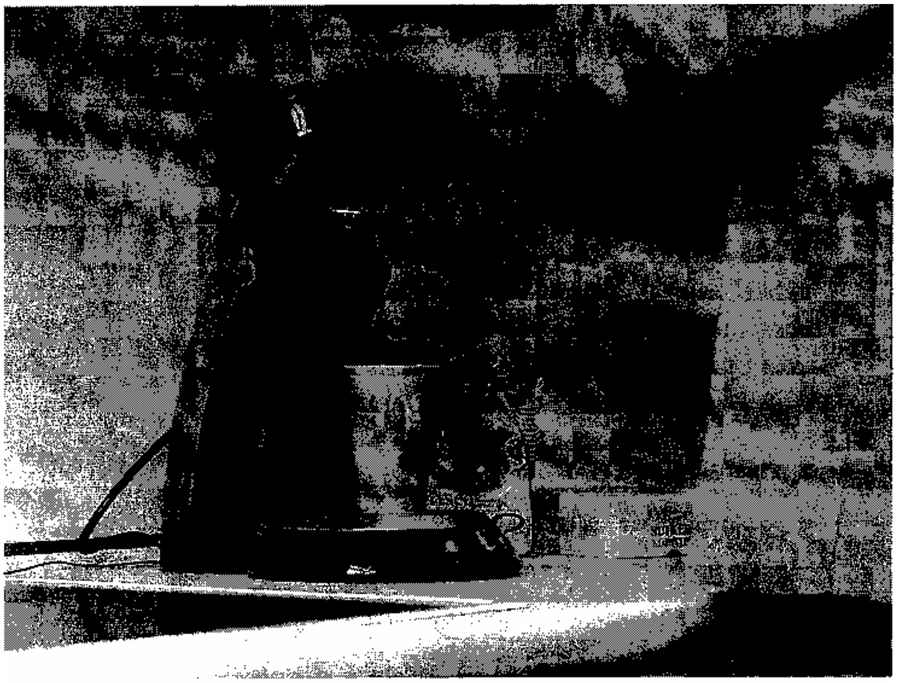
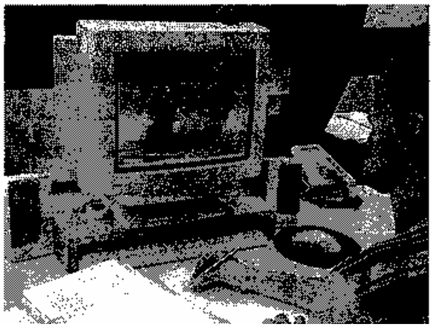
OBSERVATION AS DESIGN TOOL

H. Kanis, MJ. Rooden

January 2005





TU Delft

Delft University of Technology

Faculty of Industrial Design Engineering
Section Applied Ergonomics and Design

0 Introduction

This text summarises in two chapters theoretical considerations, empirical findings and practical guidelines concerning the observation of user-product interaction as a design tool.

In the first chapter, conceptual issues are discussed. Topics involve theoretical perspectives on user-product interaction. The chapter outlines current insights, including some theoretical reflection, which should be seen as supportive for carrying out an observational study (user trial) as outlined in chapter 2.

Chapter 2 briefly highlights the consecutive steps in doing observational research. These steps involve the explication of presumptions, phrasing research questions, the operationalisation in a research set-up including for example the instructions to participants and the role of the researcher in asking questions, the significance of observations in small samples, the analysis, (re)design consequences of the findings, and the communication of the findings.

A number of scientific papers which can be consulted for further information is added to this document (this is indicated throughout the text).

1 User-product interaction

1.1 Usage: some definitions and a representation

Usage is conceived as an ongoing sequence of user activities - i.e. perceptions/cognition and action, with any effort involved - and the concurrent functioning of a product, in context. With respect to the user, this sequence is seen as being constrained by human properties, capacities and limitations (sensory, mental, physical). For the product, featural (form) characteristics are seen to possibly work out as constraints, e.g. weight, dimensions, graphics. See Figure 1, which highlights usage as consecutive user activities within human properties and product characteristics as boundary conditions.

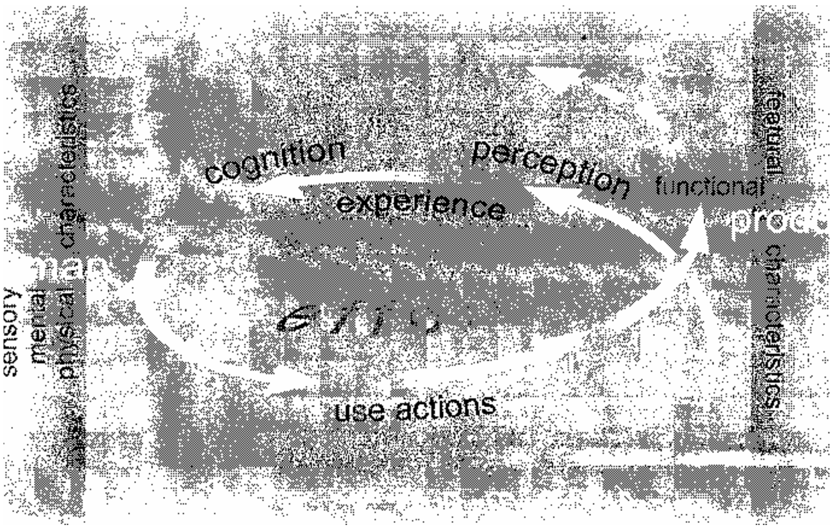


Figure 1. *Usage as consecutive user activities, within human properties and product characteristics as boundary conditions*

The distinction in Figure 1 between perception and cognition is not meant to be clear-cut. Theoretically, perceptive and cognitive activities are difficult to separate (Neisser, 1976). Think of the "mute' function on a remote control. Only people who *knowVnat* there is such a function, without being *acquaintedm^|* the sign, will have trouble in *looking^* it. Purely perceptual reasons for missing the sign would be lack of contrast or minimal graphics. In user trials, participants often demonstrate that perception and cognition can to some extent be separated from each other. Obviously, product characteristics (featural, functional) not being *noticed|^* users as opposed to not being *understood,* may make all the difference in design.

In Figure 1, experience is conceived as the operationalisation of any kind of knowledge related to user-product interaction. Knowledge is seen to be activated by recognition and association, as emerging in some particular context, involving user-product interaction. In this way, experience can be conceived as situated knowledge.
Experience may be precipitated as a mental characteristic, amenable to recognition and association (see above), as well as a psycho-motoric condition, for example typing skill.

"Environment' should be given a wide interpretation. It comprises not only physical circumstances but also bystanders, as well as a "task-environment[7], including any aims to be attained by using a product.

The role of boundary conditions as constraints in user-product interaction is discussed on the basis of a representation of that interaction, see next paragraph.

1.2 Representing user-product interaction

The representation of usage in Figure 1, actually, is a simplification of a more comprehensive representation involving user-product interaction. Since the first publications of such a graphical representation (see e.g. Kanis, 1998; paper available in electronic form via the TU Delft library website), many adaptations haven been made, due to new insights from observational studies into the usage of various products. In this development, Figure 2 gives a recent version of the way to conceptualise user-product interaction.

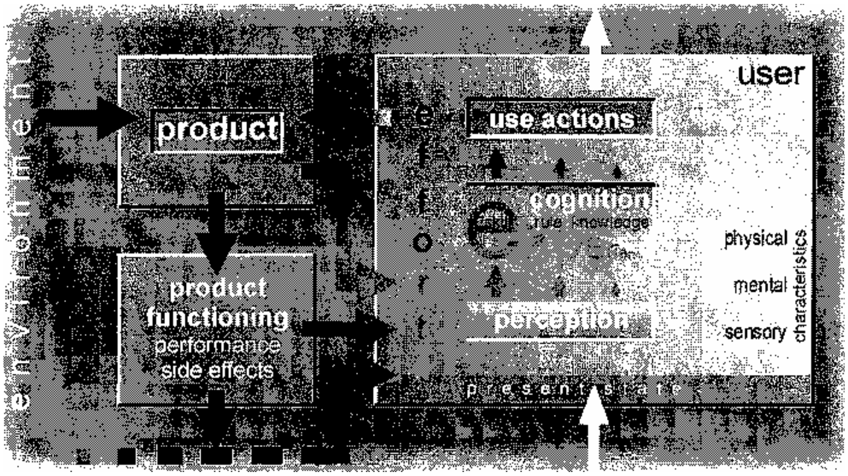
The ins and outs of the graphical representation of user-product interaction are brought together in the legend of Figure 2. The fact that in Figure 2 the position of the product and of the user (human) have been swapped compared to Figure 1, is purely accidental (due to ad hoc considerations).

The only explicitly referenced theoretical perspective is Rasmussen's distinction between various levels of cognitive activity (skill-, rule- and knowledge-based). As global indications, these different levels of cognitive activity appear to accommodate smoothly empirical results as indicated in the legend. Note that a direct coupling of [...•=> perception •=> action ^ perception •=> action <=>...] as derived from ecological studies, is approached in Figure 2 at the lowest level of cognitive involvement, e.g., automatic actions, or skill-based in terms of Rasmussen's model.

In Figure 2, a distinction can be seen between, at the right side, the technique (the product and its functioning), and, at the left side, human behaviour. In this view, the consequences of use actions are conceived as technical, physical phenomena, e.g. emulatable by a robot, without any so-called reification, or dehumanisation, of the human being as the origin of use actions (despite the "mechanistic' impression all those arrows possibly mount up to). Similarly, the fact that featural and functional product characteristics may be perceived and interpreted is no incentive to "humanise' the source of these characteristics, i.e. in conceiving a product as "intelligent', "emotionally aware' and the like.

1.2.1 Further theoretical considerations

As to the user, in Figure 2 the production of user activities is embedded in human characteristics (sensory, mental, physical) and in what is indicated with the collective term "present state'. Only the latter one, capturing changeable, individual conditions at the time (see legend Figure 2), is meant to be directly influential on actual usage. Such a direct link has been left out deliberately in Figure 2 for human characteristics (capacities, limitations). These characteristics are of a general nature rather than being defined in regard to any user-product interaction. Time and again, research has shown that the relevance of these characteristics is largely limited to setting boundary conditions, mainly indicating what users will not do since they are unable to. How users perform activities within these boundaries usually has little to do with their limitations and capacities (Kanis, 1998). Thus, for the actual user activities within the boundaries set by human capacities/limitations, these characteristics of a general nature, as a rule, appear to be uninformative. This finding reflects an observation made by Garfinkel (1991), in the sense that the actual activities of a user cannot be specified by detailing general characteristics or regularities.



(black arrows) Product functioning as the result of a technical/physical process: the outcome of the co-occurrence of a product with use-actions in a physical environment. The functioning of a product is distinguished into
- performance as protection, support, replacement, extension of human activities;
- side-effects: vibration, noise, accidents, ... affecting the user, the environment.
(grey arrows) Product functioning due to user activities, i.e. perception of featural/functional product characteristics, cognition and use actions, including any effort involved.
Perceptions and cognitions may be seen to be triggered by <u>usecues,</u> conceived as meanings given by users to product characteristics in terms of what functionalities a product has and how these functionalities can be activated (Kanis, Rooden and Green, 2000; see Reader PUUE). Usecues, or product's "tellings', are not to be seen as existing "out there' (they are not specified in the figure) but should be conceived as opportunities, to be realised conditionally in relation to individual, situated predispositions of users.
Experience is seen as traded off against cognition in terms of being skill-, rule- or knowledge-based (cf. Rasmussen et al., 1994), with the effort involved as low as possible as the default for perception/cognition. Achieved familiarity with a product tends to "automate' and short-cut cognitive activities, that is: proceeding as much as possible on a skill-based level. This tendency is indicated by the increasing size of the vertical grey arrows between "perception' and "use actions' in going from the right (cognition knowledge-based) to the left (cognition skill-based).
(white arrows] Additional paths involving the environment for possible effects of product functioning on users, and vice versa. "Present state' serves as a collective term, referring to the current physical condition (e.g. wet hands), to sensory conditions (e.g. otherwise wearing glasses), or to being in a particular mood (e.g. irritated, in a hurry, pleased).

No link is indicated between human characteristics (sensory, mental, physical) and user activities; human characteristics involve general, more or less stable human capacities and limitations such as sensory thresholds, memory capacities and exertable forces.

Figure 2. *User-product interaction*

As indicated above (see par. 1.2.1, see also the capture of Figure 2), a link between human characteristics and user activities, is deliberately avoided. In addition, no reference is made in Figure 2 to any presentation of users as rational operators in terms of problem-solving, decision-making and plan-based acting. Similarly, some kind of unfalsifiable concept of representations in people's heads like mental models are not addressed in Figure 2 (let alone the confusion about what such models would represent, see e.g. Wilson and Rutherford, 1989).

In view of the considerable diversity in activities between users, as observed in empirical studies, usage is, primarily, best seen as 'situated' activities, prompted by what people encounter from moment to moment in a partly self-made situation (cf. Suchman, 1987). Here, the notion of 'plan' as possibly able to specify more general human behaviour (cf. Activity Theory in Nardi, 1996) seems largely undirective for actual user activities in context, see Garfinkel's observation (quoted above) that no actual practice is specified by detailing a generality. Hence the absence of the notion of "plan' in Figure 2.
The concept of "situatedness', as counterpart of predictability, thrives on the absence of invariant structures applying across situations (Suchman, p.67). This is not to say that user activities would proceed in an unconditioned way. No way of acting is "baggage-free'. Consider for example the tendency to adopt a low level of cognitive involvement as the default (i.e., automatic actions or skill-based in terms of Rasmussen's model, see capture Figure 2), and the possible fixating role of a one-sided experience (Standaert, 2004; Kanis, 1998).

Finally, the user in Figure 1 and Figure 2 is not identified according to possible personality traits like autonomy, flexibility and intro-/extroversion. Our empirical studies have not yielded indications of their relevance in user product-interaction. Personal traits are primarily "psychological', in-depth personal characteristics, and do not take account of contextuality, such as the practicalities of user-product interaction. Hence, as for human characteristics, Garfinkel's observation (see above) seems to apply for personality traits as well in considering their non-association with observable user activities.

1.3 Design supportive research into user-product interaction: some general considerations

Due to its situatedness as indicated above, actual usage of products often comes as a surprise for designers. And what is worse, in many cases the surprise is an unwelcome one as unanticipated users' operations undermine designed functionalities or may lead to accidents. The difficulty for designers is that user activities (perceptions, cognition, action including the experienced effort) tend to be largely unpredictable on the basis of theoretical considerations or characteristics of a general kind. Data rather than theory appear to be needed in order to anticipate future usage. Therefore, designers must resort to empirical research for a specific design at hand. In order to be design-relevant, that is: to be applicable in a particular product (re)design, empirical findings should meet the following requirements (see Kanis, 2003, added to this document): (/) measurements/observations should be concerned with elements constituting usage such as depicted in Figure 1, whilst these findings (//) should be sufficiently detailed, rather than blurred and contextually stripped summative measures or averages, so that links are specified between user's activities and featural (form) and functional product characteristics at issue, since these characteristics are what product design is all about.

1.3.1 Explorative research

The obvious way to meet both requirements *(I,* //) is by explorative research into user-model/-prototype/-product interaction. The actions that users may undertake in operating a (new) design can be observed. Self-reports (in thinking aloud and concurrent/retrospective interviewing) may shed light on whether featural/functional product characteristics have been perceived, how these characteristics are understood, which role is played by skill, routine or by experience as situated knowledge, and what makes individuals (dis)satisfied in terms of the mental or physical effort experienced. Preferably, this type of explorative research should be carried out on the spot, in as natural a context as possible, which begs for variety as to both users and situations. Reduction of the situatedness of everyday usage calls for explanation when assessing the generality of findings, whilst rock-solid explanations typically tend to be unavailable as this would presume a deep insight into human-product interaction, the absence of which is the very reason to resort to empirical observation of this interaction (cf. Kanis, 2001; added to this document).

1.3.2 Individuality

Usage oriented design is based on insight into activities, experiences and judgements of future users. As outlined above, such insights are typically gained by observation of actions, together with individual's self-reports of her/his reactions to external phenomena (e.g. what is perceived), of internal activities (e.g. ways of reasoning, on the basis of what experience etc.) and of internal references (e.g. to assess activities as (in)convenient, (dis)satisfactory). In other words, usage oriented research is substantially reliant on the exploitation of "subjectivity' in terms of individual accounts. Searching after "objectivity' by reducing appropriate user involvement in the production of data would also reduce or even eliminate the design relevance of such data - see requirements /and ii_r think of the limited significance for ordinary product design of human characteristics such as anthropometries and perceptual thresholds (see above), especially if narrowed down to averages. See also the uninformativeness for designers of mean performance times and the total number of errors observed in a study into voice-operated information services (Kanis, Weegels and Steenbekkers, 1999; paper added to this document).

1.3.3 Observing interactivity

Usage can be seen as situated activities prompted by what users encounter from moment to moment in a partly self-made context (see Figure 2), whilst at the same time being constrained by their individual predispositions (e.g. physical characteristics, experience, use habits). Empirical studies show that, interindividually, diversity in users activities is the rule rather than the exception, given the resolution description required to meet //(see 1.3). The identification of user-product interaction on the basis of users' expressions about internal processes and references relies itself on interaction, think of the meaning or interpretation of terms. Self-reports will at least partly rely on individual frames of reference. Similarly, social desirability may colour the answering of questions (cf. Foddy, 1998). Thus, in a sense, interactivity may thwart its observability (Kanis, 2001, 2003).

Keep this inherent limitation in mind when generating insights by empirical research into user-product interaction, see chapter 2.

2 Doing usage observation: some practical guidelines

Figure 3 gives an overview of the consecutive activities in doing usage observation. The topics mentioned are briefly discussed in this chapter.

research objective (2.1)

presumptions (2.2)

research questions (2.3)

operationalisation: research design (2.4)

- models (2.4.1)
- research environment (2.4.2)
- participants (sampling, selection) (2.4.3)
- representativeness (2.4.4)
- instruction, unobtrusiveness (2.4.5)
- ways of observing (2.4.6)
- self-reports (thinking aloud, retrospective interviewing) (2.4.7)
- asking questions (2.4.8)
- identification of experience (2.4.9)
- measuring human characteristics (2.4.10)
- carry-over (2.4.11)

doing observations (2.5)

- who is incharge (2.5.1); intervening (2.5.2); end of session (2.5.3)
- pilot (2.5.4)
- number of participants (2.5.5)

analysis (2.6)

- approach (2.6.1)
- quantitative-qualitative (2.6.2)

redesign consequences (2.7)

communicating results (2.8)

Figure 3 Consecutive research activities in usage observation

2.1 Objective of usage observation

The aim is to gain insight into user activities in terms of what, how and why, in order to apply these insights in usage oriented product design.
Observation may involve existing product(s)/situations; simulation with the help of concepts, drawings; models, partly functioning; prototypes.

2.2 Presumptions

There is no research without presumptions, implicit as these may be. Presumptions provide the reason for what will be investigated, such as

- why designed *usecues (see* capture Figure 2; paper concerned: Kanis, Rooden & Green, 2000, see Reader PUUE) are supposed to work, or may not work at all, by hindsight,
- why the possible role of experience is considered and which experience, and
- why the operation of a product is expected to be strenuous, or particularly smooth, possibly depending on body measures or physical capacities or sensory limitations.

Try to be sparingly with the number of presumptions raised, give way to plausibility rather than mere hunches. Do not let your explorative research diminish to just checking the presumptions: keep an open eye, keep an open ear, instead of secluding yourself from the unexpected. Avoid talking hypothesis testing: "In testing an hypothesis, you know what you Ye going to discover/', as Kirk and Miller (1986) put it. Often this type of research is "aimed at preventing discovery." (same authors, p. 15).

2.3 Research questions

A research question is a sentence with a question mark. Research questions should be as concrete as possible without sliding into such a detail that (too) many of questions are needed to cover the whole study.

Possible research questions in the observation of product usage (partly obvious in view of various considerations discussed in chapter 1):

- What are participants in a study actually doing in operating a product, a model (etc.), what usability problems are they facing (in terms of perception, cognition, action, experienced effort)?
- Which role is played by featural (form) and functional product characteristics (or presumed/ designed *usecues,* where appropriate) in the emergence of usability problems, or in a user-product interaction that appears to progress smoothly?
- Are participants possibly misled by false cues?
- What is the role of experience? (This topic may also be addressed in one of the previous questions.)
- What is the role of human characteristics such as anthropometries in the emergence of usability problems?
- To what extent are observed usability problems ephemeral? And are there usability problems which may be missed in a first confrontation as these problems tend to emerge after prolonged usage?

Be keen on a proper delineation: something is wrong with a presumption that is not addressed, in one way or another, in a research question. Similarly, something is wrong with a research question raising the issue of possible differences in whatever usability aspect between elderly people and the rest, or between men and women (etc.) without a corresponding presumption (see paragraph 2.4.3 for the disputable relevance of such demographic characteristics as age, gender and education in design oriented observational research).

Try to limit the number of research questions to five, at the most, in order to avoid fragmented reporting.

2.4 Operationalisation

2.4.1 Mod els

When dealing with design models, an issue is which characteristics of the intended design are to be represented in the model and which are not. A difficulty is that it is not known beforehand which characteristics of the design will play a role in the interaction, or, rather, may be adhered to as usecues by different users. Try to represent explicitly designed cues into a design model (e.g. product graphics). It does not seem necessary to make very refined design models such as precisely simulating the looks of the intended design.

Non-functioning design models are a drawback. They are inescapably approached as such by users, rather than as functioning products. As a result there is a shift in users' intentions with more emphasis on being in a research situation: participants may set out to demonstrate their understanding of the design, to be critical, to be solving a puzzle, or to show their competence. This can be prevented to some extent by building a situation as naturally as possible with realistic tasks for the participants.

See Rooden (1999, 2001) for further information about observational inquiry on the basis of models.

2.4.2 Research environment

Preferably, the research environment should constitute a natural context, e.g. in people's home, providing the possibility to create an informal atmosphere. See paragraph 1.3.1 about generality: the more natural/complete the research environment, the less reason to presume artificiality and the more confidence in generalisation. If the interaction mainly takes place In the head', e.g. the user sitting behind a screen where it is 'all happening', some artificiality of circumstances may be less disturbing.

Artifacts which are often unavoidable: the researcher, the camera, any instruction, ... Possible remedies: suggesting to participants a different focus, giving participants a distracting instruction, ... Afterwards, participants should be fully informed and be given the opportunity to require deletion of any recording (see 2.4.5 for some other suggestions to create unobtrusiveness). In order to enhance comparability between participants, make sure that the circumstances are identical in as far as possible between participants: present products, accessories in same way, on the same spot, in same sequence (unless there are good reasons not to do so, for example when the naturalness of a context prevails). However, the ongoing research may provide good reasons to waive comparability, in the case of a very informative observation which urges to address the phenomenon at issue explicitly with the participants to come. Then it is important to document carefully what is changed in the research design.

2.4.3 Participants (subjects)

Try to achieve heterogeneity as to the issues raised in the presumptions and research questions. If the role of experience in using particular software is at stake, it may make sense to look for both elderly and young participants as an efficient help in the selection in order to end up with participants differing widely in experience. Note that age is a so-called *proxy,* in this case for knowledge and/or experience. Similarly, gender may be used as a proxy in order to achieve variation in force exertion. Age and gender are never "factors' in the sense of a causal, explanatory variables; the same holds for other demographics such as level of education, income etc.. Here the term "explanatory[7] means that the (non-)emergence of let's say usability problems equates with differences in the type of experience or in body measure (for which the indicated demographics are proxies).

Practical difficulty. Characteristics which are presumed to be relevant in a particular study, e.g anthropometries, type of experience, visual acuity, should be established or measured after the observation of usage in order not to emphasize the research context (see 2.4.2). This means that any intended heterogeneity (e.g. participants with big hands as well as participants with small hands, idem participants familiar with surfing the Internet and others having no clue) should only be casually addressed in the selection of participants. Keep in mind that it is no option to select participants with the "same experience': knowledge, experience tend to be very diverse, multi-faceted etc. amongst people.

2.4.4 Representativeness

This is a widely misused concept. Especially in observational research, the usage of the term "representative' is suspicious.
Representativeness should always be specified as to which (user) characteristic(s). The claim ""my sample is representative" is meaningless.
And what about the claim ""My sample is representative demographically for the future user population."? This claim may only make sense if a demographic variable at issue (e.g. age, gender, profession, income) is associated with a user characteristic which turns out to be explanatory in the explorative study, such as a body measure or some kind of experience, see previous paragraph.
Note that the significance of user characteristics in terms of (not) being explanatory often involves one of the questions to be answered by the explorative study. As long as this is not known, any claim like ""My sample is representative demographically for the future user population." can be dismissed as premature and dispensable.

2.4.5Instruction, unobtrusiveness

The instruction should be the same to participants and should, obviously, help to produce the data that suffice to answer the research questions.
In order to avoid emphasising the research environment: keep the instruction as short, as global as possible, in everyday language (straightforward, no complications, no going astray). Do not specify every single activity you would like to observe, such as ""Could you open it up, could you install this, would you do that, ...?." WRONG, say only what the end result should be. See whether the need to give an instruction can be avoided by presenting the design that has to be operated in the wrong default (too low/high, empty, deviant setting etc.).
Question the need to give participants a manual: wouldn't it be much more interesting to observe
participants struggling with the (absence of) "tellings' by the product? And if a manual is deemed necessary, consider whether the one accompanying the product is appropriate indeed: the information presented in that manual may be too much, hidden, even wrong. If so, make a succinct one yourself.

2.4.6 Ways of observing

In most cases the best way to record observations is by video. If video is not available, taking pictures may be an alternative, otherwise paper and pencil. To some extent, people tend to become accustomed to the presence of research devices.
DO NOT use lists of (possible) premeditated "observations' which only have to checked: if you think that you know that much, stop doing exploratory research as this, apparently, is not your "piece of cake'.
It is difficult to combine the operation of a camera with being in charge, i.e., communicating with the participant. It may be helpful that someone (if feasible the researcher her-/himself) makes notes of particular events which should be discussed afterwards with a participant, see next paragraph.

2.4.7Self-reports: thinking aloud, retrospective interviewing

Information about users' perceptions and cognition (see Figure 2) is indispensable in linking use actions or usability problems to characteristics of the design, see par. 1.3.
Users' perceptions and cognition can not be observed directly. Only by extra utterances can insights be gained into these activities: people may make remarks on what they perceive and think. Concurrent thinking aloud with a passive observer generally produces little insight into why

users do what they do, and retrospective accounts easily suffer from retrospective construction and rationalisation.
A drawback of interviewing is that a research setting is emphasised when the researcher asks questions with answers supplied by the user. It is preferable that during the observation a different setting is created, with the user being regarded as the expert speaker and the researcher in the role of an active listener (Boren and Ramey, 2000). With subtle probes (e.g. *"Mm hmmmni* and *"Uhutf)* the listener keeps the conversation going. The information revealed in this way suffers less from biases which contaminate retrospectively collected data (i.e. memory processes, rationalisation, social desirability). Occasionally, a question may be asked in order to clarify a problematic situation. However, beware of the drawbacks of interfering abundantly such as by concurrent interviewing, as this may obscure views on natural usage.

2.4.8 Asking questions

* Do not work with completely elaborated questions such as a questionnaire in a survey. Just prepare points to raise, e.g. the origin of observed usability problems, which may have to do with the perception and understanding of usecues, and also with the effort experienced (physically, mentally). You may also wish to find out why particular operations were observed to proceed smoothly.

+ A list consisting of points of interest to discuss with the participants tends to grow during a study due to observations prompting new ideas. Thus, to some extent the comparibility between participants may be sacrificed, see 2.4.2. The comparability is least affected if in the retrospective interview with a next participant new points of interest are raised at the end (if possible).

* If you want to ask something which is not covered by the research questions, then there is something wrong with these questions (adjustment needed).

+ Start with open questions, avoid leading questions, do not use adjectives and adverbs.

* Never interrupt a participant, and immediately stop talking when the participant interrupts you.

* Be aware that people who are asked the same question may answer different questions, mainly due to various possible interpretations of words, see par. 1.3.3. In addition, participants may have different frames of reference, see Kanis (2003; added to this document) about the interpretation of scales. Finally, people's reactions may be inspired by social desirability and the eagerness to make a competent impression, or can be the result of rationalisations (i.e., by hindsight).

* Make sure that a participant talks about her-/himself, about her/his individual perceptions, cognitions, experienced effort etc., rather than falling back into alleged experiences of other people.

2.4.9 Identification of experience

Asked about experience, people may come up with something which is completely different from the product under investigation. Find out what makes things similar in the eyes of the participants.
Do not stop asking whether people do have experience; find out with what product, which product part, which program exactly etc.
If possible, have participants' experience demonstrated using their own products.
Generally, the kind of experience is much more important than an estimated frequency of usage. If you do ask about the latter, than also ask how long ago it was that something was last used.
As mentioned above (2.4.3), always do the detailed charting of experience after the observations.

2.4.10 Measurement of human characteristics

In the type of research dealt with here, body measures (e.g. stature, hand width) do not have to be very precise (for instance stature may be asked, as a reasonable estimate). The same applies for exertable forces such as measured by a hand dynamometer.

Note that right/left handedness may differ for writing, throwing a ball, firing a gas stove etc. As noted above (2.4.3), always do measurements after the observations.

2.4.11 Carry-over

People learn, get acquainted, bored, tired. So, in the case of more than one task, there may be so-called carry-over: the result of a task may reflect (some of) the experience in a previous task. By reversing tasks (if possible or appropriate) different types of carry-over are created. In general, it is illusory that these different types of carry-over would cancel out each other. Often,
textbooks propagate random assignment of tasks to participants in order to neutralise carry-over. This is a misleading advice.
There is no carry-over in a so-called *between-subjects research design,* that is: one task for each participant. There are two considerable advantages of the so-called *within-subjects research design* (all tasks to all subjects): much less participants needed, and the possibility to compare the carrying out of tasks intra-individually.

2.5 Doing observations

2.5.1 Who is in charge?

For designers testing their own "brainchild', it often appears to be hard not to interfere if participants do something "stupid'. Avoid presenting yourself as the designer as this may induce "socially desirable' reactions (see 2.4.8).

2.5.2 Intervening

If a participant gets stuck, do not provide help too quickly. For instance, ask first how things are going, followed by questioning what the difficulty exactly is. Do not be too generous in giving clues, ask e.g. ""Have you seen this?" or ""Have you thought about that?".

2.5.3 End of session

If you have announced the end of a session, do not switch off the video (or sound recording) immediately: on leaving, often participants tend to make some reflective remarks containing new information.

2.5.4 Pilot

This is a compulsory try-out in order to see whether the research-design is OK: objects (if any) in the right position, presented/offered in the right sequence? camera's in good position? recorded sounds clear? smooth retrospective interview?
Use a checklist of all the things that should be done, should be in place, should be ready before an observation starts.
Only if everything is going allright, without any changes in the research-design which could possibly affect the observations, the results of the pilot may be used in the analysis as if generated by a regular participant.

2.5.5 Number of participants

Figure 4 illustrates the significance of the observations gained on the basis of a small number of participants: relatively rare cases tend to be missed whilst recorded observations probably are not rare.
In the Appendix, an algorithm is given to estimate the total number of usability problems given the observations after any number of particiants.

Figure 4 The low probability of being rare, or
the high probability of being there

2.6 Analysis

2.6.1 Approach

Step 1: data per participant

Observational studies may yield hours of video-recordings. The first step is to transfer the relevant information to paper. In transferring and summarizing the data, selections are made. The research questions delineate which data are relevant. In general, the relevant data consist of

a list of use actions and manifestations of usability problems at a detailed level. The link with the visual data should be maintained by annotating where on the tape/disc the specific clip can be found. In cases it will be necessary to later review certain clips to understand what actually happened.

Go through your video-data in a systematic way, and be consistent in your registrations. Avoid interpretation, or at least try to keep it separate from the "actual' data. You can highlight specific or remarkable observations and add your comments, but an overview of all data is needed for final conclusions.

When users' actions are difficult to describe in words, such as with physical manipulations, pictures (video stills) can be added to the summaries. Pictures are also good to refresh your memory during further analysis.

Step 2: data per research question

The data are combined across participants. Information related to each research questions is gathered. Do not consider the original phrasing of the research questions as a straightjacket: change the wording if that covers the research more satisfactorily, by hindsight. With each research question a balance should be sought between general descriptions and specific examples from the data with pictures and quotes. Do not sacrifice 'the exploitation of subjectivity in terms of individual accounts' (see 1.3.2) to the presentation of summative measures, try to stick to so-called low inference descriptors, such as literal quotation of participants' utterances.

In combining data from various participants new insights may be gained, prompting additional analyses of parts of the video recordings.
In answering a research question on usability problems, a table showing which participants experienced which usability problems is a good starting point in presenting findings.

2.6.2 Quantitative-qualitative

See the paper on this issue by Kanis, Weegels and Steenbekkers (1999; added to this document), see also Kanis, 2001, 2003 (both added to this document).
Lord Kelvin (mathematician, physicist) is often quoted by social scientists as follows: ""When you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind (1883, in Sydenham, 1979)."
However, this is what Kelvin wrote: ""In physical science [underlining added, ...] when you can measure what you are speaking about, and express it in numbers, you know something about it; but [...] when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind [...]."
In 1883, for instance psychology as a science was at its infancy.

2.7 Redesign consequences

Think of a "solution space' between two extremes:

- such a (re)design that users can do what they are used to or what they like, wherever they are, think of double insulation of electrical devices, and
- a (re)design which enforces particular use actions, excluding others, think of hair driers fixed to a wall in order to prevent driers falling into the bathtub and causing electrocutions (Mauro, 1978).

2.8 Communicating results

All kinds of reporting are conceivable: informing members of a design team by a video compilation or a strip, a report to a contractor, a paper for a scientific journal ... Some recommendations for the present Master course (if feasible, such as in view of the available space).

* Visualisations tend to be indispensable, such as a product rendering, a storyboard of a possible use sequence, illustrations of results; use tables, diagrams rather than text.
* Always describe the research design, be clear about instructions (preferably literal text, in an appendix), mention consequences of the pilot.
* Deal with the results of the analysis in a logical sequence of the research questions.
* If appropriate, briefly reflect on the presuppositions: did they come true, or can nothing be said?
* Do not report per participant (such information may be annexed), answer each research question involving all participants. Never present primary data without any context, not even in an annex.
* Present pictures with captures, referred to in the text.
* Mention remarkable events (e.g., a retrospective denial of an action which is on the video or something unobservable which a participant claims to have seen).
* Quote participants literally, without upgrading their sayings to what is deemed cultured language.

Appendix

Estimating the total number of usability problems

In Kanis and Arisz (2000, added to this document), the following algorithm is derived for estimating the total number of usability problems

$F_{CG} = F_1.F_{n\_1}l(F_1 + F_{n\_1}\text{-}Fn),$ with -------------------- .... --------.................. -------------- (1)

F_{ro} the "total' number of distinctive usability problems found after an unlimited number

Coo^1) of participants, F_1 the average of the *n* series of

observations *(n* participants), and

F_{n_1} the average of the ^combinations of the findings over *n-1* participants.

Note that the denominator in expression (1) gives, on average, the overlap between the number of the different usability problems found after *n-1* participants and the number of usability problems found with an extra participant (the rP^1*).*

Example (four participants, eleven usability problems (a ... k)).

Participants				*observed usability problem*							
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	7	*k*
Pi	x	x	x	x							
P2		x		x	x	x	x				
P3				x			x	x	x		
p4		x			x		x			x	x

F_1 = (4+5+4+5)/4=4.5

F_n^ =

(9+9+10+9)/4=9.25

F_n = H

F_{ro} = 4.5x9.25/(4.5+9.25-

II)=15.1 F_n/F_{aD}xIOO% =

In Kanis and Arisz (2000; added to this document) two biasing mechanisms in estimating F^ are discussed: the involution in the binomial model of an average probability to come across usability problems, and the fact that usability problems with the highest probability to occcur will, on average, be discovered first. It is concluded that expression (1) yields an informed guess as to F^ , rather than firm predictions.

Of note is that concurrent monitoring of the required number of participants in a study begs for concurrent categorisation of the observations after each participants. Often this is not feasible, given the way in which fieldwork has to be scheduled. In addition, categorisation - in terms of which observations are conceived as similar and what makes observations distinct - is an emerging phenomenon, based on continuous and reflexive adjustment in the ongoing involvement of consecutive participants. Hence, in practice the computation of F_w may be a conclusion, to be drawn after an in-depth analysis of all observations.

References

Boren, T.M. and Ramey, 1 (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication,* 43, 261-277. Foddy, W.H. (1998). An Empirical Evaluation of In-Depth Probes Used to Pretest Survey Questions. *Sociological Methods & Research,* 27, 103-133.

Garfinkel, H. (1991). Respecification [...], in G. Button (ed.), *Ethnomethodology and the Human Sciences.* Cambridge University Press, 10-19. Kanis, H. (1998). Usage centred research for
everyday product design. *Applied Ergonomics,* 29,
75-82. Kanis, H. (2001). Scientific credentials for qualitative and quantitative research in a
design
context, in Contemporary Ergonomics, 389-394. Kanis, H. (2003). Research in Design:
Situated Individuality versus Summative Analysis.
Proceedings IEA Kanis, H., Rooden, MJ. & Green, W.S. (2000). Usecues in the Delft
Design Course. In
Contemporary Ergonomics. Taylor & Francis, 365-369. Kanis, H., Weegels, M.F. & Steenbekkers.
L.P.A. (1999). The uninformativeness of quantitaive
research for usability focused design of consumer products. Proceedings HFES Annual Meeting,
401-405. Kanis, H & Arisz, HJ. (2000). How many participants: a simple means for concurrent
monitoring.
In IEA/HFES proceedings, 6.637-6.640.
Kirk, 1 & Miller, M. (1986). *Reliability and Validity in Qualitative Research.* Sage. Mauro,
C.(1978). Can hairdriers be safer? Research says 'yes'. *Industrial Design,* 25, 38-43. Nardi,
B.A. Studying Context: A Comparison of Activity Theory, Situated Action Models and
Distributed Cognition. In Nardi, B.A. (ed.), *Context and Consciousness,* MIT. Neisser, U.
(1976). *Cognition and reality.* W.H. Freeman. Rasmussen, 1, Pejtersen, A.M. and Goodstein,
L.P. (1994). *Cognitive Systems Engineering.*
Wiley. Rooden, MJ. (1999). Prototypes on trial. In W.S. Green and P.W. Jordan (eds.), *Product Design*
and usability Current Practice and Future Trends. Taylor & Francis, 138-150. Rooden, MJ.
(2001). Design models for anticipating future usage (thesis). Faculty of Industrial
Design Engineering, Delft University of Technology. Standaert, A.A. (2004), Cognitive
Fixation in Product Usage. Delft University of Technology
(thesis). Sydenham. P.H. (1979). *Measuring instruments: tools of knowledge and control.* Peter
Peregrinus
Ltd (Stevenage, UK).
Suchman, L.A. (1987). *Plans and situated actions.* Cambridge University Press, Cambridge. Wilson,
J.R. & Rutherford, A. (1989). Mental Models: Theory and Application in Human Factors.
Human Factors, 31, 617-634.

Research Context the set-up of user trials

Annelise de Jong

This theme contains a variety of aspects related to choices to be made on the set-up of user trials. Ideally, a research context is pursued in which the conditions are similar to those of a real-life context. For instance, research on usage of a house-hold product would ideally take place in people's homes, possibly with several family members present. However, even in such situations, the researcher may influence the process simply by being present at the time of use, or, when the research takes place after usage in the form of retrospective interviewing, self-reports may be biased.

further reading Suchman, LA. (1987). *Plans and situated actions, the problems of human-machine commun/cat/on,* Cambridge University Press

introduction

In setting up user trials there are a number of issues related to the research context. Three issues will be highlighted in this part of the reader:

1 the composition of participants: single participants are often asked to verbalize their thoughts. However, this may be difficult to maintain during the research. Joining participants in groups of two or more may influence the type and amount of information given by participants.

2 the tasks for participants of the research: in previous researches most user trials were done to identify problems of particular aspects in a design. Therefore, participants were asked to perform well-defined tasks or questions. More recently, scenario-driven instructions are used, in which participants can come up with self-determined tasks during the exploring of the product.

3 time and duration of the research: when presenting participants with new products in user trials they are confronted with new ways of using the product which may lead to problems. The next time the product is used, the new ways of usage may be more familiar to the participants, thus giving less or maybe other problems. Therefore, repeated user trials in time may provide information on permanent or temporary problems.

1 composition of participants

The composition of participants and observers can be varied in user trials, depending on the nature of the product and the research questions and set-up. For instance, single users can be accompanied by a observer who acknowledges their utterances and actions by statements such as "uh-huh' to stimulate the Thinking-out-Loud paradigm (TOL paradigm, another abbreviation for this technique is TA: Thinking Aloud). However, doing research with groups of participants without an observer present may be a more effective way of gathering information on the product. Also, participants are thought to fall silent when faced with problems. Hackman and Biers (1992) conducted a comparative study with three different compositions of participants (Single, single with Observer, Team of two persons) to determine the quantity and quality of verbalizations in a TOL paradigm.

"7776 *use of the TOL paradigm is not without its problems however...Users in the standardsingle-user-TOL paradigm do not spontaneously verbalize-out-loud. Some users do not verbalize at aII despite instructions to do so. Others constantly have to be reminded and prodded to verbalize throughout the study."* Citation from Hackman and Biers, 1992, p. 1205

They conclude saying that the three compositions did not differ in quantity of information but there were differences in the quality of information (on a scale from low-moderate-high value), namely the number of remarks expressing uncertainty on the steps to be taken in groups of two users (Team) were higher than in the other two compositions.

^{11} *More importantly, on the more difficult tasks, where users experienced problems in utilizing the software, the Team spent more time making moderate/high value verbalizations than did either the Single ...or Observer... conditions... Thus, with a team of users, TOL resulted in a greater amount of time spent making comments which were of value to the designer in a situation where the user was having problems in using the software."* Citation from Hackman and Biers, 1992, p. 1207

reference Hackman, G.S. and D.W. Biers (1992). Team usability testing: are two heads better than one? In: *Proceedings of the Human Factors Society 3€f^{h} Annual Meeting,* vol 2, p. 1205-1209.

further reading Kemp, J.A.M. and T. van Gelderen (1993). *The co-discovery method: an informal method for iteratively designing consumer products,* IPO Annual Progress Report 28, Institute for Perception Research, Eindhoven, p. 143-150.

2 tasks for participants

In defining user tasks attention should be given to the way in which people are used to discover the meaning and functions of products. A method to perform such research is scenario-driven research in which participants are presented a small story in which they play a role and then are asked to achieve a goal without specifying the task as such.

Taking this one step further, the goal of the research is stimulating participants to explore a product first to find out what it does and its functions and if needed to ask them to perform several tasks later in the research. Information on a comparative study on these types of tasks can be found in Vermeeren (1999).

Also, the researchers' role may influence the course of the research. To maintain a situation in which user trials take place outside the lab, a passive role of the researcher might seem appropriate. However, to gain more information from the research a more active role might be preferred including asking questions during the research and doing interviews afterwards.

reference Vermeeren, A.P.O.S. (1999). Designing Scenarios and Tasks for User Trials of Home Electronic Devices, In: *Human Factors in Product Design,* W.S. Green and P.W. Jordan (eds.), Taylor and Francis, p. 47-55.

further reading for more information on the researchers' role and verbal information see the part on Thinking aloud in this reader.

3 time and duration of the research

It may of interest for designers to understand the nature of problems that were found in user trials. Are they of a permanent nature or do they occur in initial use only? Regular use in contrast to first use may provide information on other (types of) problems such as the permanence of problems. A study by Loopik et al. (1994) reported on the permanence of problems of vacuum cleaners in on-site user trials with repeated visits. The authors categorized problems in fleeting operational difficulties and (quasi-) persistent difficulties and difficulties in actual use and discussed these problems for each part of the vacuum cleaner, thus differentiating between initial problems, which users solved by trial and error, and permanent problems, which urged subjects to consult external sources of information.

reference Loopik, W.E.C., H. Kanis and A.H. Marinissen (1994). The operation of new vacuum cleaners, a users' trial, In: *Contemporary Ergonomics, Ergonomics for All,* p 34-39.

further reading Rasmussen, J. (1983). Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models, *IEEE Transactions on systems, man and cybernetics,* Vol SMC-13 (3), p. 257-266.

TEAM USABILITY TESTING: ARE TWO HEADS BETTER THAN ONE?

and

George S. Hackman
IBM Corporation,
Bethesda, MD

David W. Biers
University of Dayton

The purpose of the study was to compare a team usability testing paradigm with that of the typical single user paradigm in terms of the quantity and quality of the user's verbalization (i.e. . thinking out-loud) and performance. The study employed a three group design in which the type of usability paradigm (Single, Observer, Team) was manipulated. Users first learned to use an off-the-shelf database management package by means of a short tutorial and then engaged in six structured tasks. While engaging in the tasks, the users either thought-out-loud alone (Single condition), in the presence of an observer (Observer condition), or as participants of a team working on the tasks together (Team condition). Results indicated that there were no significant differences among the three conditions in terms of performance nor any extensive differences in their subjective evaluation of the software. However, users in the Team condition spent more total time verbalizing than those in the Single or Observer condition. More importantly, results of a verbal protocol analysis revealed that the Team spent more time making statements which had high value for designers than did the other two conditions (which did not differ from one another). When broken out by .individual users in the Team, there were no significant differences between individual team members and users in the other two conditions in making high value comments. The results suggest that the Team paradigm may be more efficient in extracting high value information without any noticeable differences in performance or subjective impression of the software.

INTRODUCTION

Usability testing is the means for assessing the ease of use of computer systems. The traditional usability test is conducted in a laboratory environment with a single user working in isolation on a contrived structured task scenario in front of a one-way mirror and cameras. While usability testing has become a standard part of the system development process in many corporations, there is little systematic research on the process of usability testing itself. The present study is part of a research program that systematically investigates usability testing methodology and the factors which affect the outcome of a usability test.

This study examined the utility of a "team" usability testing paradigm patterned after elements of Bell Northern Research Corporation's Co-discovery technique (Kennedy, 1989). The purpose of the study was to compare this team paradigm with that of the typical single user paradigm in terms of the quantity and quality of the user's verbalization (i.e. thinking out-loud) and performance.

One of the most frequently used paradigm in software usability testing involves Thinking-Out-Loud (TOL) (e.g., Jorgensen, 1989). TOL is a technique whereby a single user is asked to think-out-loud as he/she interacts with the software. This verbal protocol information is then used to gain a better understanding of the user's insight into the software and provide information on the source of problem difficulty. This verbal protocol data provides valuable information to the software designer that is not garnered from performance data.

The use of the TOL paradigm is not without its problems however (e.g., Hoc & Leplatt, 1983). Users in the standard single-user-TOL paradigm do not spontaneously verbalize out-loud.- Some users do not verbalize at all despite instructions to do so. Others constantly have to be reminded and prodded to verbalize throughout the study.

One possible reason for the low quantity and quality of verbalization is the unnaturalness of the situation. The user is brought into a room with bright lights and video cameras and told to talk out-loud to people watching behind a one-way mirror in the next room. Usability practitioners have attempted remedy this situation by placing a second "neutral observer" in the room with the user at the time of the usability test. The passive observer then acknowledge the user's verbalizations with an "uh-huh" or "OK" but does not comment on the correctness or incorrectness of any action. The thinking behind this paradigm is that the user would feel more comfortable thinking out-loud with another person present and would therefore verbalize more frequently. Whether or not this improves the quality of the verbalization is open to question.

One new approach to encouraging verbalization is to have two users work on the computer at the same time, as a team. The users are encouraged to communicate with one another as they work on the task and these communications are used as information for possible system redesign. There is no scientific evidence as to whether this team approach at all increases the quantity and quality of verbalizations in a TOL paradigm. Moreover, there is no information on performance of teams. For example, is a team of users more likely, less likely, or equally likely to encounter problems using software than a single user?

METHOD

Design and Subjects

The study employed a three group design in which the type of usability paradigm (Single, Observer, Team) was manipulated. Forty subjects participated as users with 10 users per group. In the case of the Team condition, this represented 10 pairs of users. The users were selected using a sampling plan in which the age and education of the user was varied and balanced across conditions. For purpose of assignment of subjects, age was divided into three categories (18-26, 27-45, and 46-55), as was as education level (high school only, 2 years of college or less, more than two years of college). For the team condition, the two users were homogeneous with regard to the age and education category.

Materials and Facilities

The software platform used in the study was an off-the-shelf database management package called Reflex (Version 1.0). The study was conducted in the University of Dayton Information System Laboratory, a facility specifically designed to test the usability of consumer software. All sessions were video taped for later verbal protocol analysis.

Procedure

Users learned to use the software by means of a short tutorial. Then they engaged in six structured database management tasks (i.e., creating a form, data entry, changing the view, creating a graph, filtering the database, and crosstabulation). Users were given a maximum of 15 minutes to complete each task.

While engaging in the tasks, the users were instructed to think-out-loud. In the Single condition, the user verbalized out-loud to himself/herself with no one else present in the room. In the Observer condition, the user thought-out-loud with an observer present in the room. Users in this condition were told that the observer was present to take notes and that they were to verbalize as if they were talking to the observer. In the Team condition, the users were encouraged to think-out-loud to one another as they worked together as a team to accomplish the six tasks. They were instructed to verbalize out loud within the context of communicating with one another. One member of the team volunteered to sit at the keyboard and served as the keyboard operator (T-KO) while the other member of the team served as the advisor (T-ADV).

Prior to the start of the study, users in all conditions were shown a short video tape on thinking-out-loud to increase the overall level of verbalization. In addition, this tape highlighted the types of verbalizations which are more valuable to the software developers (e.g., comments on design of the interface, comments on the evaluation of the interface, statements which give insights into the user's thinking behind actions)* Users were encouraged to give these higher level verbalizations.

Dependent Measures

The primary dependent measures were the frequency of verbalizations, the total time spent verbalizing, and the rated quality of information generated, Subsequent to data collection, the video tapes were subject to a verbal protocol analysis. A relatively low-level verbal protocol analysis was conducted in which the basic unit of measurement was a single sentence. Each utterance (sentence) was classified along two dimensions—the type • of verbal statement (declarative, declarative plus explanation, evaluative, evaluative plus explanation, uncertainty, reading, data entry) and the referent (personal, task/instruction difficulty, steps/procedures/actions, interface). In addition, each utterance was rated according to its value to the designer on a 4 point scale (0 = no value, 1 = low value; 2 = moderate value, and 3= high value). Value was defined as the degree to which the statement provided insight into: the design or redesign of the system, the person's understanding of system terminology, and the user's mental model of the software. Only the data with regard to value are reported here.

The principal author coded and evaluated all utterances. A second trained evaluator coded and evaluated 20% of the utterances (two randomly chosen users from each condition) as a reliability check. There was 94% agreement among the two evaluators.

The primary measure of task performance was task completion time. Two additional measures, user rating of task difficulty and evaluator judgement task success, were taken at the conclusion of each task. For this latter measure, user performance was classified as either a success, success with major problems, or a failure.

Subjective ratings of software usability were taken at the conclusion of the study using a modified version of the Chin, Diehl, & Norman (1988) questionnaire. The team members completed the questionnaire separately. The Chin, et al. questionnaire contains 9-point bipolar scales which can be grouped into five major categories; learning, screen, terminology and system information, system capabilities, and overall evaluation. For purposes of data analysis, the mean rating for the scales comprising each of the five categories was used as the dependent measure.

RESULTS

Based upon user ratings of task difficulty, the six task were collapsed into two groups of three tasks each—easy tasks (creating a form, data entry, changing the view) and hard tasks (creating a graph, filtering the database, and crosstabulation).

User Performance

User performance did not differ as a function of TOL paradigm. There were no significant differences among the three conditions in the evaluator judgement of task success (F(4,54) = 0.63, £ = .918) or in task completion time (F(2,27) = 0.08, £ = ,926). There was no significant difference in user ratings of task difficulty as a function of TOL paradigm (F(2,27) = 0.93, £ = .406). There were no significant interactions of paradigm with task difficulty for any of these measures.

Hard tasks (M = 706.96 s) took longer to complete than easy tasks (M = 405.31 s) (F(2,27) =58.76, £ < .001). Fifty-five percent of the hard tasks were classified as task failures whereas only fourteen percent of the easy task were so classified.

Overall Level of Verbalization

Since the results of the statistical

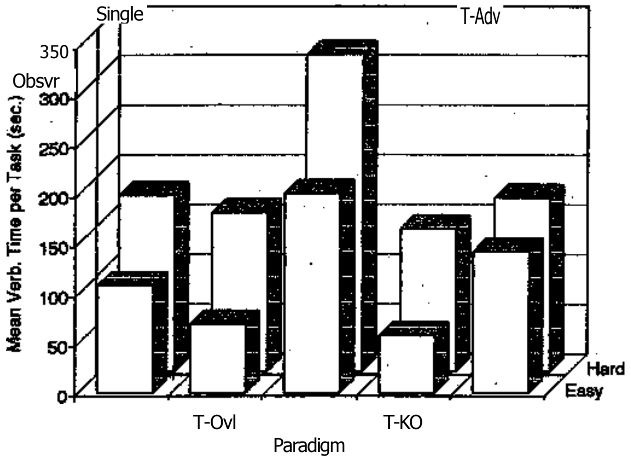


Figure 1. Mean verbalization time per task as a function of TOL paradigm and task difficulty.

analysis on the frequency of verbalization and the total time spent verbalizing were comparable, the presentation of the results will focus upon the time based measure. Figure 1 presents the total time spent verbalizing for each of the three TOL paradigms on easy and hard tasks. Although not shown in the figure, the average percent of task time spent verbalizing was approximately the same for easy(M = 32%) and hard (M = 31%) tasks.

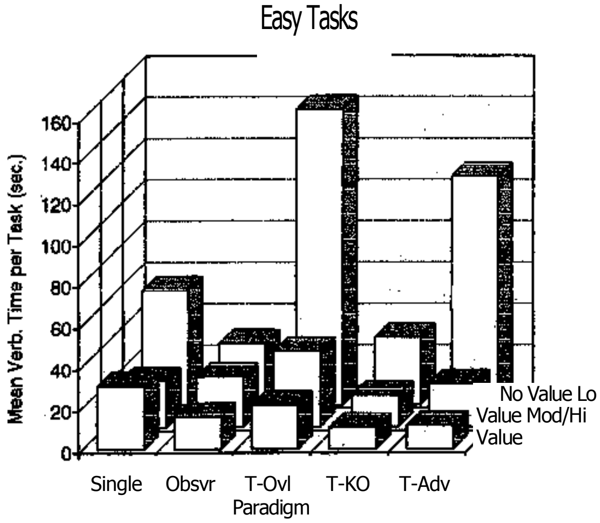
Users in the Team condition spent more total time verbalizing than those in both the Single (F(l,27) = 15.66, £ < .001) and Observer (F(l,27) = 24.09, £ < .001) condition. The latter two conditions did not did not differ from one another (F(l,27) = 0.90, £ = .350). This Team effect, however, simply can be attributed to the fact that there were two individuals who could verbalize rather than one. When broken out by individual users in the Team, there were no significant differences between individual team members and users in the other two conditions except that between the Observer condition and the Team-Advisor. Overall, the Team-Advisor spent more time verbalizing than did users in the Observer condition (F(l,36) = 5.15, £ = .029). Although the interaction with task difficulty was not significant, this effect was more pronounced for the easy tasks where the Team-Advisor spend time reading data values to the Keyboard Operator. The difference between the keyboard operator and the advisor was not significant overall (F(l,9) = 4.91, £ = .054).

Value of Verbalizations

Each verbalization was rated on a four point scale in terms of its value to the designer. Since the frequency of verbalizations given a value rating of 4 (high value) was low, the moderate and high value rating categories were combined for purpose of data analysis.

Figure 2 presents the time spent making no, low, and moderate/high value verbalizations as a function of TOL paradigm and task difficulty. Overall, only 21% of the time spent verbalizing was spent making statements rated as having moderate to high value to the designer.

More importantly, on the more difficult tasks where users experienced problems in utilizing the software, the Team spent more time making moderate/high value verbalizations than did either the Single (F(l,27) = 11.73, £ = .002) or the Observer (F(l,27) = 7.26, £ = .012) conditions. The Single and Observer conditions did not significantly from one another (F(l,27) = 0.53, £ = .472) in making moderate/high value verbalizations. Thus, with a team of users, thinking-out-loud resulted in a greater amount of time spent making comments which were of value to the designer in a situation where the user was having problems in using the software. However, when broken out by individual users in the Team,



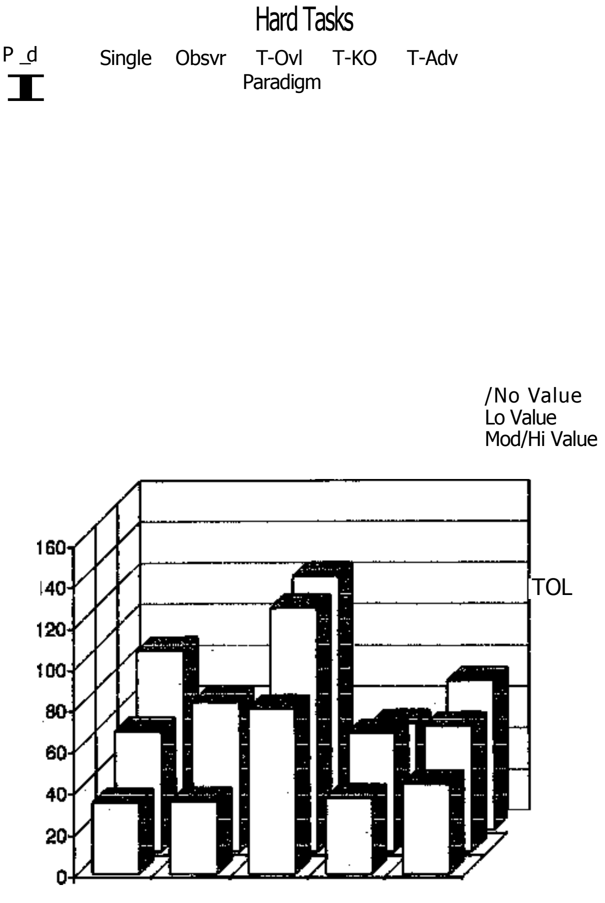


Figure 2. Mean verbalization time per task as a function of value, paradigm, and task difficulty.

there were no significant differences between individual team members and users in the other two conditions in the time spent making statements of moderate to high value.

Subjective Evaluation of Usability

Although there was some evidence to suggest that the user's evaluation of the software varied as a function of TOL paradigm, these differences were confined to just two factors. As shown in Table 1, the Team-KO (F(l,36) = 9.67, £ - .004) and the Team-Advisor (F(l,36) = 8.28, £ = .007) both found the software more difficult to learn than users in the Single condition. In addition, the Team-KO gave a more negative evaluation of the terminology and system information than did users in the Observer condition (F(l,36) = 4.81, £ = .035). However, there were ""no significant differences among the three TOL paradigms in terms of evaluation of screen characteristics (F(2,27) » 2.41, £ = .109), system capabilities (F(2,27) = 0.24, £ « ,792); and overall evaluation of the software (F(2,27) = 0.89, £ = .422).

Table 1

Mean subjective evaluation for five question categories as a function of TOL paradigm.

	Paradigm			
Category	Single	Obsvr	T-KO	T-Adv
Screen Terminol . &	5.37	5.90	4.20	5.17
System Information	5.73	6.57	4.80	5.37
Learning Syst. Capabilities	5.37	4.40	3.62	3.48
Overall Evaluation	5.69	5.60	4.45	4.79
	5.69	5.03	4.56	5.41

Note; Higher numbers indicate evaluations . more positive

DISCUSSION

This study is significant in that it employed a wide range of users, both in their age and in their educational level. The major result was that the team gave significantly more verbalizations of high value to designers and spent more time making high value comments. Although this can be reduced to the fact that the team spoke more overall and that there are two people talking rather that one, this finding is not trivial. If one is looking for a more efficient TOL methodology, one can extract more high value information in a shorter period of time with a team of users than with a single user.

There was a qualitative difference in the types of statements made by members of the Team in comparison to users in the Single and Observer conditions. The Team made many more statements expressing uncertainty about steps to be taken and uncertainty about the interface than did users in the other two conditions. As a matter of fact, much of the high value exchange between two team members represented an interchange where one member expressed uncertainty about an action or about the meaning of something displayed on the screen and the other team member responded with a declarative statement ("Why don't you try this") or with a declarative statement plus an explanation ("Try this because...").

A second noteworthy finding was that the team did not differ from the single user conditions in performance or differ appreciably in their subjective evaluation of the software. Although there were some differences in subjective evaluation, these differences were confined just to two factors. The fact that team members .gave lower ratings on ease of learning is understandable since the interchange between the two members brought their lack of understanding to the surface. However this uncertainity did not affect their overall evaluation of the software product.

The team was no more, or less, likely to experience problem difficulty or to evaluate the software differently. Perhaps this was due to controlling the age and educational level of the users across conditions. In addition, by using a team of users which were homogeneous with regard to age and education, no one team member was likely to dominate and thus possibly give the team an advantage.

It is also noteworthy that the Single and Observer condition did not differ in the number of verbalizations, total time spent verbalizing, and the value of their verbalizations. This finding contradicts popular belief and case studies reported in .the popular literature. However, the lack of a difference could be due to the verbalization training given at the outset of the study. Maybe users in the single condition were more willing to think-out-loud to themselves (an unnatural situation for most users) after seeing other users do it.

Finally, despite the verbalization training, most of the verbalizations were at a low level in terms of their value to designers. This, in part, may be the result of the low level of verbal protocol analysis where the basis unit of analysis was a sentence. No attempt was make to organize the utterances into higher order episodes and then rate the episodes for their value. Perhaps analysis at this higher level would yield a higher incidence of valuable utterances. However, taken at face value, the overall low level of verbalization in the present study indicates that the TOL methodology is a very inefficient technique for extracting information of high value to designers. Perhaps it would be better to utilize a more directive approach such as direct intervention or use retrospective TOL (Ohnemus, 1992).

ACKNOWLEDGEMENTS

This research was partially funded by a grant from BNR Corporation and by a State of Ohio Research Challenge Grant to the University of Dayton Center for Business and Economic Research.

Chin, J. P., Diehl, V.A., & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of an human-computer interface. Proceedings of ACM CHI 1988, ACM Press, 213-218.

Hoc, J. M., & Leplatt, J.(1983). Evaluation of. different modalities of verbalization in a sorting task.International Journal of Man-Machine Studies, 18, 283-306.

Jorgensen, A. H. (1989). Using the thinking-aloud method in system development. In G. Salvendy and M. L. Smith (Eds.), Designing and using human-computer interfaces and knowledge based systems. Amsterdam? Elsevier Science, 743-750.

Kennedy. S. (1989) Using* Video in the BNR Usability Lab. SIGCHI Bulletin, 21, 92-95.

Ohnemus, K. (1992) Retrospective vs. concurrent ' thinking-out-loud; The effect of delay. Unpublished master's thesis, University of Dayton, Dayton, OH.