Proceeding studies on behavior - not only a challenge for professional tools

Pascal Mangold
Graf von Deym Str. 5
D-94424 Arnstorf
pascal.mangold@behavioral-research.com

Abstract: The following insights are based on my company's long term empirical experience as system developer in the field of behavioral research. The paper discusses several aspects of data collection and analysis in day to day studies on behavior. It points out the necessity of using specialized software tools in behavioral research. It shows why video recordings are very beneficial for analysis and not only for documentation purpose. It discusses the advantages of using structured coding schemas instead of taking notes only. Finally the possibilities of the INTERACT software tool environment are sketched.

Keywords: Video analysis, physiologic data analysis, usability test, benefits of video, event logging, eye tracking, screen capturing, user interfaces, usability evaluation, software INTERACT, software DataView

If I write in terms of "we", I like to refer to our common experience as a company.

1. Using software tools in behavioral studies is more than a good idea

Doing a behavioral study seems to be simple while looking at the following steps:

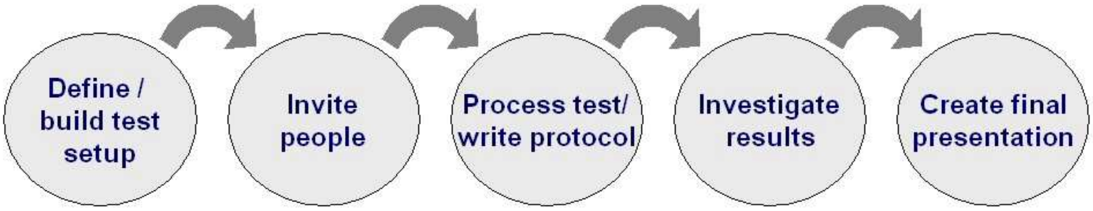

Define / build test setup
Invite people
Process test/ write protocol
Investigate results
Create final presentation

Figure 1. Main process of behavioral research

The "only" thing that has to be done is following the steps shown in Figure 1. But the reality looks different: Most people we met over the last decade in this research area are neither electrical engineers, information scientists nor computer specialists. They are psychologists, designers, sociologists, ergonomists or any kind of similar non technical oriented professionals. Not only for them doing a behavioral study means much more than that simple view! A lot of questions have to be answered before, during and after the test:

- How can the scenario be set up technically?
  Which plugs fits into which device (audio / video and computer technology knowledge necessary). How can a video with a duration of more than 90 minutes be recorded or how can a video camera be remotely controlled (limitation of standard equipment / up to date knowledge about commercially available equipment)…
- How is the collected data going to be processed further?
  How can the videos get into a computer. How can the data from system X be retrieved (a used device may have no data export except to its own analysis software)...
- How is the data going to be integrated?
  How to deal with a different density of data? Data might be recorded on different time scales (physiologic data in e.g. milliseconds / video in 30 or 25 frames per second). How can a sync start point of all data sources be identified? Is a software available that can handle X-thousand data values[1] from e.g. physiologic or eye tracking recordings? How can a hand written protocol be synchronized to a video?…
- What should the analysis process look like and who is involved?
  If videos are stored on classical video tapes a limitation might be the availability of equipment to review those videos. Several people might not be able to watch the same or different videos in parallel at their own office. Thus video analysis gets a serious bottleneck. Which tools (software / paper and pencil) should be used for collecting the findings / codings / transcriptions and in which format (with / without time stamps, based on absolute day time or relative duration only, as structured lists or free textual annotations etc.)? Is it satisfactory to base the findings on the accuracy of a stopwatch (and the person who is handling the watch) or on a video players' mechanical counter? How is this data going to be shared in a team (is everyone working on the same document or on their own copies)? How can the results be manipulated and integrated for comparability during interpretation or for reporting (graphs, statistical figures, any kind of specialized visualization)?…

All of the above fragmentary listed items show that behavioral research can not be seen as simple as displayed in Figure 1. It demonstrates, that the most problems arise not *in* the main process stages (Figure 1: e.g. doing a 1 hour video recording or writing a protocol) but *during* the things that have to be done in between the stages (e.g. putting different kinds of data sources together, reformatting time scales, collecting findings in a structured way…). The process approximating this reality in a better way is visualized in Figure 2.

[1] The typically used Microsoft Excel can handle 65536 data lines only…

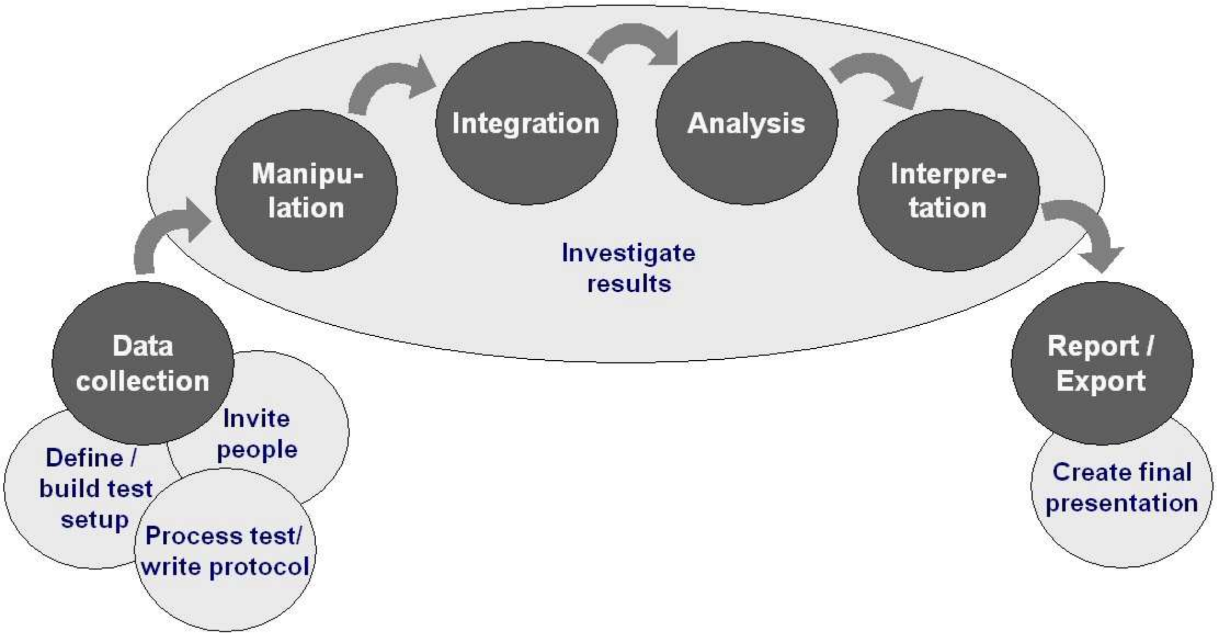


Figure 2. The "real" process of behavioral research

Figure 2 also gives an anticipation of what happens, if one finds out that the test setup was incorrect, some information is missing or has been accidentally restructured in a wrong way. The effects are shown in Figure 3. In the worst case, one finds out that the results are not reliable or somehow curious: Thus you have to start all over again, not only "collect some more data" but to go through every single step as shown in Figure 1 again (depending on how big the problem is). This is what takes most of your time and is a real pain in the neck.

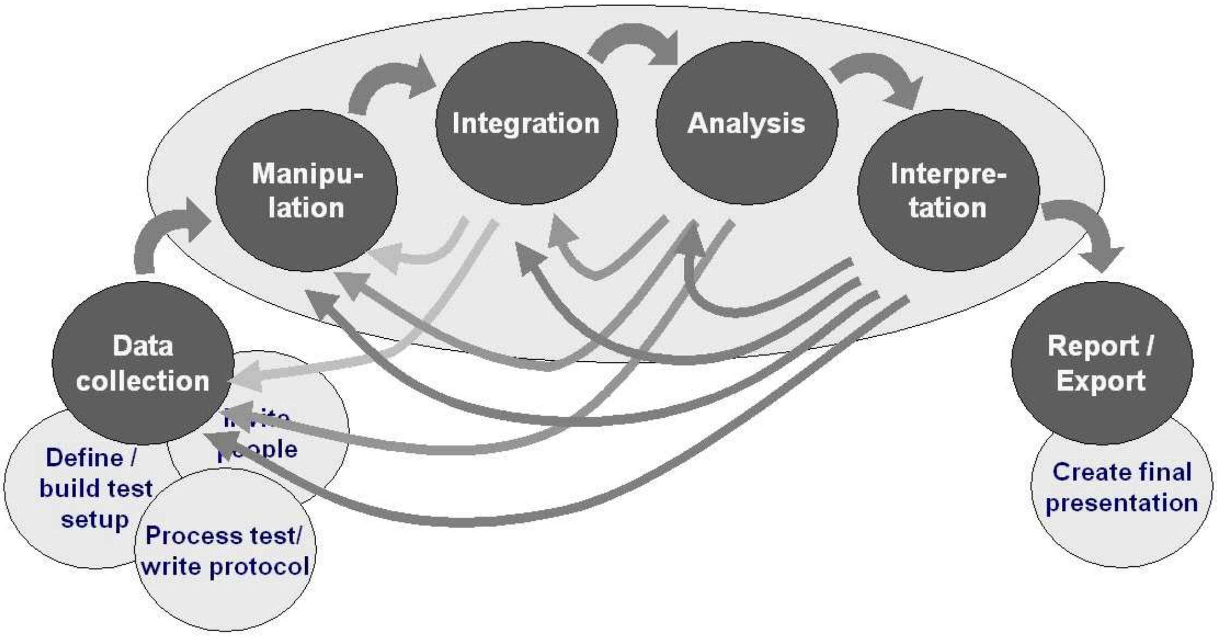

Integration
Analysis
Manipu-lation
Interpre-tation
Data collection
Report / Export
Define / build test setup
Process test/ write protocol
Create final presentation

Figure 3. The worse cases in a behavioral research study

Because of all of the above, we claim that the field of “behavioral research” (whether expressed in usability tests, psychological studies, studies on ergonomics etc.) needs to have specialized tools. By using such tools, a lot of problems as discussed at the beginning of this chapter could be avoided or at least reduced to a minimum. A so called “chain of tools” (tools that “understand” each other without too much direct intervention of the user) can help to:

- Collect data in an easy but error free and structured way
- Manipulate the data at any time so that it is comparable to one another
- Correlate results and source data with a simple mouse click (e.g. “show me the position in the video and physiologic data that matches a specific comment made by an human observer”)
- Transform and export the data easily to any other program

This helps to:

- Save time
- and, what I think is the most important aspect, improve the quality of your results.

1.1. Structured data collection and quality of results

Even without using tools one can create interesting results by simply watching the scene and writing down comments/findings using paper and pencil. This is a common practice, done by “experts”. They have obtained an subconscious knowledge about “right” or “wrong” (in a general meaning) through their long term experience. Experts often can not explain in a formal way how they come up with their findings. This results in an important question: How can the quality of results that have been identified by an expert be assured?

A methodology has to fulfill at least the following criteria to be regarded as “scientific” [1]:

- Objectivity: The results have to be revisable
- Repeatability: By repeating the test under the same conditions the results have to be the same
- Comprehensibility: The way of finding the results and the results itself have to be easily understandable

Example: A study on web usability with nine independent “usability experts” showed that their findings were in fact completely different. In total the experts found 174 different usability problems. 61% of all problems have been mentioned by only a single person! Only 18% of all problems have been reported by at least a third of all experts. This also resulted into the fact that each of the experts only identified 31% in average of all relevant results (the above 61%) [1]. This is an alerting issue.

Also the requirements of objectivity and comprehensibility are not fulfilled satisfactory by the method of using “expert knowledge”. Thus, measuring the quality of results based on “expert knowledge” is far from easy.

Secondly, the question arises on how this knowledge can be transferred, shared and combined with knowledge of others to improve and create new insights? A detailed description about knowledge and so called “knowledge management” can be found in [2].

Regarding those questions, formalizing and structuring the way of achieving interesting findings becomes very important.

1.2. Taking notes and transcribing versus coding

Among others, the following data collection methods are widely used during observation:

- Taking notes: The observer is writing down things that he/she thinks being of interest. This can be anything between not interpreted “raw data” (e.g. simply what a person said) and already worked out conclusions.
- Transcribing: This method is used to exactly write down what people say and/or what happens, in form of a well defined syntax (e.g. “[exhalation] I don’t understand this [sigh] [~] oh my gosh!”). As one can imagine, transcribing can get very complex. Some schemes are shown in [3,4,5,6,7].
- Coding: During coding already well defined labels in form of an identifier or a number are attached to certain points or periods in time. “Coding schemas” (also often called “category schemas”) are mostly hierarchical and can get quite complex. E.g. someone saying positive things about the color of an optical stimulus: “[EXPRESSION] – [verbal] / [CONTENTS] – [color] / [QUALITY] – [positive] …”. Some schemes are shown in [8,9].

All of those methods have their pros and cons. Some of them, in regard to analysis of behavior, are shown in the following table:

Method	***Pro***	***Con***
Taking notes	+ Very fast	- Because of lacking in structure it is often unclear in later analysis what a specific note “really meant” - How the notes have been originated is not necessarily comprehensible nor objective
Transcribing	+ Automated further analysis possible, if the transcription has been made very structured + Very good for textual context analysis + Objective + Comprehensible	- Time consuming - Waste of time if it just reflects what obviously can/could be listened from the video again. - Not even the best method in case of observing behavior - Requires a lot of research to define a good transcription syntax
Coding	+ Very fast + Further automated analysis possible + Objective + Comprehensible	- Requires a lot of research to define a good coding scheme

The differences between really structured transcription and coding are not too big. But in reality, a common idea exists that writing down things in a more or less structured way would overcome the need of first having to develop a good coding scheme. “We can’t use codes because we don’t know what’s happening!”. If this would be true then no study could ever be compared to any other because they would have nothing in common! From our experience, building up a coding scheme is always possible. At least a rough one that gets refined during the studies.

We have seen students doing slave jobs by writing everything down they see or hear on a video tape. Just for the supervisor to get a MS Word document which reflects in words only a fragment of the recorded reality. No gesture, posture, facial expression, no possibility to hear the real stream of speech or the words between the line. Secondly, until this process stage, no interpretation has taken place at all. Thus, the task of getting results has still to be done - but now based on an even more fuzzy data set.

In our experience the above listed pros of coding are the most important aspects during day to day studies on behavior.

- Fast: Analysis of x hours of video recording takes at least x hours in general. Thus time is a critical factor!
- Comprehensible: To train coders successfully it is necessary that everyone clearly understands *when* to give *which* code.
- Objective: In terms of quality assurance it is indispensable to be able to calculate statistical values that prove that the coders have a common understanding and that they see the same things at the same time (inter rater reliability).
- Possible automation: Analytical analysis on large data sets and comparing parts of the data requires automation. This prevents from making errors through manual processing and helps saving a lot of time.

2. Live observation versus video analysis

Trying to code a video tape without tool support is very difficult and time consuming. An easy to process and ongoing quality control is missing (going back and forth between written down codings and video(s)). This makes the process for the coder somehow unconfident. At least the results are very hard to prove, because finding lots of short fragments on video tapes during quality review becomes a real imposition. This is why lots of people tried to work with video but found it to be too time consuming.

In cognition science it is well known that humans have a limited perception. Therefore observing anything means that only a fragment of “what really happens” can be detected at a time. It is also known that different people watching the same scene focus on different things at the same time. That means during observation:

- a lot of information is not perceived.
- you don’t know whether you missed the interesting or the not interesting things.
- you don’t know if you focused on the “right” things.

Example: by observing a live discussion between two or more people it is nearly impossible to “listen - understand - interpret - transform (write down in notes)” what these persons say and focus on their body language (postures, gestures and gaze) at the same time. Maybe the body language is saying something very important, which even might be in contrast to the spoken word at all[2]. To prevent getting overwhelmed with information, a lot of information is simply filtered out by the human cognition system. But not only this loss of information is problematic. The possibly wrong interpretation is a serious source of error, too. If lots of things happen very fast in the observed scene *and* the observer is under pressure to write down findings simultaneously, a summary of notes may be produced that is not correctly reflecting what’s going on at all. This situation is getting more difficult by the fact that observers normally try to correct their previously made notes during the observation session. Thus, they are missing even more information and start loosing track of the contents at all. During years of giving consultancy to various projects in this field, all of the above has been observed to be a common behavior. I do not claim that live observation is not working at all. By using a simple observation scheme, live

[2] Lots of studies on e.g. politicians speaking have been done, showing this effect clearly.

observation can be helpful and reliable. But observing behavior and getting reliable results in more complex situations (several objects to observe, different items to focus on) is very difficult.

Live observation obviously does not fulfill all of the above mentioned criteria for a "scientific method". At least there is no exact repetition of a specific test possible.

If a method does not allow repeatability then another way of getting reliability could be used: A huge amount of material has to be processed to get a statistical significance, thus letting the results be regarded as trustworthy anyhow. Now, here is the practical problem: Besides the fact that the total costs are always a limiting factor, it is mostly very difficult to find a lot of test objects and/or the time to proceed all those tests. To overcome this situation a typical behavior has been identified through our experience: "We are videotaping all of our test sessions - just in case…". "But we never watch them again.", "This is too time consuming". I believe that nearly everyone who ever did video analysis without a focused methodology and specialized video analysis tools will agree to that.

But using video recordings can be a brilliant idea in behavioral studies: You can replay, slow down and pause the "reviewed reality" how often and in which ever way you want! Allowing you to make annotations and codings, go into detail, correct the collected information at any time, at any speed of the reviewed video. This allows for the focusing on different aspects of behavior in several turns of the video without getting overloaded with information and tasks. The repeatability mentioned above is definitively given. Also the video material can be used to achieve comprehensibility and make the results revisable. How tools can support this video analysis process, to minimize the time spent and maximize the outcome, will be shown in the following chapters.

3. Video analysis tools

Some software tools for video analysis are on the market today. But in our experience, to be really a valuable software tool in behavioral research, it needs to integrate at least the following different methods:

- Qualitative + Quantitative
- Predefined + Explorative
- Structured + Free annotation

Especially in regard to the arguments mentioned above, about developing and using a coding scheme versus annotating, the combination of *structured and free* data collection is indispensable.

In [10] some commercially available software systems for video analysis have been evaluated: "By exploring the facilities and features of the INTERACT software system in conditions that replicate the future design studies, we have ascertained that it meets our requirements to a high degree." As a result "…the use of INTERACT was found to be effective and timesaving and appears to offer a significant advantage to the analyst and hence the efficiency of the research process.". Based on their study, I like to sketch a few facts of the INTERACT system and some of its add-ons.

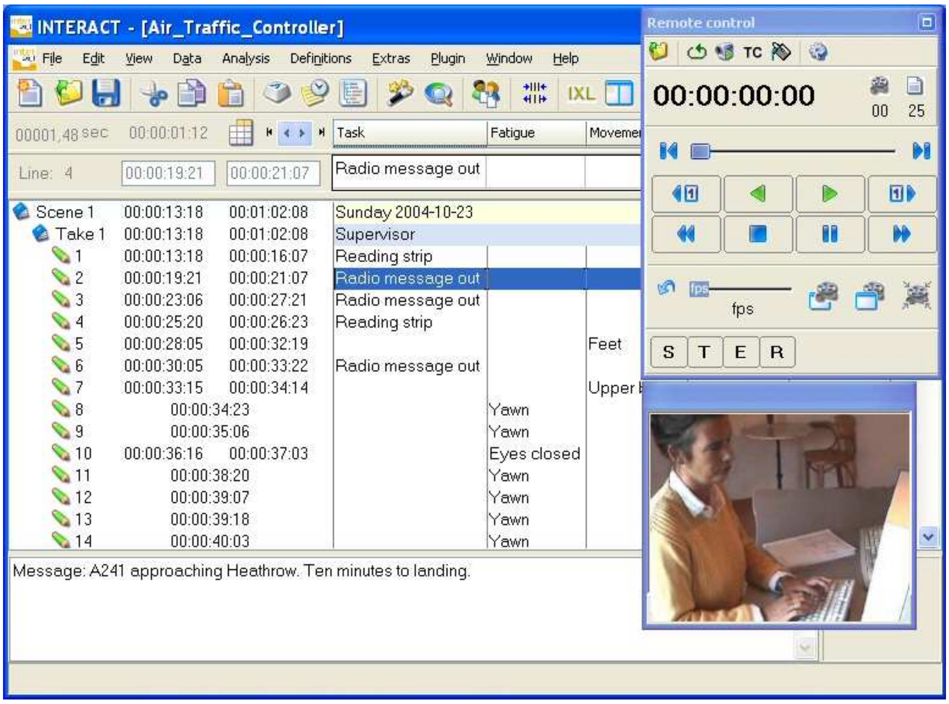
With INTERACT [11] the user can collect data in different ways and at any time during the analysis process:

- Enter a predefined code with a single key stroke
- Enter any kind of textual new code
- Enter any kind of free textual annotation

This information is always stored in combination with a start and an end point in time. Thus the exact correlation of data and video is given at any time. Secondly the time format used is a so called standard time code (hours : minutes : seconds : video frames). This allows to identify each single picture of the video and is essential in studies on emotion, facial expressions, gesture and posture. A screen-shot of the INTERACT user interface is shown in Figure 4. It demonstrates the collection of data (each line is a so called *event* with start and end time information, that contains *codes* in so called *categories* [columns]).

INTERACT can deal with a practically unlimited number of video sources simultaneously and control them synchronously. This is necessary, if several video sources have been recorded simultaneously and there is no possibility to mix them into a single video or the original size and quality of the different videos should be preserved each.

INTERACT offers a so called plug-in technology allowing the user to add any functionality that is not yet integrated (special data import, export or manipulation routines). Hence making INTERACT a tool of practically unlimited possibilities. The collected data can not only be exported for any other purpose but also any other kind of data can be imported into INTERACT. At least by writing a special import filter (plug-in). Thus, data that has been manually collected within other systems or data that has automatically been created by the test environment, can be used to enrich the analysis process.



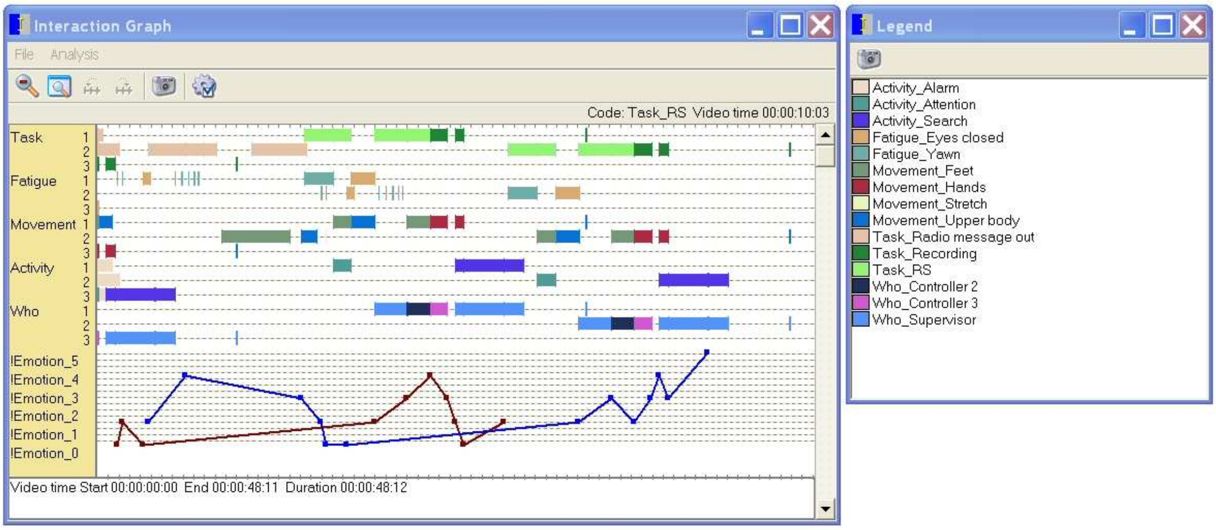
Line: 4 | 00:00:19:21 | 00:00:21:07 | Radio message out

			Task	Fatigue	Movemer
Scene 1	00:00:13:18	00:01:02:08	Sunday 2004-10-23		
Take 1	00:00:13:18	00:01:02:08	Supervisor		
1	00:00:13:18	00:00:16:07	Reading strip		
2	00:00:19:21	00:00:21:07	Radio message out		
3	00:00:23:06	00:00:27:21	Radio message out		
4	00:00:25:20	00:00:26:23	Reading strip		
5	00:00:28:05	00:00:32:19			Feet
6	00:00:30:05	00:00:33:22	Radio message out		
7	00:00:33:15	00:00:34:14			Upper
8	00:00:34:23			Yawn	
9	00:00:35:06			Yawn	
10	00:00:36:16	00:00:37:03		Eyes closed	
11	00:00:38:20			Yawn	
12	00:00:39:07			Yawn	
13	00:00:39:18			Yawn	
14	00:00:40:03			Yawn	

Message: A241 approaching Heathrow. Ten minutes to landing.

Figure 4. INTERACT user interface

INTERACT offers a lot of analysis functionalities. Among others, the “inter rater reliability” can be calculated easily, giving an international well accepted measure of data quality (did the observers see the same things at the same time?). The sequence analysis shows which codes follow on which codes and with what statistical significance. A standard literature on that is [12]. Statistics such as frequency, duration, percentage over time, variance and standard deviation are among others integrated, too. A special time line plot of the coded data is shown in Figure 5.



Code: Task_RS Video time 00:00:10:03

Video time Start 00:00:00:00 End 00:00:48:11 Duration 00:00:48:12

Figure 5. Time line visualization of coded data

We call data collected in INTERACT “sparse” (some events from time to time). In contrast to that, we call data “high frequency” if a lot of values are generated in short time intervals (such as physiologic data, recorded every 10th of a second). Importing or integrating such high frequency data into INTERACT would not make too much sense. That’s why a separate data presentation tool named “DataView” [11] has been developed. This allows for the visualization of numerical data in an practically unlimited number of graphs. This software tool is automatically time synchronizing with a running INTERACT. Whenever the video is moved to another position from within INTERACT, the data graphs are moving, too.
An Impression of a complex analysis desktop with three video windows (test person, his environment and his screen contents) several bio data visualizations and the time plot of coded data is given in Figure 6. The Screen has a resolution of 1400 * 1050 pixels, taken on an Acer TravelMate 6000 laptop computer during an analysis session. As one can see there is not enough space to see everything clearly. Especially a screen recording video needs to be reviewed in original full recording size (here 1024*786). Otherwise the image gets too fuzzy. Therefore, attaching a second display to the computer during analysis is highly recommended.

Those unlimited data collection possibilities, the practically unlimited visualization of synchronized audio, video and data files make INTERACT and its add on tools a standard application in behavioral research.

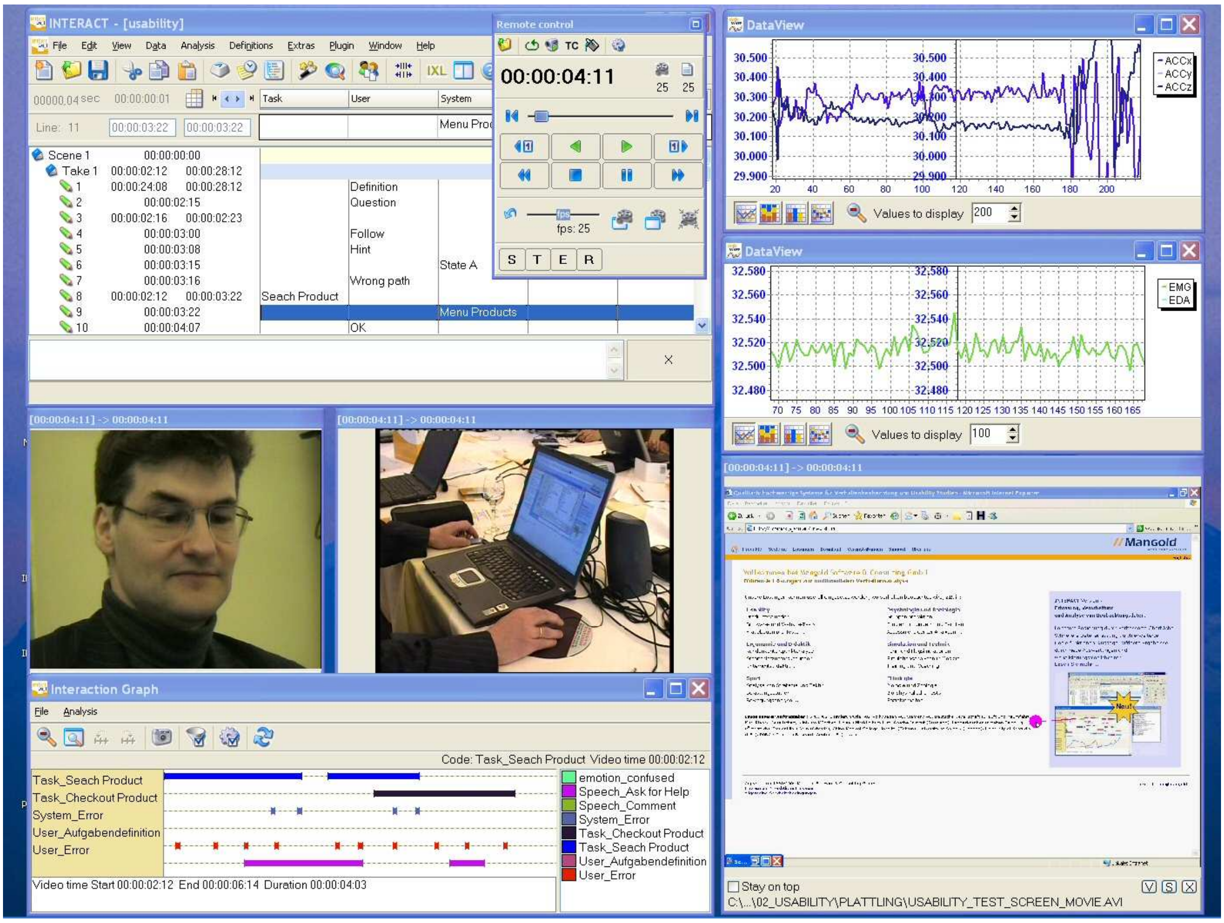


Figure 6. Impression of a complex analysis desktop with several videos and bio data

3.1. Measuring the return on invest

Looking at the impact of time on the quality of result is a good starting point for measuring the return on invest. I would like to define that the feeling of having exactly enough time for a task is a state that does not exist in real day to day work. If people feel to have plenty of time then they mostly:

a) Try to put more things in the current task than initially requested.
b) Simply stop the task, when “they did their job”.

Now, the ones in case a) have no time to take care about quality of results because they get messed up in their self produced overload of tasks/data/results. Which is the same for the ones who did not have enough time at all, right from the start. The ones of case b) have no time left or, if there is time left, are satisfied with “doing their job”.

From all of the written above and our experience, we know that using tools definitely helps to save a tremendous amount of time. But we also know, this time is no spare time. It will be occupied by new tasks immediately. That’s why users never have “the feeling” to have saved time, even by using specialized tools – except they did exactly the same things before without using those tools. But this mostly does not happen: We found out that people using tools who did the same or similar things before manually, say, “its much easier and makes more fun – but it doesn’t save any time” – Yes, that’s true, because (if we observe them) they are doing completely different things. The main tasks don’t take the main parts of time any more. That’s why there is plenty of time for doing other things (getting into more detail, formatting the data/documents, "playing" with the data and charts etc.)

Our finding is, that the factor time definitively has an impact on the quality of results.

- If there is less time the quality nearly always suffers. Especially in studies where highly accurate coding is necessary (facial expressions, detecting emotional states, observing the interaction of several objects in group discussions, identifying things that follow / have an interdependence to one another).
- If there is (would be) plenty of time it is sometimes used to improve quality of results. That's the ideal case. It also appears that this time is used for things that have no influence on the original intended task at all.

My conclusion is, that using professional tools can have a very good return on invest but the return can not necessarily be seen immediately.

What we can identify always is a "soft impact", if the user has done video analysis manually before ("Ahh. This is much better then before!"). This reduction of stress definitively has positive effects on other things in the project. It is also true, that some scenarios could not be done at all without tools (complex frame accurate video analysis / analyzing different data sources in sync).

However, measuring the return on invest is very difficult and strongly depends on the scenario where video analysis and tool support takes place.

4. Discussion

As we know, as an international operating company, the variance of studies on behavior is extremely high. Technical equipment used, methodology of data collection, complexity of hypothesis', knowledge level of observers etc.. This makes it difficult to create a "matrix" that shows under which circumstances what method is the best. However, a good point to start though would be to systematically investigate into the following:

What are the differences between live and video based observation under what circumstances? Where is the necessary changeover from one method to the other in terms of costs and quality?

Hopefully this paper could show the need and advantages for systematic data collection and video based analysis.

5. Recommendations

To understand the complexity of usability testing in relation to usability engineering, some recommended further articles are [13,14,15,16]. Information on information system design in general is given in [17]. Useful hints about Project Management in regard to chapter 1 can be found in [18]. For understanding the problems of knowledge management, especially *sharing knowledge* in regard to chapter 2, reading [19] is recommended.

6. Acknowledgment

I would like to thank my colleagues Petra Vegter and Reinhard Grassl for reviewing this paper.

7. About the Author

Pascal Mangold has a diploma in information science and has been active in the IT sector since 1989. He has a company of his own developing IT systems for scientific and industrial applications focused on behavioral research. He is a book author, runs training seminars, lectures at universities and is a much sought-after industrial consultant.

8. References

1. Blume, M., Stokar, D., Seewald, F. (2005): Usability Evaluation: Egal wer's macht? Ein schweizer Fallbeispiel, Usability Professionals 2005, German Chapter der Usability Professionals Association e.V., Hassenzahl, Marc (Herausgeber)
2. Denning, S. (1998): What is knowledge management? (A background paper to the World Development Report 1998). www.stevedenning.com Available at: http://www.stevedenning.com/knowledge.htm [visited 2005-09-01]
3. Verbmobil: www.bas.uni-muenchen.de/Forschung/Verbmobil/VMTrlex2d.html, [visited 2005-09-01]
4. SpeechDat: www.speechdat.org/speechdat/deliverables/public/SD132V24.PDF, [visited 2005-09-01]
5. Bayerisches Archiv für Sprachsignale: http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html [visited 2005-09-01]
6. SmartKom: www.bas.uni-muenchen.de/Forschung/SmartKom/Konengl/engltrans/engltrans.html, [visited 2005-09-01]
7. MATE: www.ims.uni-stuttgart.de/projekte/mate/mdag/, [visited 2005-09-01]
8. Steininger, S., Lindemann, B. & Paetzold, T. (2001): Labeling of Gestures in SmartKom - The Coding System. In I. Wachsmuth & T. Sowa (Eds.): Gesture and Sign Languages in Human-Computer Interaction, International Gesture Workshop 2001, London, UK. Berlin: Springer, pp. 215-227
9. Steininger, S. (2000): Transliteration of Language and Labeling of Emotion and Gestures in SMARTKOM. Workshop. Proc. of the Second International Conference on Language Resources and Evaluation: Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources. Athens, Greece, pp. 49-51
10. Candy, L., Bilda, Z., Maher, M.L., Gero, J.S. (2004): Evaluating Software Support for Video Data Capture and Analysis in Collaborative Design Studies, University of Sydney , Key Centre of Design Computing and Cognition, Presented at "QualIT conference, Australia 2004"
11. Mangold Software & Consulting GmbH: www.behavioral-research.com , [visited 2005-09-01]
12. Bakeman, R., Gottman, J.M. (1997): Observing interaction: An introduction to sequential analysis, 2nd edition, Cambridge University Press, Cambridge UK
13. Holzinger, Andreas (2005): Usability Engineering for Software Developers. Communications of the ACM (ISSN: 0001-0782), Vol. 48, Issue 1 (January 2005), 71-74
14. Holzinger, Andreas (2004): Application of Rapid Prototyping to the User Interface Development for a Virtual Medical Campus. IEEE Software. Vol. 21, Iss. 1, January 2004, 92-99. (ISSN: 0740-7459)
15. Holzinger, Andreas (2003): Experiences with User Centered Development (UCD) for the Front End of a Virtual Medical Campus. In: Jacko, J.; Stephanidis, C.: Human-Computer Interaction, Theory and Practice, Volume 1. Mahwah (NJ): Lawrence Erlbaum (ISBN: 0-8058-4930-0), 123-127
16. Holzinger, Andreas; Geierhofer, Regina; Ackerl, Siegfried; Searle, Gig (2005): CARDIAC at VIEW: The User Centered Development of a new Medical Image Viewer. In: Zara, J.; Sloup J. (eds.) Central European Multimedia and Virtual Reality Conference CEMVRC 2005, 63-68 (ISBN 80-01-03232-9 and available in Euro Graphics Library)
17. Holzinger, Andreas (2001): Basiswissen Multimedia Band 3: Design. Entwicklungstechnische Grundlagen multimedialer Informationssysteme. Würzburg: Vogel (240 pages, ISBN 3-8023-1858-0). www.basiswissen-multimedia.at
18. Mangold, P (2004): IT-Project Management kompakt, 2. Auflage, Spektrum Akademischer Verlag, Elsevier Publishing, Heidelberg – Berlin, (ISBN: 3-8274-1502-0)
19. Denning, S. (1998) What is knowledge management? (A background paper to the World Development Report 1998). www.stevedenning.com Available at: http://www.stevedenning.com/knowledge.htm [visited 2005-08-10]