time.
- *Learnability*. The effectiveness, efficiency and satisfaction with which specified users can achieve a competent level of performance on specified tasks with a product, having already completed those tasks once previously.
- *Experiences user performance*. The effectiveness, efficiency and satisfaction with which specified experienced users can achieve specified tasks with a particular product.
- *System Potential*. The optimum level of effectiveness, efficiency and satisfaction with which it would be possible to complete specified tasks with a product.
- *Re-usability*. The effectiveness, efficiency and satisfaction with which specified users can achieve specified tasks with a particular product after a comparatively long period away from these tasks.

Again there are similarity on some of the components or characteristics between Jordan and Nielsen. Nielsen's "Memorability" is for example comparable to Jordan's "Re-usability" although they use somewhat different names to define it.

But although the different definitions of what usability is all have there similarities, they are still only describing usability in measurable terms, or in terms that can be converted into measurable objectives. “Quickly and easily” as Dumas and Redish describes usability is not very measurable until you define what quickly and easily is. None of the definitions are however saying anything about how usability is achieved in the design of the interface or the system that is being developed.

Correlation between the different elements of usability

Since user’s perspective of usability is very subjective achieving one element of usability does not guarantee that you achieved the others. Research based on ISO standards definition of usability has showed that there is not naturally a given correlation between the different characteristics of usability (Frøkjær 2000). Their research shows that the three aspects of usability, defined by the ISO standard (effectiveness, efficiency, and satisfaction) are not necessarily related in such a manor that you can predict one from the result of the other. This means that if two different applications are compared in a usability test, in order to find the best suited application which had the best usability. Then the application that has the best score in efficiency is not necessarily the one that has the best score when it comes to satisfaction or visa versa. Saying something about the effectiveness based solely on how the application scores in efficiency and visa versa is not possible either unless domain specific studies suggest otherwise.

The fact that the different elements in many situations does not correlate also means that in order to thoroughly test usability all the different elements has to be tested not only one or two elements.

The mental model of usability

The previous described definitions might say something about how usability can be described. Another way at looking at usability is how it is being perceived and what is going on the mind of a user while the person uses and interacts with a given interface and application.

Just as a application and interface is build according to a certain design model with functions and features designed to work in certain ways according to the design specifications, the mental model (Kuniavsky 2003) or cognitive model is in short the way the user thinks the interface or system works. The better a user’s mental model matches the actual design model of the system the higher the usability of the system. If a user instinctively knows either from experience or by the way the interface is build, that A happens when the person presses button B. Or in order to complete task C the person has to pres button D. Then there is a good chance that some of the characteristics of usability apply to that application and interface.

Allow me to give an example to illustrate what typically will happen when the mental model of the user does not match the design model of a application, with a personal case which I stumbled over the other day as I was editing some recorded video.

Mental model vs. design model mismatch

In the software I use to transfer video recorded with my camcorder to my PC and edit the video, there is a functionality that can auto-detect the different scenes recorded.

At the time where this went on I had very little experience using the software – and I had at that point never tried that specific functionality. Nor had I read any manual or "help" when I used the software.

After the software had detected all the scenes in the clip I had selected – all the scenes appeared on the right side of the window (see Figure 1)

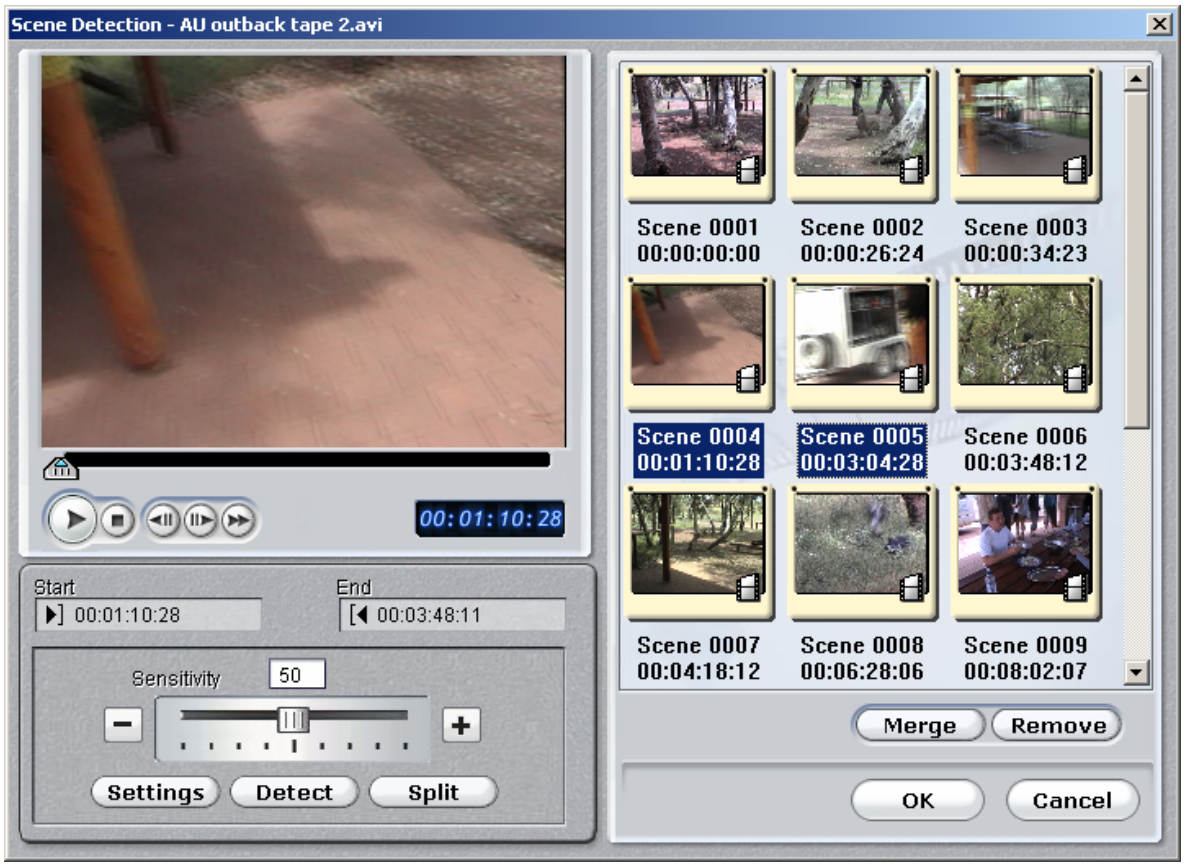


Figure 1 – Scene detection in Power director.

Since the software has a tendency to detect more scenes then when the record button has been pressed in order to start and stop a recording (on the actual video tape) you can merge the scenes together by selecting the scenes you want to be one scene instead of several small scenes.

During this process there was one scene I did not want – so instead I pressed the remove button and clicked “Yes” on the “mandatory” are you sure message. To my surprise the result was that all the scenes disappeared.

After running the detection ones again (took about 30 min) I again marked the scene I wanted to delete – since my first thought was that perhaps I had not selected a scene at all, so it deleted all of them instead, but again - same result. By that time something told me that the program might not work as I thought it would. So the third time I actually took my time and read the “mandatory” are you sure message (see Figure 2)

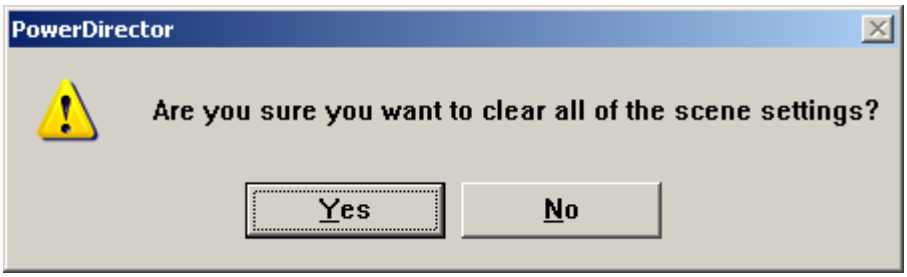


Figure 2 – Warning message when wanting to delete all detected scenes

As you can see there is one specific word that is worth noticing here – the word “all”.

Because the merge button worked on an individual level in the sense that it only merged the scenes that was selected, I automatically assumed that the same was the case for the remove button. This was however not the case. Instead I wasted the better part of an hour.

What I found out rather quickly afterwards is that to remove a scene you have to right click on the scene and select “Remove scene” (see Figure 3)

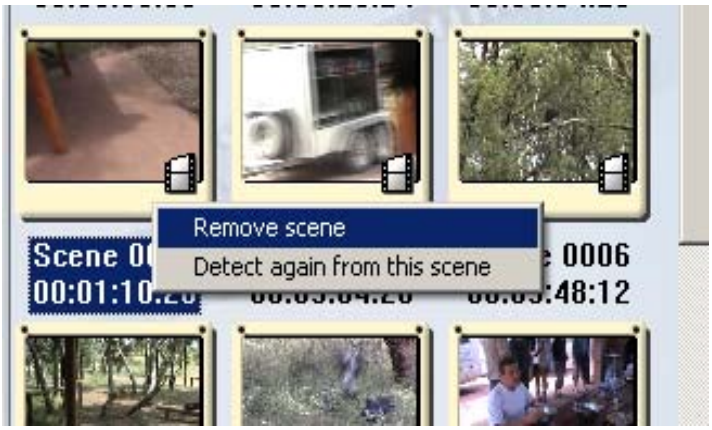


Figure 3 – How to remove only one scene at a time by right clicking with the mouse on the scene.

I have no doubt that the software works at it was designed, and that the “remove button” purpose is to remove all the scenes that has been detected. But my understanding of how the program worked (my mental model) did not match how the program was design and build.

In my case this would not have been a problem if the button had said “remove all” instead of just “remove”. That little change would automatically have let me to investigate other means of just removing one scene instead of them all. For somebody else this issue I have described above might not even have been a problem, for them the “right” way of removing one scene would have been the obvious one. Others might even still have encountered the same problem even if the button had said “remove all”.

What makes a mental model?

The explanation for why people can have different perceptions of how interfaces and systems work shall be seen in the mental “baggage” the user brings with him. The mental model, the user is making of a given system, depends on a lot of different factors (Jordan 1998) such as

- Experience
- Domain knowledge
- Cultural background
- Age and gender

All these are factors that have a large influence on what is stored in our minds as memories, and therefore a large influence on how we as people perceive things, interoperate situations we are in, and make assumption about how things work.

Domain knowledge and experience are vital factors when it comes to how tasks should be solved using a given interfaces and systems. In systems such as certain types of computer games, the solutions to a problem might not be obvious by deliberate reasons and situations where the user is required to rely on intuition or on getting insight to the solution, by piecing different kind of obtained knowledge together in order to complete the task, will occur.

If a user’s mental model corresponds to the systems design model he is using, the user will be able to use the system without much training. A users ability to change his or hers mental model so that it corresponds to the systems design model will therefore be a good indicator of how easy the system is to learn. The essential part of this is however, that although it is the user’s mental model that changes

during use of an interface or system during the learning face, failure to understand the systems design model is not the users fault.

The point about usability

The essential part of usability is to understand that that usability is more then one thing. But since there exist several different definitions of what usability is and although some, like the ISO definition of usability, seems to be more widely used then others, it is highly necessary to state how one defines usability, especially when it comes to the test part of a system.

Just stating that you have completed "a usability test" using the think-aloud method and thereby tested the usability of a system, with the use of several different users, does not really say much unless you describe how you define usability, and furthermore describe which element of usability you have tested for. The only thing it states, is that the reader can be sure of, is that it is not m*emorability* that has been tested since m*emorability* obviously needs a second test where the user tries the interface again after a given period.

As shown earlier, usability consists of many different things, dependent on who you ask, different characteristics will be part of the description on what defines usability. However none of these characteristic elements by themselves explains what usability is from a cognitive aspect. Instead they are the measurable end result of determining if the user's mental model corresponds to the design model of the interface and application with which the user is interacting with.

The diverse description of usability and the many aspects to usability along with the cognitive understanding of usability in form of the mental model, is one element that can be credited as one significant element to the diverse use of the think-aloud method. But before we take a look at how the think-aloud method is used to test usability it is first necessary to know how the method started and what its original purpose was. Knowing where the think-aloud method came from and the research made within cognitive physiology will give fundamental clues to the dilemma about having people talking aloud while performing different tasks.

3.2 The original use of the think-aloud method

Having people thinking out loud, in the form of verbal protocols have been a way for cognitive psychologist to attempt to understand and trace the processes in the human brain since the end of the 19th century. The big breakthrough came however first in 1980 when Ericsson and Simon published *Protocol Analysis: Verbal reports as data*, where they proposed a theory on verbal protocols or thinking aloud. The bases for the proposed think-aloud method was based on a model consisting of 3 levels of verbalisation.

Level 1 verbalisations: This type of verbalisation requires no processing in order to be verbalised. The information obtained can directly be verbalised the same way as it was perceived by the brain and is a valid account of the information that is stored in the short term memory. An example of this could be the recitation of numbers while a math equation is being solved.

Level 2 verbalisations: This type of verbalisation involves description or explanation of what it currently held in memory[1]. In order to be verbalised the content has to be decoded or transformed into a form that can be verbalised. An example of such content could be information about odours. The only cognitive processes between the information held in short term memory and the actual verbalisation is the actual decoding process.

Level 3 verbalisations: Third level verbalisation involves the explanation of thought processes or thoughts, such as ideas or hypotheses. Additional cognitive processes are therefore needed other then those required for task performance and verbalisation, in order to link the present information together with previous obtained information or earlier thoughts (i.e. accessing information now stored in long term memory).

When conducting a think-aloud test only level 1 and 2 verbalisation would according to Ericsson and Simon be a reliable source of data, because these types of verbalisations according to their hypothesis would not change the cognitive processes in relation to the task at hand.

Data gathered through third level verbalisation would in contrast not be suitable to reveal the cognitive processes. This is because third level verbalisation forces test persons to alter their thought sequences, by generating non-task oriented cognitive processes or tapping in to the long term memory, in order for them to verbalise the information present (what they are thinking). The verbalisation will therefore ultimately change the way they would perform a task in contrast to where they did not verbally explaining what they where doing.

[1] Short term memory

Although level 2 verbalisation generates new processes, Ericsson and Simon argued that these would not change the structure of task performed:

"Since explication or recoding requires processing time for the subject but does not replace other processing involved in the task performance, a subject who is verbalizing at this second level can be expected to take more time for the task than one who is not verbalizing. However, we would hypothesize that such recording does not change the structure of the process for performing the main task" (Ericsson and Simon 1993, page 79)

Keeping to level 1 and level 2 verbalisation

In order to learn more about the cognitive processes in the human mind, Ericsson and Simon argued that level 3 verbalisation should be avoided and only verbalisation at level 1 and 2 should be used as data. The technique to avoid level 3 verbalisation and focus the test person to generating level 1 or level 2 verbalisation is the use of non-directive reminders like "keep talking" and in general avoiding other form of interaction like comments and question to the test person. Probing for information will according to Ericsson and Simon, only lead to level 3 verbalisation, which is why this should be avoided.

What is important to remember as described in the introduction is that the purpose of the think-aloud methods is to produce verbal data, psychologist then can analyse in order to get a better understand of the cognitive processes in the human mind, a purpose it still fulfils today. But the soul focus point of the method is the test person, the think-aloud method and the tasks the test person is given is just tools involved in the process of investigating the unknown i.e. the cognitive processes in the human mind.

In 1982 Clayton Lewis at IBM's Thomas J. Watson research Center (Lewis C. 1982) described how the method could be used to study the cognitive problems that people have in learning to use computer systems in his paper on: how to use the "thinking-aloud" method in Cognitive Interface Design. This marked the beginning of the think-aloud methods use in usability testing. But as chapter 3.4 will show the direction the method has taken, has evolved into a method quite fare from the original description by Ericsson and Simon on how to generate reliable verbal data.

An important point that is worth keeping in mind is that the method itself whether it is being used within cognitive psychology or usability testing is that the method itself only produces data. In order for the psychologist to understand the processes in a test persons mind, the data has to be analyses and interpreted. The same is the case within usability testing, in order for the software developer to find eventual usability errors, the data has to be interpreted and analysed in order to locate eventual problems located in the test.

3.3 Verbal overshadowing

Before we take a closer look at usability testing and the think-aloud method, in order to analyse if verbal overshadowing has any influence on the data generated, we first have to get familiar with what verbal overshadowing actually is.

The term verbal overshadowing is used in situations where verbalisation, either in verbal form or in writing, is having an impairing effect on task performance. The term "verbal overshadowing" was first suggested in a study of face recognition, done by Jonathan W. Schooler and Tonya Y. Engstler-Schooler (Schooler, Engstler-Schooler 1990). In the study Schooler and Engstler-Schooler found that verbally describing a face would diminish the possibility of recognising the face again at a later point. Verbal overshadowing has also been found in other domains of non-verbal knowledge, such as taste memory, map memory and insight problem solving (Ryan & Schooler 1998).

But in order to understand how this phenomenon relates itself to usability testing and the think aloud method, a more deliberate description of verbal overshadowing is needed.

Verbal overshadowing and its influence on memory

One aspect of verbal overshadowing is, that if a person is asked to describe something that has previously been perceived, verbalisation can alter the obtained memories and diminish the person's ability to recognise the describe memory shortly after the verbalisation. This could be an object like a face, scent or colours.

However as shown by Kimberly Finger (Finger 2002) this effect will disappear over time, and could be prevented if a non-verbal task, like listening to music, was performed before the person was asked to recognise what was perceived. This effect of verbal overshadowing is therefore only temporarily.

This is however not the case if people are verbally describing what they perceive in a given situation i.e. seeing, tasting, hearing etc. In that case people are only able to remember things as well as they are able to describe them verbally during the perception period.

In both cases the effect of verbal overshadowing sets in when people's ability to perceive what they for example are looking at, is greater then there ability to describe it verbally. It has been shown in a series of articles (Fallshore and Schooler (1995), Melcher and Schooler (1996), Ryan & Schooler (1998) and Meissner, C.A et.al (2001)) how a disparity in people's perceptual expertise and their verbal expertise can predict the appearance of verbal overshadowing.

The situation can be illustration as seen in Figure 4. When what is perceived is being verbalised during perceptions (in this example a car), a person is only able to remember what was being verbalised as seen in situation 3.

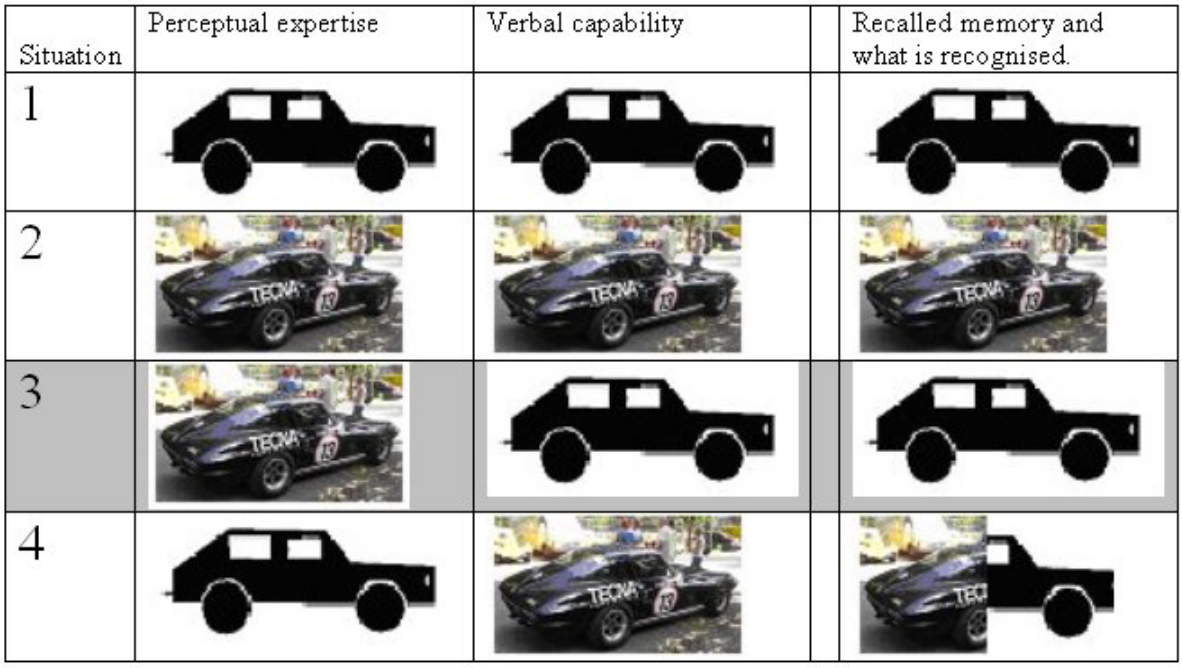


Figure 4 - The effect of verbal overshadowing on memory

Vocabulary within a given domain has is therefore a key factor in order to predict if people are subject to be influenced by verbal overshadowing.

Verbal overshadowing and solving insight problems

Another side of verbal overshadowing is how verbalisations affect the solving of insight problems.

Schooler et al (1993) discovered that performance dropped, when people where asked to think-aloud in test situations, while they where trying to solve so called insight problems. I.e. where the solution is reached in what could be described as a sudden moment of clarity or a eureka-moment. (A more comprehensive description of insight will be made later during the analysis).

What segregate these findings, from what Ericsson and Simon argued about the longer time on task regarding level 2 verbalisation, is that the found impaired performance is not a result of longer salvation time. The impairment is a result of people not being able to solve the task at all.

"There was no evidence in any of the four experiments that the verbalization increased the time taken to solve the insight problems. It appears that the case of insight problems, you either get them or you do not, and if you are verbalizing you are simply less likely to get them" (Schooler et al. 1993, page 178)

They furthermore found that there was a different in how many that could solve the different insight problems used in there experiment. The insight problems used could therefore be ranged according to difficulty.

To answer the overall question of whether or not the aspects of verbal overshadowing, as just described, is a problem when using the think-aloud method and therefore could be a source of error, in regards to the liability of the usability data collected, it is however necessary to look at how the think-aloud method is being prescribed used, in order to see if the effects of verbal overshadowing applies to any of these situations. A closer look at the advantages and disadvantages being described could also reveal if the effects of verbal overshadowing, was something that indirectly was being dealt whit.

3.4 Usability testing and the think-aloud method

One should initially think that it would be a rather simple task to see if verbal overshadowing would be a problem when using the think-aloud method. All you needed to do was to find out what the prescribed method for using the think-aloud method was, and then see if that was in conflict with the findings made concerning verbal overshadowing. But as the following will show, to do so is not such a clear cut case as one would have hoped it to be.

Just as difficult it is to find a common opinion on how to describe and define usability, just as difficult is it to find a theoretical basis for how and when to use the think-aloud method as a method to test usability. The task is not made any easier since the method is being referred to differently different places in the litterateur.

Looking at litterateur on usability testing and on the think-aloud method some initial confusion arises. When talking about a usability test, do one automatically talk about the think-aloud method, or is the thinking aloud just part of a series of methods or a special procedure within the field of usability testing?

According to for example Jacobsen (Jacobsen 1999) it is all in one, so when you are talking about usability test you would also be talking about the think-aloud method. The think-aloud method is however just one of many usability evaluation methods such as cognitive walkthrough, focus group, expert reviews, etc. But one can be very much in doubt when you look at authors like Carol M. Barnum (Barnum 2002) and Jeffrey Rubin (Rubin 1994) who combined have written over 700 pages on usability testing and thinking aloud takes up a total amount of approximately 7 out of the 700+ pages.

What the think-aloud method is good at

There seems to be a consensus in the literature on the fact that the think-aloud method has proven itself as a method that is very useable in providing usability data on a system or interface, especially in the early stages of development, where it is used in exploratory testing as stated by Barnum (Barnum 2002, page 235). Used during the development of prototypes the method can effectively give data regarding usability issues to the developers. It can furthermore be done with very few test persons as described by Nielsen (J. Nielsen 1994), and even without a big elaborate setup (J. Nielsen 1989) witch means it gives great value for money on uncovering usability issues in relation to the resources spend, early in the development face. The fact that the think-aloud method is a very usable tool can also be seen in the wide spread use describes by Boren & Ramey (Boren & Ramey 2000).

What is currently being research when it comes to the thinking aloud method.

The problem about the think-aloud method seen in relation to the potential problem with verbal overshadowing as discussed in this thesis, is however that there does not seem to be any literature, that spends much time on discussing when the method should be used and when it should not be used, in order for the data to be reliable and in synchronisation with a "real world use". As I will discuss later in this thesis there is very little said on when the think-aloud method should be used and for exactly what purposes. Even more interestingly is the lack of "warnings" and when not to use the method and just as important why.

One author that gets close to discussing something that could relate itself to verbal overshadowing and the problems about the reliability of the data provided in contrast to Ericsson and Simon 3 level verbalisation is Sally Abolrous (Abolrous 2001) when she looks at how probing effects the data provided by verbal reports[2]. (A link between Ericsson and Simon's three levels of verbalisation and verbal overshadowing will be discussed later in chapter 6).

The current discussion concerning the think-aloud method and general litterateur on the method has as far as I can tell a focus point on the test setup (i.e. use of video etc.) and how the different persons involved in the test should behave in order to get the best data possible. This could for example be how to get the test person to keep speaking aloud and how the person "interviewing" the test person should behave, in order not to influence the way the test person would solve a problem or perform at task. The wide area of approaches that has evolved within usability testing, on how the method should be carried out, so that the best data was gathered from the test person, has according to Boren & Ramey's raised the need for "getting back to basics" and adjust the theoretical approach in thinking aloud spearheaded by Ericsson and Simon in their "Verbal reports as data"[3]. This is needed so the conditions and needs of usability testing are being reflected in the theory. This should be done based on additional research, either as a supplement to the theories by Ericsson and Simon or as a replacement.

What to test using the think-aloud method!

Just as there are a multitude of different descriptions on what usability is, the definition of what usability testing using thinking aloud is, just as multifarious and diverse. One reason, as discussed with what defines usability, is that usability, in comparison to defects, can be a very subjective thing that will vary from user to user. Some users might find a particular system "satisfying" some does not understand the logic behind it since they are used to something else, for example windows vs. Macintosh or Linux.

[2] Much of her discussion is based on M. T. Boren's PhD dissertation – a paper that I unfortunately have not been able to acquire!

[3] Ericsson and Simon 1984 / 1993

The second reason is that how you define usability will neutrally define what it is you are testing for and therefore control how you would use the think-aloud method. If one defines usability in purely measurable term, then that is what you use the method for. If you on the other hand define usability by how the user understands the interface or application by there mental model, then you would use the think-aloud method to test or find out what that model would be like, and thereby discover usability defect in the form of inconsistent between the design model and the mental model.

So what does it mean to test for usability and how is it defined in the literature?

According to Barnum (Barnum 2002) there seem to be a general consensus on how to use the term usability testing since Barnum chose to define it the same way as Rubin (Rubin 1994), and Dumas and Redish in that

"Usability testing means the process that involves live feedback from actual users performing real tasks" (Barnum 2002 - page 9).

Rubin (Rubin 1994) defines 4 types of test with different purposes which involves usability testing (Rubin 1994 page 30-46), which according to Jacobsen (Jacobsen 1999) would involve the use of thinking-aloud.

Exploratory Test: Exploratory test are typically used in the beginning of the product development cycle in order to explore or examine the initial design models or concept in relation to the users mental model of the product.[4]

Assessment Test: According to Rubin, this type of test is conducted the most. These are conducted early or midway into the production cycle, and the object of these test are to expand the findings of the exploratory test by evaluating the usability of lower-level operations and aspects of the product.

Validation Test: This form of test is conducted late in the production cycle and is used to test how the product (interface and/or application) scores performance vice (i.e. *efficiency, effectiveness, learnability, satisfaction* etc.). Either by the performance requirements set in relation to a given project or in relation to predetermined benchmarks or competitive applications. (Although Rubin acknowledge that interaction whit the user in this type of test is very minimal, he does not say anything about the user not being able or asked to speak aloud at the same time)

[4] Rubin also refers to the user's mental model as the conceptual model.

Comparison Test: As the name states the test is used to compare two or more alternatives, whether it being to alternative designs or 2 competing products against each other, so see how they each score in relation to the usability specifications set for the interfaces / applications.

The Exploratory test form is also described by Mike Kuniavsky (Kuniavsky 2003), and again to map the mental model of the user and to understand what the user thinks of the interface. Kuniavsky does this by a setup that requires a propping like interview during the test, in order to get as much information by the user on what the user think of the interface. This is done by asking a lot of "Why" and "What" questions to what the user would do and why the user did it and so on[5].

Barnum (Barnum 2002) instead chooses to define what usability testing is NOT

- *Function testing.* Verifies that users are able to perform certain tasks
- *Reliability testing.* Verifies that the product performs as designed
- *Validation testing*. Verifies that the product performs without errors or 'bugs'.

These definitions by Barnum on what usability testing is not are however in my mind somewhat problematic. But because a further discussion on how Barnum defines test, for this thesis is irrelevant, such a discussion will not be made here.

In relation to the initial thought that inspired the problem of this thesis, it is therefore interesting to see if the potential problems with verbal overshadowing might be dealt with beforehand by a precise description on the different situations where the data collected with the think-aloud method should be dealt with caution since there might be a problem relying on the data.

Current advances / disadvantages with speaking aloud

However - because so little time is spend in thinking aloud, very little time is also spend on given specified description on the Method in relation to the disadvantages and advantages given by different authors. Barnum (Barnum 2002) and Rubin (Rubin 1994) each spends over 300 pages talking about usability testing but combined they only have about 7 pages on thinking aloud. Off those 7 pages less then 1 page is about the advantages and disadvantages of using the think-aloud procedure within usability testing. A third author Jordan (Jordan 1998) who discusses usability in general and among others describes the different methods to evaluate usability through different methods, has about 1½ page on thinking aloud, where 1 page is on advantages and disadvantages.

What in this respect is interesting to look at is the sources, or in some instances the lack of sources, they use for describing the think-aloud method and its potential and problems, since this would have

[5] In relation to Ericsson and Simon's model on verbalisation this would be a clear cut case of third level verbalisation

been useful in locating possible new theories for the use of thinking aloud, in relation verbal overshadowing.

Rubin list the following advantages and disadvantages (Rubin 1994, page 218):

Advantages:

- "You are able to capture preference and performance information simultaneously, rather than having to remember to ask questions about preference later."
- "The technique can help some participants to focus and concentrate. They fall into a rhythm of working and speaking to you throughout the test"
- "You are constantly receiving early clues about misconceptions and confusion before they manifest as incorrect behaviours. These early clues help you to anticipate and trace the source of problems more easily"

Disadvantages:

- "Some participants find the technique unnatural and distracting since thinking aloud is very different from their own learning style. If a participant is not an "analytical" learner, he or she may feel severely inhibited."
- "Thinking aloud slows the thought process, thus increasing mindfulness. Normally, this is a good effect, but in this case it can prevent errors that otherwise might have occurred in the actual workplace. Ideally, you would like your participants to pay neither more nor less attention to the task at hand than they normally would."
- "Regardless of learning styles, preferences, and other considerations, it is just plain exhausting to verbalize one's thought process for two or three hours."

What Rubin uses as a source to these advantages/disadvantages is a good question since there is no direct source reference. But looking at the reference list reveals on potential source "Ericsson 1980" from which the second disadvantage could have been derived or used as input. Going through the reference list only reveals one other potential source Rubin's source "69" (Knox, S.T. et.al 1989). But although this might have been used as input to the book, it reveals no specific clue to where the advantages and disadvantages might have come from[6].

Jordan is the second of the three authors that describes different advantages and disadvantages with the think-aloud method as he does with all the other method of testing that he describes in his book (Jordan 1998, page 58-59).

[6] In fact none of the sources I was able to get my hands on had any direct reference to Ericsson and Simon.

Advantages:

- "Participants' verbalisations make it possible to understand not only what problems they have with an interface, but also why these arise. This means that think aloud protocols can be an excellent source of prescriptive data, which can lead directly to design solutions"

- "Because think aloud protocol sessions where tasks are set can mirror controlled experimental sessions in their design, it might also be possible to use the session to gather objective performance data, such as task success and number of errors made"
    - "However, it may not be possible to reliably collect data with respect to more sensitive performance measures, such as time on task, as having to make verbalisations may slow the participants down."

- "Think aloud protocols can also be an efficient way of obtaining a lot of information from only a few participants. This is because each participant can provide such rich prescriptive information."

Disadvantages:

- "A possible disadvantage of thinking aloud protocols is the relation to the possible interference between participants' verbalisations and the tasks that they are performing. It could be argued that, in a sense, participants in think aloud protocols are performing two tasks – not only using the product under test, but also trying to verbalise what they are doing whilst using the product. The problem, then, is that this second task may interfere with the first and any difficulties that the user encounters could, possible, be connected with the distraction caused by having to make verbalisations."

- "Another potential disadvantage is that because participants are explaining their actions to the investigator, they may feel tempted to 'rationalise' what they do. This could mean that, for example, where a participant's approach to exploring an interface or trying to complete a task was really rather random, he or she might be tempted to give verbalisations that suggested that he or she was taking a fairly logical approach. Indeed, this effect may also work in reverse, with participants becoming 'trapped' by their verbalisations. If, for example participants give verbalisations that indicate that they are following a particular strategy, they may then feel obliged to continue with this strategy throughout the think aloud session"

- "The way in which the investigator prompts participants can have an effect on whether these problems occur. In particular, too much prompting can lead to the participant 'making things up' in order to respond. However, this has to be balanced against the risk of prompting too little, which may lead to the data gathers being less rich that could otherwise be the case.

> Having a feel for what is the right level of prompting, then, is a skill that as central to running an effective think aloud session."

But as was the case with Rubin there is not any direct reference to any sources that might have provided the input to specify this disadvantage, and the reference list in the back does not provide any obvious revelation to where these might have come from.

Do they match?

As can be seen from the different advantages and disadvantages listed by Rubin and Jordan, some of them are somewhat contradicting. Looking closer at the disadvantages listed by Rubin and Jordan also shows that they in some instances even contradict themselves.

Where Jordan speculate that the added task of speaking aloud will make it more difficult for the user to perform the primary task, Rubin says that speaking aloud although it will slow the thought processes down, doing so will increase mindfulness which will make the user more conscious of what he/she is doing and therefore prevent the user of making mistakes that the user would have been making, were the user not speaking aloud.

Jordan's first disadvantage (see page 30) is rather interesting in that it in many ways support the thesis of this paper that thinking aloud has an influence during the test phase and could prevent the user from solving the task that the user normally would solve. Unfortunately Jordan contradicts himself in that he also states under his second advantage that

> *"It might also be possible to use the session to gather objective performance data, such as task success and number of errors made*". (Jordan 1998 – page 58)

Although he says that the performance measures as time on task might not be totally accurate since the verbalising slows the user down. If verbalisation as he states in his first disadvantage has an influence on the primary task of solving the tasks in the interface that is being tested, then the method does not provide "objective" data on task success and number of errors made.

Barnum

When it comes to Barnum's 1 ¼ page of thinking-aloud[7], she only has one issue about using the think-aloud procedure during usability testing, which is the issue of time on task. But just as Jordan and Rubin in regard to whether time on task is a problem, the conclusion Barnum makes is somewhat "fishy". But in contrast to the other two authors Barnum actually has references that she uses. The use of them and the references is however in my mind somewhat questionable. As it turns out that

[7] This is without counting the 2 pages she's taken directly from the www. (page 236-237)

Barnum actually uses the same initial source (Ericsson & Simon 1980) to say that there is research that suggests that thinking out load has an effect on timed tasks as well as it does not have an effect (Barnum 2002 - page 238).[8] As a result of the research Barnum has reference to, which in her mind obviously is somewhat ambiguous about whether speaking aloud actually slows the test person down, she concludes that

"if absolutely accurate time on task is an issue, you should probably choose a method other then thinking out load" (Barnum 2002 - page 238)

But this recommendation has in my mind more to do with her being cautious then the actual research Barnum uses as an argument.

What is essential about all 3 authors, is that they all spend much time on usability and how it can be tested, and they all have a tendency to jump over the "think-aloud part" very quickly. Although they list advantages and disadvantages with the method, due to the lack of references in Rubin 1994 and Jordan 1998 case's, one gets the impression that the things they lists are more based on what the author thinks is common sense and the obligation to have x number of disadvantages and advantages, then they are based on actual hard researched evidence.

Conclusion on current prescriptions on the think-aloud method

What can be concluded about the current prescribed use of the think-aloud method is that there is a good possibility that the method is used in almost all aspect of testing the usability of applications and interfaces. From investigating the mental model of the user through "heavy" propping to the assessment of usability as well as usability performance data in the form of *efficiency* and *effectiveness*. Although it is not described that the think-aloud method should be used in what Rubin defines as a validity test. It is on the other hand not prescribed directly that the method should not be used in this type of test, just that is should be used with caution.

The current literature, when it comes to those prescribing how to conduct usability tests (Rubin and Barnum), uses very little time even describing how to conduct the thinking aloud part of the test[9]. In relation to the reliability of the data collected through thinking-aloud, the warnings given on the data are in some cases contradicting and at the same time difficult track to any base theory such as Ericsson and Simon's. It is therefore in my mind safe so say that it is very hard to exclude any

[8] Barnum referents Ericsson and Simon by the use of Van der Meij, H. 1997 as one argument, and with a citation to Hix, D. & Hartson, R. H. 1993) , the last of which in my mind only can come from Ericsson and Simon. There is however now direct reference in relation to the text cited, but Ericsson and Simon is listed in Hix, D. & Hartson, R. H. 1993 reference list.

[9] So little that one might even question if usability testing and the think-aloud method should be considered as being one in all as described by Jacobsen(1999)

situation on how the method is being used in the real world based on what is being prescribed from at "theoretical" standpoint.

As, there is not any situation the think-aloud method is not being used, when it comes to the testing of the different elements of usability. We can not beforehand rule out any potential situations where verbal overshadowing can not affect the reliability on the data provided.

4 Is verbal overshadowing a problem when using the think-aloud method?

Analysis of background theories and assumptions on verbal overshadowing

As it has been described in chapter 3.4 there are no prevailing descriptions or guidelines when to use the think-aloud method within the field of usability testing, a point also made by Hertzum & Jacobsen (Hertzum M. & Jacobsen N. E. 2003). So even though we have somewhat of an answer to sub-problem 1.1 of this thesis, the lack of guidelines makes it somewhat difficult to analyse in relation to how verbal overshadowing affects the think-aloud method as it is being prescribes at present. But at the same time it does not limit the possible ways in which the method might be used in real life, which very much leaves the possibility open for the think-aloud method being used in situation where there could be serious problems with the reliability of the data produced from the usability test.

In order to get a better understanding of what those situations are, we need to take a closer look at verbal overshadowing and how the findings made within psychology, particularly within cognitive psychology, can be transferred over to the field of usability testing and software production in general.

Because verbal overshadowing, as described in the previous chapter, as a term just covers over the fact that verbalisation in some instances has an altering influence on people, and the term could be described as somewhat elusive, the effects of verbal overshadowing are numerous. The following analyses will therefore first look at what these effects are in closer detail on an individual user basis and look at how this affect the liability of the data derived through the think aloud method. This part of the analysis will therefore look at the question of whether or not verbal overshadowing should inflict on how the user is chosen for a given test session that involved using the thinking aloud method and if the effect of verbal overshadowing might be used in a way that will provide more data in form of usability issues from using fewer test persons (sub question 1.2 and 1.3).

Choosing the right user for thinking aloud: Does verbal overshadowing strike everybody?

One of the essential problems in the analysis of verbal overshadowing in relation to problem 1.2 and 1.3 is to know if this phenomenon is something that everybody are influenced by whenever they think out load or if it defers from human to human. But in order to make this analysis it is necessary to divide this analysis in two to parts depending on the situation or scenario in which verbal overshadowing could be a factor. Based on the different situations in which verbal overshadowing can occur it is in my mind necessary to distinguish between verbal overshadowing caused by vocabulary skills and verbal overshadowing in relation to insight problems.

The reason for making this distinction is that verbal overshadowing in relation to vocabulary skills is something that everybody can be exposed too, but whether or not they will actually suffer from verbal overshadowing in a given situation will very much depend on what exactly they are talking about in that given moment, and whether or not they have the right vocabulary or not to be under the influence of verbal overshadowing.

Verbal overshadowing in concern to insight problems is as far as I can tell from the research found on the subject, and the theories made on insight problems and verbal overshadowing, not a case of verbal skills but has more to do with human physiologically and the mental processes that happens in the right and left side (hemisphere) of the human brain.

4.1 Vocabulary predicts verbal overshadowing

Different research strongly indicate that there is a clear connection betweens peoples verbal skills defined by there vocabulary within a given domain and if they are going to be influenced by verbal overshadowing or not (Ryan & Schooler 1998, Melcher & Schooler 1996). As briefly described earlier in chapter 3.3, this research show that you can predict verbal overshadowing in the remembrance phase when a person's perceptual expertise surpasses the person's verbal expertise and the memories are obtained while thinking aloud. Another way to describe the problem is that people can only remember thing as well as they can describe them verbally if they think aloud during the "experience" phase.

The big question in relation to usability testing is how this is relevant?

Transference of the scenario in which the findings made by Melcher & Schooler (1996) takes place, can be made in relation to how test persons remember the use of an interface or application and how

they remember the look of the interface[10]. If a test person is thinking aloud during the first time use of an interface or application and after a period of time subsequently is being asked to use the application to see how well the person remembers the application, in for example a performance or benchmark test. Then this situation is comparable with the situation being tested by among others Melcher & Schooler (Melcher & Schooler 1996) in regard to people's ability to remember and recognise different wines.

The key to how well the test person was able to later identify the wines tasted in Melcher & Schooler's research, lay in there ability to describe what they tasted. The test persons could be defined either as a novice, intermediate or expert according to the persons experience in wine tasting and drinking.

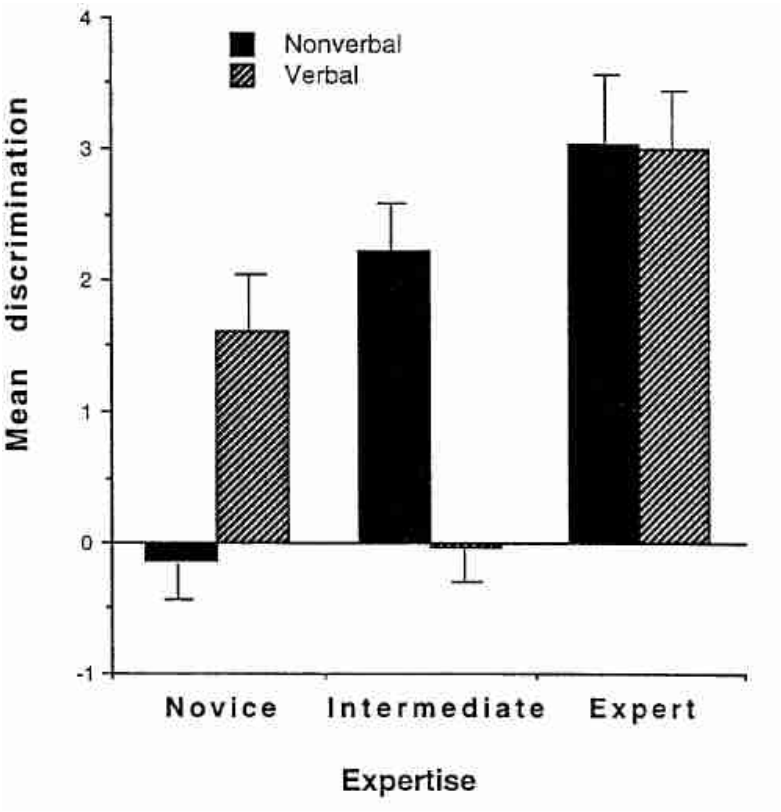


Figure 5 - Mean discrimination found as a function of difference in perceptual and verbalisation expertise[11]

[10] In the experiment people where asked to taste a wine and then describe it so that others might recognise the wine from the description. This was followed by a blind test where they should regognice the wine they ha previously tasted

[11] Melcher and Schooler 1996, Fig 1

As Figure 5 shows the experts performed the same way whether they talked aloud or not, when it came to remembering what they had perceived. For novices it actually helped to talk aloud in relation to remembering what they perceived. But for those with intermediate verbal expertise there ability to remember and recall what they perceived dropped significantly.

Other research has shown that there is in general a problem with performance when it comes to non-verbal knowledge. According to Ryan & Schooler (1996)

> *"Verbalization has been found to impair the implementation of task critically relying on non-verbal knowledge or processes, more propositional domains have been shown to be invulnerable to verbalization and to in fact often benefit from it*". (Ryan and Schooler 1996 , page 107)

Though at first it seems to limit the problem to include non-verbal knowledge, I would argue that it also emphasises that vocabulary is central in this issue, since non-verbal knowledge or tacit knowing (Polanyi 2005) is defined by its inability to be verbalised and therefore to be transferred from one person to another – one reason being that we as humans don not have the vocabulary to do so. A further analysis in relation to the more logical domains and how this affect the problem of verbal overshadowing will be done later in chapter 4.3.

One possible explanation to the findings shown in Figure 5 could therefore, as I see it, be that since the "expert" have the vocabulary to express what he or she perceives the person will not spend extra mental resources on finding the right words in order to verbally say what is perceived. Hence the zero effect in respect to non-verbal control group. The novice will, since the person has no knowledge in regard to terminology and specific vocabulary to the domain, just say what the persons perceive and think, and not spend time or mental resources wondering if he uses the right words etc. since the person is not aware that the word exist in the first place. The intermediate person will, since the person has some knowledge to the terminology used within the given domain, use mental resources to use the right words for what it is the person perceives and thinks during the task. A logical explanation for this behaviour, as I see it, is that although it might not have been instructed during the introduction face to the test the person will put pressure on her or himself to use the "right" words in the thinking aloud process.

In relation to Ericsson and Simon it could be argued that the verbal report given by the intermediate person is being reduced to "3 level verbalisation" because the person uses cognitive processes in the quest to locate the right term from the person's long term memory.

As shown in Figure 5 it would therefore be predictable to assume that people with intermediate verbal skills when it comes to explaining how they used the application, what they see on the screen etc. would have an impaired performance compared with a user that were a complete novice or an expert in relation to the verbal skills needed in a given situation. More importantly they would have an impaired performance compared to users with intermediate verbal capabilities that did not speak aloud during a usability test.

Although I have no direct evidence[12] I will argue that the effect of verbal overshadowing will diminish or even disappear as a person get used to the performed task. Simply because it is the same vocabulary that is needed, as long as the task does not change too much. And ones the person has verbalised his or her thoughts ones concerning the task. The person does not have to search for the right words again since they already are present in memory. Although not described specifically the effect is mention by Meissner and Brigham (Meissner, C. A., & Brigham, J., C. 2001) with reference to research made by Schooler and Engstler-Schooler in 1990 (Schooler, J. W. & Engstler-Schooler, T. Y. 1990). The effect can further more be seen in the experiment made by Ericsson on problem solving processes with the 8-puzzle (Ericsson, K.A. 1975) where the subject had an impaired performance[13] in the first couple of attempts to solve the puzzle (described in Ericsson and Simon 1984 / 1993).

The hypothesis that impairment due to verbal overshadowing disappears as long as the task remains the same does however raises one fundamental problem that future research hopefully can answer: When does a task change enough for it to be a new task that again is subject for verbal overshadowing, manifested by a longer time on task time, in the first attempt to perform a given task?

Can tasks for instance be categorised in a way so as soon as a test person has become familiar with one type of tasks and therefore now longer should suffer from verbal overshadowing, they can perform a new task of the same type without suffering from verbal overshadowing?

4.2 Can the effect of intermediate vocabulary skills be minimized?

A possible explanation to the relationship between vocabulary skills and verbal overshadowing that would also suggest a possible solution to keeping this potential problem on a minimum is given by Meissner & Brigham (Meissner & Brigham 2001) in their meta-analysis of the verbal overshadowing effect in face identification. They find that in relation to post descriptions the effect of verbal

[12] After hours of searching in vain I have not been able to locate the article I at one point had found, describing how the effect of verbal overshadowing disappeared when a person got used to a task.

[13] They used more moves then the non verbal group.

overshadowing is greatest when the person is instructed to give a more elaborative and detailed description of the faces they saw (also described in Meissner, C.A et.al (2001).

There is therefore clear indications that the more a person's vocabulary is tested in order to explain something that the person does not have the vocabulary to do, the person will suffer from verbal overshadowing.

In regard to the findings described by Meissner & Brigham these finding support the necessity to give an introduction to the test person so that the person knows the form and format of what he or she says during the test (thinks out loud) does not matter and is not important, what is important is that the person thinks aloud. It is impeccably important to the "intermediate" person – although such an introduction to a person with the intermediate characteristics has the inevitably chance of making the person more self-conscious of this matter if it is not done correctly, and not given in the right way!

The other key point to remember here is that because verbal skills seems to be the key to whether or not people will suffer from verbal overshadowing, the phenomenon is very domain specific from person to person, and for that matter from language to language. One person might be an "expert" vocabulary vice within a certain domain in one language – but that does not necessarily mean that the person is an 'expert' in a foreign language.

4.3 Do vocabulary skills impair performance in general?

The next question to ask in connection to verbal overshadowing dependence on verbal expertise and its relation to how people remember perceived experiences is whether or not this goes beyond memory and whether or not it could affect the actual test as well, and will impair the performance or otherwise have influence on how the assigned task would be carried out.

In this matter there are several researches that would indicate that verbal overshadowing is not a problem when it comes to verbalising for example analytical problems (Schooler et. al 1993, Ryan & Schooler 1996). But seen in relation to usability testing this research does not in my mind rule out the possibility of encountering verbal overshadowing. This is because none of the research that I have found, or found reference to has looked at the data gathered concerning non insight problems, based on how the test person's verbal skills was in relation to the domain in which the problem / task was solved.

As can be seen by the non-insight problems used by Schooler, Ohlsson & Brooks (Schooler et. al 1993) in "Appendix A – insight and non insight problems", none of these problems involves domain specific vocabulary that would be known to some people and not known by others. What these

findings say is, that in situation where an every day vocabulary is being used, thinking aloud during the solving of non insight problems verbal overshadowing is not encountered. And as quoted earlier by Ryan & Schooler it could even in some instances benefit from it.

But when we look at the different domains in which software is being developed, I would argue that domain specific vocabulary is encountered on a daily basis. Just verbalising what you see on the computer screen alone requires knowledge to a domain specific vocabulary. There are therefore in my mind strong indications to predict that users undergoing a think-aloud test with intermediate domain specific verbal capabilities will suffer from verbal overshadowing and have impaired performance as a result thereof.

4.4 Insight problems predicts verbal overshadowing

Even though there is research that has not been able to reproduce the effect of verbal overshadowing in relation to insight problem solving (for example Russo et. al. 1989). There is enough evidence of people having problem with insight problems while talking aloud, that its effect on the data produced during a think-aloud test session should be analysed, even thought verbal overshadowing still is a subject for debating within the field of physiology (Ericsson and Simon 1993, Ericsson 2002).

In order to analyse how talking aloud and insight problems could be a problem in a test situation we first need to understand what exactly an insight problem is, first then can we analyse whether or not this actually is a problem or not.

But in order to answer that question on insight's relation to usability and in order to be able to evaluate it in a usability context, we first have to become familiar with what defines insight problems? Or more precise, what defines a specific answer/solution or the path to one as being obtained by insight in relation to being non-insight, and how does speaking aloud influence solving a problem using insight.

4.5 What is insight?

The common description of insight is that the solution to a problem comes in form of an "aha" or "eureka" experience (ex. Schooler et al 1993 and Jung-Beeman et al 2004).

This "eureka" moment demands however that a person encounters an impasse in relation to the problem in hand, and in order to get the insight the person has to have the competence to solve the problem as well (Ohlsson, S 1992).

In Schooler et al 1993 the following points where used to describe the insight problems that were used in there experiment: (Schooler et. al 1993, page 168)

1) *It is well within the competence of an average subject.*
2) *Has a high probability of leading to an impasse, that is, a state in which the subject does not know what to do next.*
3) *Has a high probability of rewarding sustained effort with an "Aha" experience in which the impasse is suddenly broken an insight into the solution is rapidly attained.*

The problem about insight

There is however one fundamental problem, as I see it, in relation to insight. The problem with this "eureka" description in relation to the problem discussed in this thesis is that it is not very specific or concrete, and could be used to a number of different scenarios, many of which may or may not be an "insight" that is influenced by verbalisation.

Simply describing insight's by that of an "aha" experience, would mean that a number of scenarios would fall under the definition as being insight problem solving – the question is whether or not they are? The same goes for solution where a person might not get an "eureka" moment, but coming up with the solution/answer still needs what would be described as insight in the form of bringing "forgotten knowledge" or tacit knowing[14] on a given subject into the conscious part of the mind.
The sudden recall of historic facts or remembering the name of a certain actor in a movie etc. are some examples of situations where an "eureka" moment might be experienced by the person recalling the information. But if the same answer is found by a person without the "eureka" felling is it still an insight or is it another way of retrieving the information from ones memory, or does true insight require that different pieces of knowledge are combined into new knowledge in order for it to be an "insight" and remembering past obtained knowledge is something completely different, that is not influenced by verbalisation?

I myself have numerous times experience that after a long time of not working with a piece of software, I sometimes find my self wondering how on earth I knew that in order to do the task I had just done, I had to do what I had just did, especially with software that is not very user-friendly and where the interface provides absolutely no clues to have the solution such as the one shown in Figure 6. But does the uncurious retrieval of what application-number to use in the shown AS/400 working environment fit the description of an insight or is it something completely different?

[14] One example of tacit knowing is that you for example can be 100 percent sure that what some one says is not correct, without actually being able to remember the true answer. You just know that it is not true!

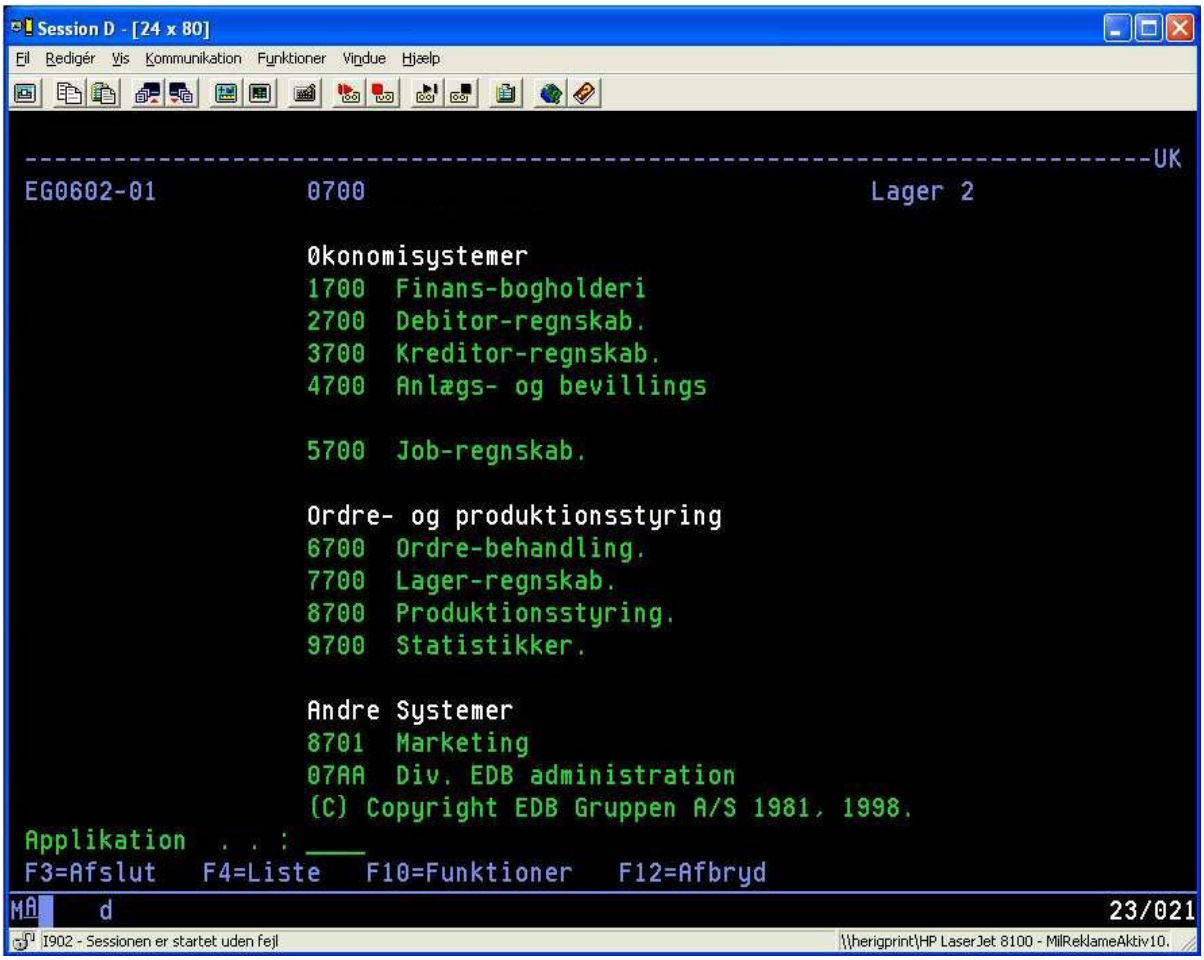


Figure 6 - AS/400 Session

In order to get a better understanding of when insight occurs and its relationship to verbalisation it is in my mind necessary to look at the processes that happen within the human brain and not how something is perceived by a given person.

The current theory made by Schooler (2002) concerning verbalisation and non verbal processes, which includes insight, is that the verbalisation of thoughts which happens in the left hemisphere of the brain somehow reduces the “mental resources” available for the right hemisphere, where it has been found that non-verbal processes occur and therefore reduces the chance of people having an insight. Research using functional magnetic resonance imaging (FMRI) and electroencephalogram (EEG) made by Jung-Beeman and others (Jung-Beeman et al 2004) actually shows how the neural activity changes in the human brain when an insight occurs.

But no research that I have come across has investigated how verbalisation during the solving of insight problems affects these neural activities. Neither have I come across any research investigating whether the neural activity found by Jung-Beeman et. al. resemble other situations where the classic “eureka” might be experienced by a person but the situation might not have

anything to with an insight on a neurological level. I have further more not found research that shows that certain types of solutions to a problem, or the retrieval of memories and past experience might still have come to a person as an insight, defined by the neural patterns in the brain, but the person might not have experienced or perceived the moment in the classic “eureka” moment or is aware that this is the way the solutions was reached.

How does insight correspond with usability in general?

In order to understand the implications of verbal overshadowing and insight problems in relation to usability testing and the use of thinking aloud during tests, we need, as mentioned before, to see if this analysis is relevant in the first place.

When we initially look at how usability is defined in the literature and hold the mental model of the user in mind, having to figure out how something works by being forced to have insights. Not being guided by the interface and the design of the application does not seem like something that is very desirable. And it would initially be something that should be avoided by all means, if we want to have software and interfaces with the best possible score in usability.

Although this is true when it comes to most software and equipment that is being used in relation to work situations, it is necessarily not the case when it comes to software within the entertainment industry, an industry that also uses the thinking aloud method in their tests of software. IO-Interactive[15], which produces the game series Hitman, does for example use the thinking aloud method when they test there games. And thought being forced to figure out how a work related software application works by having insights in order to complete a task is, is something that initially is very undesirable. This does not mean that no one has ever had insights when they where working with there text editor, spreadsheet or ERP system, and through an insight solved a problem they had. It might even have saved them the trouble of getting hold of a manual or getting other kind of help.

When it comes to the control of a game you want something that is as logical to the user as possible, but when it comes to how the game is played and completed many games such as adventure games is based on the user solving task that in many cases fits the description used by Schooler et. al. on there definition on insight problems perfectly. Whom has not found themselves completely stuck in a computer game only to get the solution to the problem at a time where we were not even think about the game (perhaps because we where so frustrated about not to being able to move on in the game and for that reason did not want to think about it).

Getting a better understanding of insight and in what situation insight occurs is therefore in my mind very relevant, in order to understand when verbal overshadowing could become a potential problem

[15] www.ioi.dk

in relation to the gathered data, and how they should be interpreted. But as mentioned earlier it is in my opinion necessary to get a better understanding of insight as based on the neural activity end their relationship to speaking aloud. First then can we get a better picture of the full implications of speaking aloud while solving problems or task that require past obtained memories.

But since insight is part of the working environment in some types of applications and is something to be avoided in others, a further analysis of how insight problems might be avoided or perhaps exploited is therefore necessary.

4.6 Insights and verbal overshadowing on an individual level

It seems somewhat peculiar that the software industry in regards to the test persons they use during the test of their products, have their test person perform multiple task at ones (i.e. thinking aloud, use of motor coordination, and performing a given task) much like a fighter pilot that has to perform mutable task in his cockpit (control his plain, talk in the radio or to his navigator and at the same time perhaps engaging an enemy), and the software industry have not asked themselves whether or not the test person was able to so, so that it was the application that was being tested and not the user. When the Danish air force for example tests people who applies as pilots to see whether they are capable of performing different task at the same time before you even consider putting them in a cockpit, why is the same not done within the software industry so that we are sure that it is in fact the usability of the software and interfaces that are being tested, and not the mental capability of a random chosen user who happen to come by the day we were performing a test?

To make this analysis we have to look at the individual user and not at the population as a whole, because just as there is a big difference in peoples skills when it comes to for example simultaneous capacity, another area that also has to do with peoples ability to get the brains two hemispheres to work with each other, the same difference could be the case when it comes to peoples likelihood of being influenced by verbal overshadowing.

The possible difference in people's ability is especially important if we only use a handful of test person or less, where the individual characteristics of a given test persons is not evened out by the use of a large number of test persons.

So does this mean that users are selected completely randomly with no opinion to their skills?

No it does not. But it is rather the test person's domain specific IT skills and knowledge that is the base for selection and grouping, than it is the person's ability to provide good and solid test data. To make the conclusion that no one within the software development business looks and select test

person's as to their ability to be good at talking aloud and use the software at the same time, is impossible. It would seem logical that software developers from pure experience have found what they regard as a good test person (i.e. a specific person that they perceive to provide good test data) and at the same time have selected not to use people again in later test sessions because they found them unusable as test persons. But getting that kind of experience, inevitably cost resources in form of time and money, resources that could have been spend on other projects or could have been saved.

The alarming point in this connection, is that I have not been able to find any research or litterateur within in the field of Usability testing or elsewhere that addresses the problem of selecting and screening for users that can provide you reliable data. Not from a theoretical standpoint or in form of looking at how users are selected in the "real world" outside the research laboratory.

What is known on an individual level about insight and verbal overshadowing?

The research made by Schooler et. al (1993) provides several elements that could be very interesting in an analysis of verbal overshadowing on an individual level. And could very well provide part of an answer in relation to problem 1.2 in form of providing significant clues to the question whether or not there is an individual difference on who potentially would suffer from verbal overshadowing in relation to insight problems. Something that would suggest the need or possibility for a pre-screening of people's ability, before they were used as test persons in think-aloud test.

When it came to solving the insight problem in the experiment set up by Schooler et. al. (1993) the test was designed so that the subject had 6 minutes to solve the problem. The difference found in how talking aloud and silent test persons was able to solve the insight problems was not so much found in that the subjects talking aloud used scientifically more time to solve the problems when the found the solution. It was more a case of solved / not solved. The people that solved the problems while talking aloud or without talking aloud, in most cases solved the problems well within the set timeframe of 6 minutes. The different insight problems used in the test could be ranked by how difficult they where to the test persons used in the test. And why is this relevant?

Take a scenario where you as a developer has to perform a think aloud test and you have 2 users to choose from and only time and resources to perform one test, and they both have the same profile regarding IT experience which one do you choose? Do you flip a coin? Take the pretties of them? Or would it be sensible to choose the one that had the best profile regarding verbal overshadowing, and look a there ability or for that matter incapability to perform a task and think aloud at the same time – and hence be giving you as a developer or evaluator the best possible and reliable data? Especially if you by the use of very little time and resources by conducting a small test could make this selection

and at the same time get the awareness that certain result from the test of your software should be interpreted with that in mind that it might be the user that was suffering from verbal overshadowing and might not your software that was defective. It might even tell you that making the users thinking aloud during the test was not the best way and would not give you any reliable data, and performing the test would be a waste of time and resources that could be spend more wisely.

The ranking found by Schooler et. al (1993) could be very interesting if it could tell us something about the relationship between peoples ability to solve the problem ranked as the most difficult found in the test and there ability to solve problems ranked as easier.

There is however one essential problem with a further analysis on this subject in relation to verify whether or disclaim the relation between the found difficulty of an insight problem and peoples ability to solve other insight problems. A problem that also excludes an analysis based on evidence and instead leads to a more hypothetical discussion / analyses instead. The data collected by Schooler et. al. in regards to the mentioned experiments does not exist anymore.[16]
So to answer the headline of this particular chapter – very little is actually known on the individual level about insight problems and verbal overshadowing.

Although it therefore is impossible to conclude that there is an individual difference, due to lack of sufficient data I will however spend some time on what I see as a possible way to pre-screen people for there ability to solve difficult or easy insight problems, and the possible problems as I see them with that particular solution. Not because it will provide much of an answer, but because the next discussion in my mind rises some questions that others, with the right expertise, might be able to answer and therefore provide knowledge to the problems raised by this thesis.

Can insight problems be ranked, and can it be exploited?

If we wanted to have a test person that was not subjected to be influenced by verbal overshadowing or for that matter the opposite[17], when it came to solving insight problems then what if we could have a simple test that would indicate a persons profile in relation to how that person would be influenced by verbal overshadowing?

With a base in the ranking of insight problems used by Schooler et. al., assume it was found that a person was not able to solve problems ranked as being more difficult the easiest the person was unable to solve. And that the person was successful in solving any question easier the most difficult he had solved Then by making a test of selected ranked insight problems, it would in theory be

[16] See appendix B for mail correspondents with Jonathan Schooler.

[17] How peoples inability to talk and solve problems could be exploited will be addressed later in this thesis

possible to see if a person used in at think-aloud test, would be subject to verbal overshadowing and potentially how much. This would mean that an evaluator would have a better basis for interpreting the data gathered in a given test.

The first problem with this is obviously that the data does not exist anymore. The second problem is with the ranking of the questions. It might be possible to rank the question on an overall level for a large population, but again looking at an individual level the variance might possible be too great for it to be useful. How a person or large population sees an insight problem could very much depend on the experience of a single individual the same way as the mental model of a computer program does.

This point can be illustrated by using one the insight problems used by Schooler et. al. (1993).
The insight problem asked to test person was the following:

"A prisoner was attempting to escape from a tower. He found in his cell a rope that was half long enough to permit him to reach ground safely. He divided the rope in half, tied the two parts together and escaped. How would he have done this?" (Schooler et. al. 1993, Appendix A., question 2.)[18]

The solution too this problem is that in order to get down from the tower you split the rope, wish is made by 2 or more smaller ropes twisted together, and tie the 2 parts together and you now have a rope that is just as long as the tower it high, and you can therefore get down from the tower.

In order to get to that solution you have to know that thick ropes are made up by smaller ropes (typically 3) that are twisted together. These smaller ropes are again made up by smaller ropes or individual fibres that also have been twisted together and so on. But this is only the case for conventionally made ropes. Special ropes like climbing ropes in particular are not made this way. Here smaller individual treads are weaved together to make the rope, and it is not possible to split these kind of ropes in two and then put them together.

[18] A complete list of the insight problems and non insight problems used by Schooler et. al can be seen in appendix A

Figure 7 - On the left traditionally made rope, on the right modern climbing rope.

The question is therefore if you take two persons, that both know that a conventional rope is made up by smaller ropes, but one is a professional climber who on a daily basis only uses the new kinds of weaved ropes – the other is a fishermen, which daily job among others consists of handling and splicing conventional ropes together (a thing that is not possible with the new kind of weaved ropes). Do both of these persons due to their past experience have an equal chance to get the solution to the insight problem, described above?

Unfortunately that question will stay unanswered within this thesis, but if we want to get a better understanding of how verbal overshadowing can influence insight problem solving, we also need to know how people perceive insight problem as being easy or difficult, in order to have an insight into how the test persons might react to verbalisation during the actual thin-aloud test.

Sub conclusion on insight

As I have argued in the previous sections, there are various pieces of circumstantial evidence that indicate that speaking aloud during a usability test, involving software with insight problems such as certain types of games, will potentially reduce the reliability of the data gathered during the test.

As seen in relation to the overall question, it seems that there is a potential problem when using the think-aloud method in relation to software, which deliberately is not made with the intent of guiding the user from A to B in order to complete a certain task, but where one of the applications premises is to challenge the user. There is in my mind a high possibility of generating data that would indicate a higher number of usability issues, but where some of these found issues, in reality are caused by the test person's inability to talk and come up with the solution at the same time.

4.7 Can verbal overshadowing be exploited to locate non-logic designs?

Arguments for an individual difference in a person's ability to solve insight, while talking can also be seen in every day life, by the way some people are having trouble in multitasking, and others, such like persons accepted as pilots by the armed forces, in contrast to the average person, are better at multitasking. A resent report from the Danish Transport Research Institute for example states that having a telephone conversation using a hands-free set is just as dangerous as not using one (i.e. holding the phone in your hand). It states:

"The effect of cognitive and visual loads is related to the conversation-phase, and is equal for the use of both types of phones[19]*. The typical effect is reduces awareness, increased reaction time, and more unsafe driving"* (Ritzau - 4 December 2006, can be seen in appendix C)

What in my mind is very striking is that the situation is not unlike the situation of a think-aloud test, where a person is asked to perform multiple tasks involving, speech, motor coordination, and problem solving.[20]

As has been analysed in a previous chapter (4.1), when is comes to non-insight problems verbal overshadowing is not present in situations where language/vocabulary is not a factor, and does therefore not impair peoples ability to solve these problems.

But based on the analysis on insight in the previous chapter, and the assumption that there is a measurable difference in people's ability to have insight during verbalisation, the following statement could be made:

Statement: *Since verbal overshadowing apparently does not have an effect on peoples ability to solve problems when it comes to analytic problem solving with no domain specific vocabulary. The effect of verbal overshadowing in relation to insight problem solving will only enhance that there is a usability issue that should be addressed in the pursuit of better usability defined as more logical applications.*

[19] Hands-free phones vs. non hands-free phones

[20] Just as verbalisation is generated in the left hemisphere of the brain, so are the motor function of the right arm (Nielsen O. & Springborg A. 2002) controlled by activity in the left hemisphere, and the right hand is the preferred hand for approximately 90 percent of the population (Hellige, Joseph B. 1993). If failing to achieve insight because the overshadowing effect of right hemisphere activity, as argued by Schooler, one could in my mind easily argue that adding motor functions controlled by the left hemisphere as well, does not help on achieving insights.

This statement does however depend on the definitions of usability in relation to the application developed. But in a test situation where you want to eliminate the possibility that the test subject is drawing on past experiences, or at least want to keep them at a minimum, when using an application or interface, it does in my mind make sense to use persons, where the forced verbalisation will have this particular effect.

By knowing the different cognitive abilities, and not just IT or other domain specific abilities, for a test subject, it will make the software developers able to select the test person (or persons) that is most likely to fail during the test due to usability problems, and thereby have the potential of getting data that would raise the standard of usability in the design.

5 New aspects to the use of the think-aloud method.

The analysis of the different aspect of verbal overshadowing has in relation to the think-aloud method, and more specifically on how it is used, exposes some new element that system developers conducting usability test, should be aware of. Ignoring these aspects will in my mind lead to a continued reduced reliability in the data gathered and subsequently leading to impaired usability.

Even though there in some areas are similarities to the current discussion between the original theory on verbal reports by Ericsson and Simon, and the current practise of the think-aloud method as discussed by Boren & Ramey and Abolrous[21], and the elements of verbal overshadowing as discussed in this thesis. The analysis has in my mind set the arguments for the following aspects for further consideration when using the think-aloud method:

1. **The verbal skills of the test person.**
2. **The test person's ability to solve insight problems during verbalisation.**
    - **The cognitive ability to perform multiple tasks at ones.**

These aspects form the ground for supplements to the current theory of the think-aloud and how it should be used to test usability. The new theory on the effect of verbal overshadowing and the impairing effect on the validity and reliability of the data produced, can be illustrated by the following hypotheses.

Verbal skills predict impaired performance

Hypothesis 1: T*est persons with domain specific intermediate verbal skills (concerning IT or the subject the application), will have either impaired performance or changed behaviour caused by verbal overshadowing in proportion to a non-verbal control group. Vocabulary novices and experts will in contrast (to the intermediate) not suffer from the same effect of verbal overshadowing and will not show any degree of impaired performance or changed behaviour.*

[21] Much of Abolrous's work is however based on the same material as Boren & Ramey, among other Boren's PhD.

Hypothesis 2*: The effect of Verbal overshadowing will be present in the beginning of every new task, with a domain specific vocabulary, that is being engaged by a test person. This effect will however diminish as the test person gets familiar with the task at hand, and the vocabulary needed is easier available due to resent use on previous similar tasks.*

However it is not just during the think-aloud test verbalisation can be a problem:

Hypothesis 3: *When testing how well people remember the use of a given application, test person's that learned the application while verbally describing there thoughts will display an altered ability in contrast to non-verbal learners. The effect will however depend on a disparity between the test person's verbal capability and their perceptual ability.*

Verbal overshadowing and insight:

Hypothesis 4: *Just as there is a difference in people's ability to multitask, there will be an individual difference in how people react to insight problems and think aloud at the same time. Selecting test persons based on there ability to solve (or not to solve) insight problems during verbalisation will make the data from the think-aloud method more valid and reliable.*

Hypothesis 5: *Choosing a test person that is less likely to get insights during verbalisation would lead to the discovery of more usability issues in the design i.e. of illogical design where the completion of a task requires heavy use of past experience.*

Hypothesis 6: *When testing application where insight is a purposely build design feature, it would be anticipated that test persons unable to get insights, would change there behaviour and instead try to get to a solution by "grinding" there way through the different possible solutions they can think of, until they hopefully get to the right one at some point.*

6 Does the knowledge of verbal overshadowing give better usability?

The straightforward question to the raised hypotheses is of cause, if they bring anything new to usability testing, do they have there justification?

As it turns out, although I have not been able to locate research on verbal overshadowing in relation to the think-aloud method, this is however not the same as saying that the phenomenon of verbal overshadowing, has not made its way into a domain of IT. I did manage to come across research that specifically speaks of verbal overshadowing within the area of IT in that Hepting and Arbuthnott (Hepting D.H. & Arbuthnott K.D. 2003) looks on the implications of verbal overshadowing for computer interface design. They have found that the pattern of how the user looks at the interface changed as a result of an interface being predominately verbal in contrast to a predominantly visual interface. Although there was no verbalisation involved in the form of thinking aloud, the internal verbalisation (reading) somehow changes the way users looks at an interface as well as it consistently slows them down in relation to users using a more visual based interface.

Just as verbal overshadowing is being argued fears fully within cognitive psychology, because the phenomenon of verbal overshadowing in many ways challenges the theory made by Ericsson and Simon on verbal reports and the three levels of verbalisation. One could make the obvious dispute that the same arguments that are used to refuse the existing of verbal overshadowing within cognitive psychology can be used when looking at verbal overshadowing, in relation to the reliability and validity of the data when testing for usability issues.

The answer to that argument is both yes and no.

When looking at hypothesis 1 and 2 in respect to the theory on verbal reports, one could very easily argue that nothing is new, because what these hypotheses essentially specifies is that for a test person with intermediate vocabulary skills, the verbalisation taking place in not level 2 verbalisation but level 3 verbalisation, and Ericsson and Simon themselves prescribe that this type of verbalisation should be avoid. This is all perfectly true as I see it, but in contrast to what was previous the case, software developers now have a potential indicator to predict when the data collected by the think-aloud method, even by Ericsson and Simons own prescriptions, is unreliable and should be taken under serious consideration when used; **the verbal skill of the test person.**

Something that is not part of the current discussion on how to get more reliable data from the think-aloud method. But also in the case of hypothesis 3, where it is hypothesised, that there might be a

problem when testing performance (without having the person think-aloud), if a person learned to use the application or interface during verbalisation.

It is however rather difficult to say anything about how often the circumstances in hypothesis 3 will arise, since I do not have any data on the subject, but I would speculate that it is not very often.

The aspect of problem solving in relation to giving verbal reports is in it self not new. Ericsson and Simon speaks in general terms of avoiding 3 level verbalisations since this will change the way people think and thereby also change the way people will solve a task. But although the problems concerning the changed behaviour have been touched also by usability practitioners, it has been in very general terms without clues to when this disadvantage would apply, and in relation to what kind of problems and what kind of user.

The following disadvantage listed by Rubin is in my mind a perfect example of this:

"Thinking aloud slows the thought process, thus increasing mindfulness. Normally, this is a good effect, but in this case it can prevent errors that otherwise might have occurred in the actual workplace" (Rubin 1994, page 218)

The aspects of verbal overshadowing, does not change with this possible disadvantage, since the right combination of user and task at hand will have this effect[22]. But the extra dimensions added by the stated hypotheses on the effect of verbal overshadowing, is that it gives a way to predict when the behaviour described is less likely to happen, whit the same task at hand, and with the same type of assumed verbalisation. Hence making it possible to interpret the data from a think-aloud test on the right premises.

The described behaviour by Rubin in relation to mindfulness is also not the case in relation to the described behaviour in hypothesis 4, 5 and 6, and the concept of people not being apple to solve certain kinds of problems as is claimed by these hypotheses, and not just slow them down, is as far as I can tell not a concept previously touched usability practitioners and theorists.

Should the evaluator be aware of verbal overshadowing?

As described in chapter 3.2 the data derived from the think-aloud method does not in itself say anything about the usability problems in an application or interface, these are first found when the data is interpreted by another human being. It is therefore important that the evaluator knows on what ground the data should be analysed.

[22] Based on what was found in the analysis it would be predictable behaviour for a user with novice verbal skills being asked to make a describing verbal report.

If the evaluator is not aware that the data gathered in some situations in unreliable in certain aspects, there is a good chance that the data will be misinterpreted. The result will be usability that would have been different where data gathered in another fashion or with another user, and therefore in those specific situations, would have been more valid.

Even though an evaluator strive to keep the verbalisation of a subject to level 2, by not using any probing, and keeping within the prescription given by Ericsson and Simon on how to conduct thinking aloud sessions, the subject might still deliver 3 level verbalisation because of the person's verbal skills in contrast to his perceptual ability.

Because the diverse use of the think-aloud method, conflicts with the original theory by Ericsson and Simon as data often is gathered with the help of level 3 verbalisation. The reliability and validity of the data is already threatened (Abolrous 2001). The problem with verbal overshadowing is that it in some cases is also a thread to the liability and validity to data that is gathered according to what Ericsson and Simon would consider as level 2 verbalisation, something that evaluators should also be aware of i.e. that could be conceived as level 2 verbalisation but is in reality level 3..

Looking back at the disadvantages listed by Rubin, Jordan and Barnum the real problem with these are that they are lacking at least one or two dimensions to their disadvantages in order to know when they apply. One possible reason being that it is not within the scope of what they are trying to achieve, but as the analysis in chapter 4 has shown, coming with specific references to when applied, is not something you just do, especially because there is still a lot to learn in order to give more specific recommendation to when the effect of verbal overshadowing will be present.

7 Discussion

It has in this thesis been the objective to understand the effect of verbalisation when conducting usability test, in order to get more reliability and valid data, and thereby getting better usability in developed software and interfaces. As it has been argued throughout this paper, there are several aspects about the current use of the think-aloud method that not only defies the base theory of how the think-aloud method is prescribed used within cognitive psychology. But new developments in cognitive psychologies understanding of how people are affected by speech and verbal reports, have also not at this point made its way into the field of usability testing. Developments that in my view are essential aspects in creating a theory for the use of the think-aloud method's use as a tool for testing usability.

7.1 Results

In order to answer the overall question of this thesis ***"Is verbal overshadowing a problem when using the think-aloud method within usability testing?"*** The following sub question has been analysed with the following result.

1.1 How does literature prescribe that the think-aloud method is being used and is verbal overshadowing a problem, in the way the think-aloud method is being prescribe used at present?

The think-aloud method is according to that part of the literature dedicated to describing how to conduct usability testing, used to a variety of purposes, from investigating the mental model of the user in relation to the design model, to looking at how the usability applies to set performance requirements. How the test is used depends however heavily on how usability is defined and in what part of the development cycle the test is being conducted. The consequence of this is that it leaves the possibility open for use in which verbal overshadowing is a problem. The result will be unreliable and invalid data, with deteriorated usability as a result. The disadvantages concerning the think-aloud method described by the common literature on usability testing are either contradicting or so universal, that they do not give specific direction to when exactly those disadvantages apply.

There is therefore a great possibility that aspects like solving insight problems and a person's verbal expertise vs. his or hers perceptual ability, both found to be elements related to verbal overshadowing, are generating impaired data by today's use of the think-aloud method. Since there is no literature that I have found that takes these element into account, evaluators are at present interpreting this unreliable data, eventually leading to the discovery of inaccurate usability problems.

1.2 Should verbal overshadowing be considered when choosing users for think-aloud testing?

Because there are individual differences in people's verbal skills and there perceptual ability, and there might be a possible difference in peoples ability to solve insight problems and talk-aloud at the same time. The effect of verbal overshadowing should therefore be taken into consideration, when choosing the test persons participating in a usability test involving thinking aloud.

Knowing how a test person will react to thinking aloud will make the evaluator able to focus on the usability issues caused by the interface or application, and not on the problems caused by the test person's inability to think-aloud and perform a task at the same time. Even if the right test person in respect to verbal skills in not found, it will make the evaluator aware of how the data should be viewed in order not to misinterpret the data produced during the actual think-aloud test.

1.3 Can the phenomenon of verbal overshadowing be used in a way that would make the think-aloud method more effective?

When it comes to the aspect of verbal skills (vocabulary) for the test person it is very difficult so see how verbal overshadowing can be anything else but a source of disturbance to the validity and reliability of the data. It is however possible that in specific situations where the usability of an interface or application is set to be as logic as possible, and where you want to avoid that a test person is using the application based on passed experience, that you would be able avoid this by selecting a person which is less likely or incapable of retrieving those past experiences, while talking aloud at the same time (illustrated by persons inability to verbalise and have insights at the same time). Past experiences, which would otherwise help the person to use the application and interface. In this specific situation verbal overshadowing would elucidate the areas that where not logical, to the evaluator, point out where the person would be attempting to perform at task by calling on past experiences.

The summarised conclusion to the overall question ***"Is verbal overshadowing a problem when using the think-aloud method within usability testing?"*** is therefore that verbal overshadowing is a problem in usability testing. Taking actions to eliminate the effects of verbal overshadowing or being aware that they exist, and in what situations they will appear, will only help to get better usability in the developed software and interfaces.

7.2 Limitations

Even thought this has been an attempt to paint a “bigger picture” of the different aspects of verbal overshadowing, in relation to usability testing and the think-aloud method. Stating that this is the whole picture rather then just one or two peaces two of the puzzle would be an overestimate of dimension.

As the adage goes: *"The wiser I get, the more I realise, how little I know"*

This is also the case when it comes to the problem of verbal overshadowing and its effect on usability testing using the think-aloud method. Trying to connect the dots between one, an area of practise (usability testing / the think-aloud method) that seems to be out of sink with the theory on which it is based upon, and second a phenomenon (verbal overshadowing) that still is heavily debated within the field of cognitive psychology, is and has not been an easy task.

Because the conclusions and argumentation within this thesis can not be any better then the data they are based upon, it is as stated in the beginning difficult to refuse that the outcome would not have been different was other material found in the process. Other conclusion would also have been the case if articles and material already included was being scrutinised by others that would have placed another emphasis on them.

Hopefully others, with a desire for developing better usability and surge of a better understanding of how and when the think-aloud method can provide the best possible data to support this development, will be able to put the conclusion and argumentation within this thesis, into a relevant perspective that will help the understanding of usability testing and the challenges involved in developing better usability.

7.3 Recommendation to practitioners

Looking at the effect of verbal overshadowing there are 2 areas in which practitioners of the think-aloud method should be aware of the possibility impairing side effect by using the think-aloud method to test usability.

1) **The verbal skills of the test person.**
2) **The test person’s ability to solve insight problems during verbalisation.**

When using the think-aloud method as a way of testing usability, practitioners should seriously consider the purpose of the test. It is hard to argue against the fact that the verbal data provided by the think-aloud method in contrast to a silent test with the same technical setup, makes it easier for the evaluator to locate usability issues and get a close estimate to the root of the problem. The reason being that the test person in many cases by him or herself will give a verbal report to what the person sees as the problem, or because the evaluator specifically probes for a cause to the problem during the actual test. But evaluators should be very aware of the fact that the price, for what at first can be seen as data easily interpreted, could very well be a misrepresentation of how the test person would have performed, were the person not asked to think aloud during the test.

As argued and analysed in this paper, there is more to the user then the standard user profile, indicating the users work or IT related skills. If more emphasis is to be oriented on the test person's ability to test the application or interface, it is necessary to look at people's ability to perform multiple tasks at ones, so that it is not a person's ability to think-aloud, and at the same time trying to perform a task, that is being used as a basis for locating usability issues.

In general the phenomenon of verbal overshadowing has illuminated the problem that making people perform multiple tasks at ones changes the way they behave. It should therefore be taking into consideration that the test person is not being asked to do so many unfamiliar tasks at ones that the person is not able to use the application. If the situation calls for multiple tasks, then it would be recommended to use persons that are more capable of performing multiple tasks at ones, including giving verbal reports.

Even thought the exact implications of verbal overshadowing has not been determined at this point, there are however ways that practitioners can minimize the effect of verbal overshadowing.

1) *Test persons should be instructed only to say what is on there mind. Any kind of verbalisation, consisting of elaborate and detailed description of what the person perceives should be avoided, in order to keep the effect of verbal overshadowing at a minimum.*

In order words encouraging a test person not to by shy of saying something in the order of ***"this round greyish thingy",*** and at the same pointing it out (for example with the mouse on the computer) instead of saying ***"the second radio button from the top in the top frame of the webpage"***. Thereby refereeing to things they see by the shape or colour instead of trying to use the precise technical terms, could be a way of terminating the effect of verbal overshadowing for persons with intermediate verbal skills.

It is furthermore recommended that:

2) *In the specific situation where insight is part of the design, it would be recommendable to compare the data and behaviour of the test persons with a silent control group. In order thereby to eliminate that the problems found in reference to not getting to a solution, was not caused by the person's inability to talk-aloud and have insights at the same time.*

To sum up the recommendation for practitioners in view of how verbalisation can affect person's behaviour, I will state what some of the authors (i.e. Barnum, Jordan and Redish) in my mind have been neglecting to do, so that it can not be misunderstood.

3) When testing usability to see if it meets any given set of measurable requirements, do so by mimicking how the application would be used in real life. Meaning that if no verbalisation is involved in the use, then the test should be done without having the test person think-aloud. If usability issues are being discovered, and if the visual data from the first test is inadequate, and a subsequent need for verbal data to locate the cause of problems exist, another test involving thinking aloud can be performed to provide the needed data on the issues discovered.

If better usability is the ultimate goal, this should in my mind apply until the effects of verbalisation are better understood and the problematic scenarios concerning verbalisation can be singled out and in other ways be prevented.

7.4 Recommendation to further research

If it is a goal to have better usability in the software and interfaces being developed, there is a need not only to know how in practice to conduct usability tests with the use of thinking aloud. There is also a need for guidance on when to use the think-aloud method, when not to use the method, and more importantly why. Something that is not available at present. The only way to better understand what is being tested, by the data produced through the think-aloud method, is to know the cognitive profile and capabilities of the person performing the test better. As described in the beginning, the think-aloud method is at present being used as where there only one unknown in the test, represented by the application or interface being tested. As illustrated in this thesis this is by fare not always the case, why new "equations" are needed in order to understand the result of the test (the gathered data) better, without having the same test performed in silence in order to verify the validity of the data produced.[23]

[23] Something I find difficult to believe is being done at present by many, since resources in many cases does not allow this extra test.

With the introduction of verbal overshadowing into the equation there is however some fundamental questions that need to be answered in relation to the hypotheses made in chapter 5 of this thesis.

If an effect of verbal overshadowing is found, either in relation to verbal skills, or in relation to insight problems, the next question that needs to be answered is:

- *How big are the effects of verbal overshadowing?*
- *And is the effect found on verbal overshadowing a factor that should be taken serious at present, (based on the size of the effect) or is the effect overshadowed by other aspects of the think-aloud test such as the evaluator effect as described by Hertzum & Jacobsen (Hertzum M. & Jacobsen N. E. 2003).*

This would in my mind make it possible to set up sufficient equations, so that data provided by the think-aloud method can reflect the usability of what is being tested, and not a combination of man and machine. This way usability can be developed according to the users need, reflected by there normal true behaviour, and not according to there inability to verbalise what they do and see.

Words will in another sense, be eliminated from getting in the way of better usability.

8 Reference list

Abolrous A. 2001 – Probing and its Effects on the Validity and Reliability of Verbal Reports
http://www.abolrous.com/sally/projects/probing_paper.htm

Barnum C. M. 2002 – *Usability Testing and Research*
ISBN 0-205-31519-4

Boren M. T. & Ramey J. 2000 – *Thinking Aloud: Reconciling Theory and practice.*
IEEE Transactions on professional communication, Vol. 43, Number 3, 261-277

Ericsson, K.A. 1975 – *Instruct ion to verbalize as means to study problem solving processes with the Eight Puzzle: A preliminary study(No. 458).*
Reports from the department of Psychology.
Stockholm: University of Stockholm

Ericsson, K.A. 2002 – *Toward a Procedure for Eliciting Verbal Expression of Non-verbal Experience without Reactivity: Interpreting the Verbal Overshadowing Effect within the Theoretical Framework for Protocol Analysis*
Applied Cognitive Psychology, Vol. 16, 981-987

Ericsson, K.A. & Simon, H.A. 1980 – Verbal Reports as Data.
Psychological Review, Vol. 87, number 3, 215 - 251

Ericsson, K.A. & Simon, H.A. 1984 - *Protocol analysis*
ISBN 0-262-55012-1

Ericsson, K.A. & Simon, H.A. 1993 - *Protocol analysis*
ISBN 0-262-05047-1

Fallshore, M. and Schooler, J, W 1995 – *The verbal vulnerability of perceptual expertise.*
Journal of Experimental Psychology: Learning, Memory, and Cognition, June. Vol. 21, 1608-1623

Finger, K. 2002 – *Mazes and Music: Using Perceptual Processing to Release Verbal Overshadowing*
Applied Cognitive Psychology, Vol. 16, 887-896

Frøkjær, E.& Hertzum, M. & Hornbæk, K. 2000 – *Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction really correlated.*
Proceedings of the SIGCHI conference on Human factors in computing systems CHI '00
ACM Press

Hertzum M. & Jacobsen N. E. 2003 – *The evaluator Effect: A Chilling Fact About Usability Evaluation Methods*
International journal of human-computer interaction, Vol. 15, 183-204

Hellige, Joseph B. 1993 - *Hemispheric asymmetry what's right and what's left*
ISBN 0-674-38730-9

Hepting D.H. & Arbuthnott K.D. 2003 – *The Implications of Verbal Overshadowing for Computer Interface Design*
Technical Report TR-CS 2003-10
ISSN 0828-3494
ISBN 0-7731-0461-5

Hix, D. & Hartson, R., H. 1993 – *Developing user interfaces: Ensuring Usability Through Product & Process*
Wiley
ISBN 0-471-57813-4

Jacobsen, N. E. 1999 – *Usability Evaluation Methods – The Reliability and Usage of Cognitive Walkthrought and Usability Test*
Ph.D. Thesis October 5 1999, Department of Psychology, University of Copenhagen

Jordan P. W. 1998 – *An Introduction to Usability*
Taylor & Francis Ltd.
ISBN 0-7484-0762-6

Jung-Beeman, M. et al 2004 - *Neural Activity When People Solve Verbal Problems with Insight*
PLoS Biology | http://biology.plosjournals.org
April 2004, Vol 2, Issue 4, Page 0501-0510

Knox, S. T, Bailey, W. A. & Lynch, E. F. 1989 - *Directed Dialogue Protocols: Verbal Data for User Interface Design*
CH1'89 PROCEEDINGS may 1989, 99 283-287

Kuniavsky Mike 2003 – *Observing the user experience: A practitioner's guide to user research*
Morgan Kaufmann Publishing
ISBN 1-55860-923-7

Lewis C. 1982 – *Using the "Thinking-aloud" Method on Cognitive Interface Design*
IBM Thomas J. Watson Research Center.
Yorktown Heights, NY 10598

Meissner, C. A., et Al 2001 – *The influence of retrieval processes in verbal overshadowing*
Memory & cognition, Vol. 29, 176-186

Meissner, C. A., & Brigham, J., C. 2001 – *A Meta-analysis of the Verbal Overshadowing Effect in Face Identification*
Applied Cognitive Psychology, Vol. 15, 603-616

Melcher, J.M. and Schooler, J, W. 1996 – *The Misremembrance of Wines Past: Verbal and Perceptual Expertise: Differentially Mediate Verbal Overshadowing of Taste Memory*
Memory and language, Vol. 35, 231-245

Metcalfe J. & Wieber, D. 1987. *Intuition in insight and non-insight problem solving*.
Memory & Cognition, Vol. 15, 238 – 246

Nielsen J. 1989 – Usability Engineering at a Discount
Designing and using Human – Computer Interfaces and Knowledge Based Systems edited by G. Salvendy and M.J.Smith
Elsevier Science Publishers B.V. Amsterdam 1989

Nielsen, J, 1993 – *Usability Engineering*
Academic Press, Inc.
ASBN 0-12-518405-0

Nielsen, J. 1994 - *Estimating the number of subjects needed for a thinking aloud test*
Int. J. Human-Computer Studies, Vol. 41, 385-397

Nielsen O. & Springborg A. 2002 - *Ind under huden: Anatomi & fysiologi*
ISBN 87-628-0271-*2*

Ohlsson S. 1984(I) – *Restructuring revisited I. Summery and critique of the Gestalt theory of problem solving.*
Scandinavian Journal of Psychology, Vol. 24, 65-78

Polanyi, Michael 2005 – *Tacit Knowing*
In Stehr, N & Grundmann, R. (eds) Knowledge; Critical Concepts II
ISBN 0-415-31738-X

Rubin J. 1994 – *Handbook of usability Testing*
John Wiley & sons, inc.
ISBN 0-471-59403-2

Russo, J. E., Johnson, E.J., & Stephens, D.L. 1989 - *The validity of verbal protocols* -
Memory & cognition, Vol. 17, 759 – 769

Ryan, R.S., & Schooler, J.W. 1998 – *Individual differences in susceptibility to verbal overshadowing*
Applied Cognitive Psychology, 12, S105-S125

Schooler, J. W. & Engstler-Schooler, T. Y. 1990 - *Verbal overshadowing of visual memories: Some things are better left unsaid.*
Cognitive Psychology, 17, 36-71

Schooler, J. W. et Al. 1993 – Thoughts Beyond Words: When language Overshadows Insight
Journal of Experimental psychology: General, Vol. 122, Number 2, 166-183

Schooler, J. W. 2002 – *Verbalization Procedures a Transfer Inappropriate Processing Shift*
Applied Cognitive Psychology, Vol. 16, 989-997

Van der Meij, H 1997 – *The ISTE Approach to Usability Testing*
IEEE Transactions on professional Communication, Vol. 40, No 3

8.1 Orienting background literature

DeShon, R.P., Chang, D., & Weissbein, D.A. 1995 –*Verbal overshadowing effects on Raven's Advanced Progreesive Matrices: Evidence for multidimensional performance determinants*
Intelligence, Vol. 21, 135-155

Ericsson, K.A. & Simon, H.A. 1998 – *Hoe to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking*
Mind, Culture, and activity, Vol. 5, 178-186

Järvinen Pertti 2004 – *On Research Methods*
ISBN 952-99233-1-7

Jørgensen P. S. & Rienecker L 1999 – *Formidlingskontoret: Specielt om specialer*
ISSN 1397-8934

Katagami, S. et Al. 2002 – *The influence of Test-Set Similarity in Verbal Overshadowing*
Applied Cognitive Psychology, Vol. 16, 963-972

Katzenelson Boje 2001 – *Universitets opgaver*
ISBN 87 7706 322 8

Lane, S. M. & Schooler, J. W. 2004 – *Skimming the surface, Verbal Overshadowing of Analogical Retrieval*
Psychological Science, Vol. 15, Number 11, 715 -719

Lund, Anders 2006 – *Komparativ analyse af tænke-høj-testen: Fordele og ulemper ved åbne og lukkede opgaver*
Unpublished dissertation from the IT-University of Copenhagen

Meissner, C. A., & Memon, A. 2002 – *Verbal Overshadowing: A Special Issue Exploring Theoretical and Applied Issues.*
Applied Cognitive Psychology, Vol. 16, 869-872

Newell A., Simon, H.A 1972 – *Human Problem Solving*
ISBN: 13-445403-0

Ohlsson S. 1984(II) – *Restructuring revisited II. An information processing theory of restructuring and insight.*
Scandinavian Journal of Psychology, Vol. 25, 117-129

Ohlsson S. 1992 – *Information-processing explanations of insight and related phenomena*.
In M. Keane & K. Gilhooly (eds.), Advance in the psychology of thinking (vol 1, pp 1-44).
London OK: Harvester Wheatsheaf
ISBN: 0745009816

Schooler, J. W. et Al. 1997 – *At a loss from words: Verbal overshadowing of perpetual memories.*
The psychology of learning and motivation, Vol. 37, 291-340

Schooler, J, W. and Melcher, J.M. 1995 – *The Ineffability of Insight*
The Creative Cognition Approach edited by Steven M. Smith, Thomas B. Ward and Ronald A. Finke
ISBN 0-262-19354-X

Van Someren M. W., Barnard Y.F., Sandberg J.A.C 1994 – *The think aloud method*
ISBN 0-12-714270-3

Appendix A – insight and non insight problems

Source: Schooler et. at. 1993

Appendix A

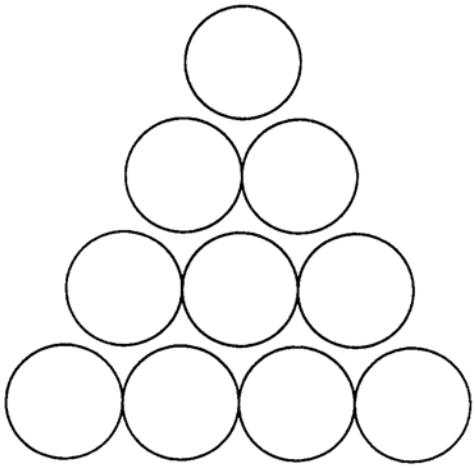
Insight Problems

1. (Experiments 1–4) Show how you can make the triangle below point downward by moving only three of the circles. [see Figure A1.]

2. (Experiments 1–4) A prisoner was attempting to escape from a tower. He found in his cell a rope that was half long enough to permit him to reach ground safely. He divided the rope in half, tied the two parts together, and escaped. How could he have done this?

Solution: He unraveled the rope and tied the two pieces together.

3. (Experiments 1–4) A dealer in antique coins got an offer to buy a beautiful bronze coin. The coin had an emperor's head on one side and the date 544 B.C. stamped on the other. The dealer examined the coin, but instead of buying it, he called the police. Why?



Solution:

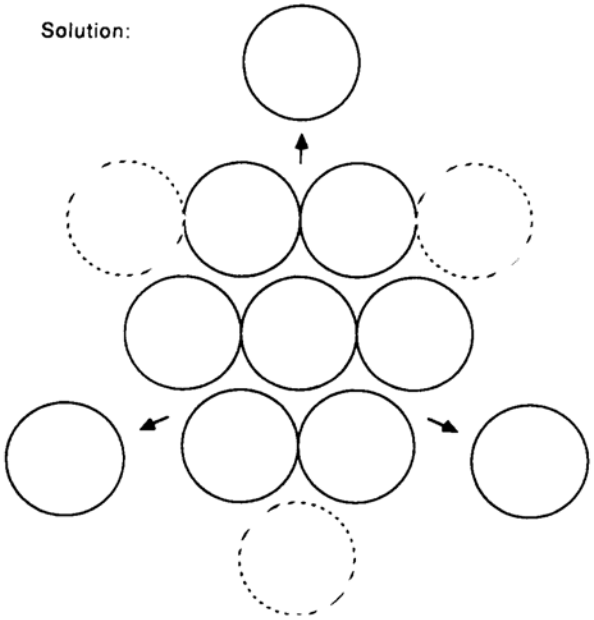
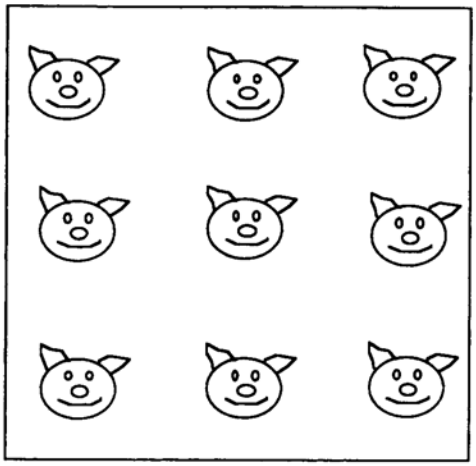


Figure A1. Diagram and solution for the "Triangle" problem.



Solution:

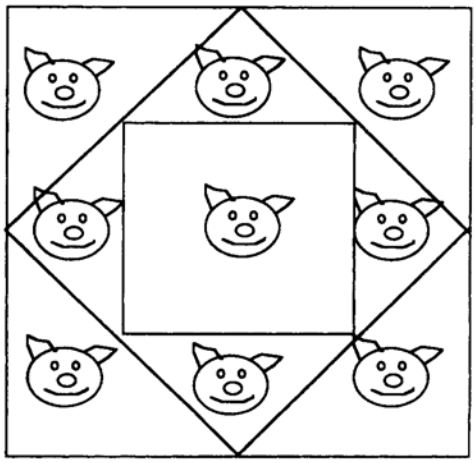


Figure A2. Diagram and solution for the "Pigs in a pen" problem.

Solution: In 544 B.C. Christ had not been born, so a coin from that time would not be marked "B.C." (before Christ).

4. (Experiments 1 and 2) Nine pigs are kept in a square pen. Build two more square enclosures that would put each pig in a pen by itself [See Figure A2.]

5. (Experiments 1 and 2) Describe how to cut a hole in a 3-×-5-in. card that is big enough for you to put your head through.

Solution: First cut a spiral path from the outside of the card to the inside. Then cut a long slit down the middle of the spiral strip leaving the ends of the strip intact. A number of similar variations to this solution were also accepted.

6. (Experiments 1 and 2) A giant inverted steel pyramid is perfectly balanced on its point. Any movement of the pyramid will cause it to topple over. Underneath the pyramid is a $100 bill. How would you remove the bill without disturbing the pyramid?

Solution: Burn or tear the dollar bill.

7. (Experiment 1 only) Water lilies double in area every 24 hr. At the beginning of the summer, there is one water lily on the lake. It takes 60 days for the lake to become completely covered with water lilies. On which day is the lake half-covered?

Solution: The lake is half-covered on the 59th day.

Appendix B

Noninsight Problems Used in Experiments 3 and 4

1. (Practice) Mary won't eat fish or spinach, Sally won't eat fish or green beans, Steve won't eat shrimp or potatoes, Alice won't eat beef or tomatoes, and Jim won't eat fish or tomatoes. If you are willing to give such a bunch of fussy eaters a dinner party, which items from the following list can you serve: green beans, creamed codfish, roast beef, roast chicken, celery, and lettuce.

Solution: roast chicken, celery, and lettuce.

2. Three cards from an ordinary deck are lying on a table, face down. The following information (for some peculiar reason) is known about those three cards (all the information below refers to the same three cards):

- To the left of a queen there is a jack
- To the left of a spade there is a diamond
- To the right of a heart there is a king
- To the right of a king there is a spade

Can you assign the proper suit to each picture card?

Solution: jack of hearts, king of diamonds, queen of spades.

3. The police were convinced that either A, B, C, or D had committed a crime. Each of the suspects, in turn, made a statement, but only one of the four statements was true.

- A said, "I didn't do it."
- B said, "A is lying."
- C said, "B is lying."
- D said, "B did it."

Who is telling the truth? and Who committed the crime?

Solution: B is telling the truth, and A committed the crime.

4. There are four coins—two heavier coins of equal weight and two lighter coins of equal weight, all of which are indistinguishable in appearance or by touch (you cannot tell them apart by looking at them or holding them). How can you tell which coins are the heavy ones and which coins are the light ones in two weighings on a balance scale? (You may only use the scale twice.)

Solution: Begin by placing one coin on each side of the scale. If they do not balance, then you have already identified one heavy and one light coin. Repeating the procedure with the remaining two coins will identify the other light and heavy coins. If the initial two coins balance, simply remove one of the coins and replace it with one of the remaining coins. This weighing will provide the remaining information needed to determine which coins are heavy and which are light.

Received January 13, 1992
Revision received September 14, 1992
Accepted September 24, 1992 ■

Appendix B – Response from J. Schooler on individual data

Hi René

I am afraid I no longer have the raw data for that study but your question is a good one. You might try writing mark beeman who has bee working with insight like remote associate problems. I have cc'd him your message.

best,

Jonathan

>Dear Professor Schooler
>
>First of all – thanks for your quick answer on my initial mail October last year, and for the article >itself – My apologies for not thanking you right away.
>
>After reading the article and research on verbal overshadowing done by you and others, I have >stumbled over a question that the data from "Thoughts beyond words:…" might be able to answer >in relations to a persons chance to be affected by verbal overshadowing in a Think-Aloud >situation.

>I am therefore hoping that the data is still available and in a form so that it can be used in relation >to my question.
>
>Now I realise that the experiments conducted leaves no clue to what causes, how little or how >much a person would suffer from the effect of verbal overshadowing in a given situation (like in >"Whom Do Words Hurt?...." by Ryan and yourself). The question is whether or not the data >indicates a relation between the difficult level of each insight problem and a person's ability to >solve an easier problem, derived from the outcome of a more difficult problem.
>
>In other words – if a person solved the insight problem ranged as the must difficult, would he/she >then be more likely to solve a lower ranked insight problem, than a person that couldn't solve the >same problem?
>
>Sincerely
>
>René Ginger-Mortensen
>Copenhagen Business School

Jonathan W. Schooler, Ph.D.

Professor of Psychology
Canada Research Chair in Social Cognitive Science

PLEASE NOTE CHANGE OF ADDRESS

Department of Psychology

3410 Kenny Hall
2136 West Mall Ave
University of British Columbia
Vancouver BC V6t 1z4

tel: 604 822-2851
fax: 604-822-6923

Appendix C – Ritzau

Copied from www.eb.dk on the 4 of December 2006.

Farligt med håndfri mobil i trafikken

Det er også farligt at tale i håndfri mobiltelefon, når man kører bil. Ny rapport sammenligner brug af mobiltelefon med kørsel under alkoholpåvirkning

/ritzau/ - 16:22 - 04. dec. 2006

Risikoen for trafikulykker stiger, hvis man taler i mobiltelefon, mens man kører bil - uanset om telefonen er håndholdt eller håndfri.

En ny rapport fra Danmarks Transportforskning fastslår, at det ikke er ufarligt at tale i telefon under kørsel, blot fordi man følger loven og benytter et håndfrit sæt. For selv om man på den måde undgår at tage hænderne fra rattet, er man stadig mere uopmærksom.

"Effekten af mental belastning og visuel informationsbehandling er relateret til samtalefasen og er lige stor ved begge typer telefoner. De typiske effekter er reduceret opmærksomhed, forøget reaktionstid og mere usikker kørsel", står der således i rapporten.

Rapporten viser også, at betjening af radio og musikanlæg er lige så distraherende som at tale i mobiltelefon, og at forringelserne i kørslen, når man taler i mobiltelefon, kan sammenlignes med kørsel under alkoholpåvirkning omkring promillegrænsen på 0,5.