An Empirical Study of the Effects of Three Think-Aloud Protocols on Identification of Usability Problems

Anders Bruun and Jan Stage(✉)

Department of Computer Science, Aalborg University,
9220 Aalborg East, Denmark
{bruun, jans}@cs.aau.dk

Abstract. Think-aloud is a de facto standard in user-based usability evaluation to verbalize what a user is experiencing. Despite its qualities, it has been argued that thinking aloud affects the task solving process. This paper reports from an empirical study of the effect of three think-aloud protocols on the identified usability problems. The three protocols were traditional, active listening and coaching. The study involved 43 test subjects distributed on the three think-aloud conditions and a silent control condition in a between-subject design. The results show that the three think-aloud protocols facilitated identification of the double number of usability problems compared to the silent condition, while the problems identified by the three think-aloud protocol were comparable. Our results do not support the common emphasis on the Coaching protocol, while we have seen that the Traditional protocol performs surprisingly well.

Keywords: Usability evaluation · Thinking aloud · Verbalization · Think-aloud protocols · Empirical study

1 Introduction

Software organizations and developers increasingly emphasize usability as a software quality characteristic [21], and usability evaluation is a key tool to improve usability [17, 21]. User-based usability evaluation is an approach to evaluation that produces two types of results [5, 21]: (1) usability measures and (2) usability problems.

Usability measures include various quantitative assessments measured directly or indirectly during a usability evaluation. The three factors in the classical ISO definition of usability are examples of this [13]: effectiveness, efficiency and satisfaction.

Usability problems are a list of problems that have been experienced by users during a usability evaluation that involves usage of the evaluated software. They are typically expressed in a format that enables developers to enhance the usability of a user interface design. Usability problems are also useful as means to understand a particular usability measure, e.g. why efficiency is low [21].

Think-aloud protocols are widely applied in user-based usability evaluations [5, 17, 21]. This technique was originally introduced in psychology to produce rich data through verbalization of cognitive processes [6]. While there is general agreement

J. Abascal et al. (Eds.): INTERACT 2015, Part II, LNCS 9297, pp. 159–176, 2015.
DOI: 10.1007/978-3-319-22668-2_14

about the strength of thinking aloud, there have also been concerns that it affects the task solving process of the test subjects. Yet the amount of systematic empirical studies of the effect of thinking aloud is limited. A recent study of three different think-aloud protocols characterized the literature on think-aloud protocols as being unclear with respect to the protocols applied. With a focus only on usability measures, i.e. effectiveness, efficiency and satisfaction, they found that different protocols affect the usability measures differently [19]. However, they did not deal with usability problems, and previous studies of this point in various directions.

This paper reports from an empirical study of the effect of different think-aloud protocols on the usability problems that are identified in a user-based usability evaluation. The study also replicates a previous study of the effect on usability measures [19], and compares the results on the two types of measures. In the following section, we present related work. This is followed by a description of the method applied in our empirical study. Next we provide the results and discuss these in relation to previous studies as well as their implications for usability practitioners and researchers. Finally, we provide the conclusion and point out avenues of future work.

2 Related Work

This section presents an overview of related literature on think-aloud protocols.

2.1 The Traditional Think-Aloud Protocol

The think-aloud protocol was originally described by Ericsson and Simon in 1984 [6] as a technique to study cognitive processes through verbalization. They divided verbalization processes in three levels. Level 1 and 2 are verbalizations that do not distract a subject from the current activity and rely only on information in the subject's short term memory. On these two levels, a participant can be probed or reminded to continue verbalizing, e.g. by saying "Keep talking" or "Um-humm". The assumption is that this probe limits distraction from the current task. Level 3 is verbalizations that distract a subject from the current activity and draw on the subject's long term memory. Ericsson and Simon argued that reliable verbal reports can be produced during task solving provided they do not require participants to carry out additional cognitive processing. They asserted that only verbalizations from level 1 and 2 are valid as rich data to understand cognitive processes, because verbalizations from level 3 distract the subject [6].

The challenge for usability evaluators is that they can see what a user is doing but not why. Thinking aloud has been introduced to overcome this. The idea is that when users are working with an interactive system they express what their intentions are and from that the reasons behind their difficulties can be inferred and compared with what they are actually doing [21].

In a usability evaluation with the *Traditional* think aloud protocol, the test moderator will have minimal interaction with the users and only use the phrase "Keep talking" when the user stops thinking aloud. The test moderator will not embark on any conversation or provide assistance of any kind. Thinking aloud is usually related to laboratory-based testing, but it has also been used in other settings, e.g. [1, 3, 11].

2.2 Other Think-Aloud Protocols

There is a variety of other think-aloud protocols. Based on field observations of usability evaluations, Boren and Ramey [2] conducted observations of usability practitioners in software companies. They found several discrepancies between the traditional protocol and the way the practitioners conducted the evaluations and in particular the probes they used. They argue that the traditional protocol with an entirely passive listener is unnatural, and "a speech communication perspective may well interpret silence interspersed with commands to keep talking as a more abrasive form of contact" [2, p. 269]. Instead, they propose a protocol that reflects the way human beings naturally communicate, with a combination of statements by a speaker and some of feedback or acknowledgment from a listener. The feedback uses a back channel from the listener to the speaker, confirming for the speaker that there is actually an "active listener". Boren and Ramey [2] suggest that the feedback that is most effective is the quiet, affirming "Um-humm" response given at relevant times; and if the test subject goes quiet, the active listener can repeat the last word the speaker said with a questioning intonation.

In usability evaluations with the *Active listening* think-aloud protocol, the test moderator will be an active listener providing acknowledging expressions. The moderator will not ask questions directly or start a conversation, but only use verbal expressions like "Um-humm" or "ahh". If users stop thinking aloud, the test moderator will repeat the last word expressed by the user.

There are several think-aloud protocols in usability evaluation practice that go beyond the traditional and active listener protocols. In practice, a test moderator will often try actively to get the test subjects to talk about their intentions, thinking, understanding and mental model. This is accomplished by using probes that are much more intrusive. Dumas and Redish [5] present "active intervention" as a technique where the test moderator asks questions in an active manner to get an understanding of the test subjects' mental model of the system and its functioning. Such an intrusive protocol with extensive conversation has been denoted as coaching [19].

In a usability evaluation with the *Coaching* think-aloud protocol, the test moderator will be empathetic and encourage the test subject, e.g. by stating that "everybody has problems with this part of the system" if test subjects express frustration or feel insecure. The moderator may also express sympathy and give confirming feedback, e.g. "well done, you are doing great" when they have completed. When test subjects come to a complete stop, the moderator may assist by asking questions, e.g. "what options do you have?" or "what do you think happens when you click on that menu item?"

2.3 Empirical Studies of Think Aloud Protocols

There is a range of empirical studies of various think-aloud protocols. An early study [25] focused on the coaching protocol, showing that the level 3 verbalizations used in this protocol improved the test subjects' performance, which made them conclude that level 3 verbalizations in usability evaluation imply a bias.

Another early study [9] examined four user-based usability evaluation methods (logged data, questionnaire, interview, and verbal protocol analysis). They found that the verbal protocol analysis was most efficient, and even using two evaluation methods made no statistically significant improvement over the verbal protocol analysis.

A study by Krahmer and Ummelen [14] compared the traditional protocol [6] with active listening [2]. They found that in the active listening condition, more tasks were completed and the participants were less lost on the website. However, the website was untypical, and their active listening protocol included elements of coaching.

A more recent study by Hertzum et al. [10] compared the traditional and the coaching protocol. They found that with coaching, the subjects spent more time solving tasks, used more commands to navigate and experienced higher mental workload.

A study by Rhenius and Deffner [20] focused on the relation between verbalization and where test subjects were looking. By using eye tracking they showed that where test subjects were looking was directly related to what they were verbalizing, and concluded that verbalizations and short term memory are synchronized.

The related work that assesses think-aloud protocols is listed in Table 1. Apparently, these focus mostly on usability measures. We have only found two empirically based assessments of the usability problems identified with different think-aloud protocols. One study compared the traditional with a retrospective think-aloud protocol. The two protocols revealed comparable sets of usability problems [8]. Another study compared the traditional and the coaching protocol. They found that coaching identified more usability problems related to dialogue, navigation, layout and functionality, but the unique problems of this condition were less severe [26]. This limited focus on usability problem identification confirms the general critique of usability research for facing major challenges [7, 24] and having little relevance to practice [23].

Table 1. Overview of the literature that assess think-aloud protocols.

	Description of protocol	Assessment of protocol based on	
		Usability measures	Usability problems
Traditional	[6]	[10, 14, 19]	[8, 15]
Active listening	[2]	[14, 19]	
Coaching	[5]	[10, 18, 19, 25]	[26]

2.4 Empirical Studies of Thinking Aloud in Practice

It has been argued that usability evaluation practice needs level 3 verbalizations despite their influence on task solving. Practitioners argue that level 3 gives the most useful data for identifying usability problems in interactive software systems [19].

There are only few empirical studies of think-aloud protocols in practice. Nørgaard and Hornbæk [18] observed and interviewed practitioners on the way they conducted usability evaluations. They found that practitioners often asked hypothetical or leading questions, which would elicit level 3 verbalizations [19]. Another study explored the use of think-aloud protocols by usability practitioners. They found that the think-aloud technique was generally appreciated, but divergent practice was reported [15].

Several researchers have requested additional research in the effects of various think-aloud protocols in usability evaluations, e.g. [2, 14], as it is not clear which protocol usability practitioners should employ in different situations.

3 Method

We have conducted an empirical study of the three think-aloud protocols in Table 1. The aim was to study the effect of these think-aloud protocols on the identification of usability problems. Part of the study also replicated a study of usability measures [19].

3.1 Experimental Conditions

Our empirical study was designed to compare the following four conditions:

- *Traditional:* the protocol originally suggested by Ericsson and Simon [6], where the test moderator is only allowed to probe with "Keep talking" when the test subject has been quiet for a while.
- *Active listening:* the protocol suggested by Boren and Ramey [2] where the test administration provide feedback or acknowledgment to the test subject by constantly probing with "Um-humm".
- *Coaching:* the protocol that is generally used in usability evaluation practice where the test moderator is engaged in a continuous conversation with the test subject.
- *Silent:* the test moderator does not to talk at any point during the evaluation sessions, except when introducing test subjects to the experiment and tasks. The test subjects are specifically asked not to think aloud.

The first two protocols, Traditional and Active listening, support level 1 and 2 verbalization, thereby providing what is assumed to be the most reliable data. The coaching protocol supports level 3 verbalizations where the probes are expected to initiate use of the long term memory of the test subjects. The Silent condition is a benchmark.

3.2 System

The system evaluated is a data-dissemination website (dst.dk) that provides publicly available statistics about educational levels, IT knowledge and skills, economy, employment situation etc. in Denmark. It is the same type of website as in [19].

3.3 Participants

Users. All participating users were invited through emails distributed to the whole university. In total we recruited 43 participating users divided on four different demographical profiles, including 15 participants from technical and administrative

personnel from different departments, 13 faculty members from Ph.D. students to professors from different departments, and 15 students in technical and non-technical educations. All participants were given a gift with a value corresponding to 20 USD.

We distributed the participants evenly according to their demographic profiles over the four experimental conditions, see Table 2.

Table 2. Distribution of participants on conditions and demographic profiles, TAP = Technical and Administrative Personnel n = number of participants.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Mean age (SD)	34.9 (10.6)	33.2 (10.9)	37.5 (16.5)	40.8 (13)
Females/Males	4/6	5/6	7/4	6/5
TAP/Faculty/Students	5/2/3	3/3/5	4/3/4	4/4/3

Test Moderators. We used four test moderators in this experiment, all of whom had previous experience in moderating usability evaluation sessions. We chose to apply a single test moderator for each condition. To avoid test moderators introducing a bias, none of them took part in planning the study, analysing the data or writing this paper.

Evaluators. The two authors of this paper analysed all video material from the 43 evaluation sessions. Both have extensive experience in analysing video data. To reduce possible bias from learning, the evaluators analysed the collected data in different (random) orders. The moderator/analyst separation is uncommon in practice, but was necessary to reduce bias from experimenter expectancy.

3.4 Setting

All usability evaluations were conducted in a usability lab, see Fig. 1. The test moderator was in the control room (A) and the user in the test room (B). They communicated through microphones and speakers. The rooms were separated by one-way mirrors so the user could not see the test moderator. This physical separation of test subjects and moderators is different from typical practice. It was introduced to remove the influence from the test moderator's body language and other visible expressions.

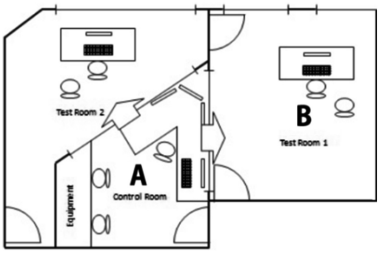

Fig. 1. Layout of the usability laboratory. The test moderator was located in the control room (A) and the user in test room 1 (B).

3.5 Procedure

The usability evaluations were conducted over the course of five days. On the first day of the four test moderators each did two evaluation sessions in order to get acquainted with their protocol. They had received written instructions on how to act in their particular protocol three days in advance, to allow for preparation. After completing their two evaluation sessions, they feedback on how they performed with respect to the protocol. The test moderators did not observe each other's sessions.

Each of the following four days was devoted to evaluations with a particular protocol. All were conducted as a between-subjects study and each session lasted one hour. In all evaluation sessions, the users were first asked to complete two tasks using a similar website, to get acquainted with the protocol applied in their session. They were then asked to solve eight tasks using the main web site. At the end of each session, the user filled in a shortened version of the QUIS questionnaire, applied in [19].

Tasks. The users solved eight tasks that varied in difficulty. For example, the first task was to find the total number of people living in Denmark, while a more difficult task was to find the number hotels and restaurants with 1 single employee in a particular area of Denmark.

3.6 Data Collection

All 43 sessions were video recorded. A questionnaire was collected for each user.

3.7 Data Analysis

The analysis of video material was divided in two parts: A joint analysis of 16 videos followed by individual analysis of the remaining 27 videos. All videos were transcribed in a log file before analysing for usability problems, cf. [22].

Joint Analysis. In order to achieve consistency in usability problem identification with the different protocols, the authors of this paper first analysed 16 videos together (four videos from each condition). These were analysed in random order.

Due to the variety in protocols we found it important to ensure consistency of problem identification in the individual analysis. To support this, we adapted the conceptual tool for usability problem identification and categorization in [22]. We distinguished used the following ways of detecting a usability problem:

(A) Slowed down relative to normal work speed
(B) Inadequate understanding, e.g. does not understand how a specific functionality operates or is activated
(C) Frustration (expressing aggravation)
(D) Test moderator intervention
(E) Error compared to correct approach.

Examples of usability problems are "cannot find the detailed information for a municipality" or "cannot get overview of the presented information".

Individual Analysis. After the joint analysis, we individually analysed the remaining 27 videos in different and random order using the same tool as in the joint analysis. Duplicates where more than one user experienced the same problem were removed (Table 3).

Table 3. Mean any-two agreement between evaluators, n = number of datasets analysed individually by the evaluators.

	Coaching (n = 6)	Active listening (n = 7)	Traditional (n = 7)	Silent (n = 7)
Mean (SD)	0.42 (0.11)	0.44 (0.12)	0.47 (0.08)	0.44 (0.17)

Merging Individual Analysis Results. After the individual analysis, we merged our two individual problem lists. We did this by discussing each identified problem to determine similarities between the two individual lists. Across the 27 videos we had an average any-two agreement of 0.44 (SD = 0.11), which is relatively high compared to other studies reporting any-two agreements between 0.06 and 0.42 [12]. For the agreement between evaluators, we found no significant differences between the four conditions using a one-way ANOVA test.

Calculating QUIS Ratings. The satisfaction score was calculated in the same way as in [19] by combining the sixteen scores from the modified version of the QUIS, each with a score on a Likert scale from 1 to 7. Thus a user could give a total score from 16 to 112 where a high score reflects more user satisfaction with the website.

4 Results

In this section we present our findings on usability measures, i.e. effectiveness, efficiency and satisfaction, as well as on usability problems.

4.1 Effectiveness

We measured the mean number of tasks completed correctly by users in the four conditions. Users in the Coaching condition had a higher completion rate than users in the other three conditions, while users in the Active listening, Traditional and Silent performed similarly. A one-way ANOVA test reveals no significant differences between the conditions for effectiveness.

We also measured the number of times the users gave up while solving a particular task. Users in the Silent condition had a tendency to give up more often than users in the other conditions, while users in the Coaching condition had the lowest rate in this respect. A one-way ANOVA test reveals no significant differences between any of the conditions with respect to the number of times that users gave up.

4.2 Efficiency

Table 4 shows the task completion times in seconds for the four conditions. For each condition, we have calculated the mean value of the total task completion time for each user on all tasks. The users in the Coaching condition on average spent most time while users in the Silent condition had the lowest time. The Active listening and Traditional conditions performed in between these.

Table 4. Mean task completion times in seconds, n = number of users.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Mean (SD)	238 (42)	219 (47)	220 (50)	175 (72)

A one-way ANOVA test reveals no significant differences between any of the four conditions. However, it should be noted that the difference between the Silent and Coaching conditions is close to significant ($p = 0.055$).

4.3 Satisfaction

The QUIS questionnaire included 16 questions where the response for each was a score on a 1–7 Likert scale. Like [19] we added these numbers together, yielding a score between 16 and 112 where a high score should indicate a high level of satisfaction. A one-way ANOVA test of our data reveals no significant differences between median ratings in the four conditions.

The simple sum combines a broad range of different factors. To complement this, we considered a subset of the questions that deal specifically with the users' overall reactions to the website (each on a scale from 1 to 7):

- Terrible – Terrific
- Frustrating – Satisfactory
- Difficult – Easy.

Table 5 shows the overall median scores for these three questions. The Coaching and Traditional conditions gave the highest overall medians, while the Active listening and Silent conditions gave the lowest scores. Note that QUIS measure constructs with more questions than those presented in Table 5. Picking out individual question items would be debatable practice. This is why we considered all three questions of the "overall reactions" subcategory in QUIS, cf. [19].

A one-way ANOVA test reveals significant differences between these one or more conditions ($F(3,119) = 5.29$, 5 % level, $p < 0.002$). We apply a Tukey's pairwise comparison test on all pairs of conditions in order to detect which are significant. The Tukey test reveals significant differences in overall median ratings between Coaching-Active listening ($p < 0.03$) as well as between Coaching-Silent ($p < 0.03$). The Tukey test also suggests significant differences in overall median ratings between

Table 5. Satisfaction ratings given by users for overall reactions to the website, n = number of users.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Terrific-Terrible	4	4	4.5	3
Frustrating-Satisfactory	4.5	3.5	3.5	2
Difficult-Easy	4.5	3.5	4	2.5
Overall Median	**4.5**	**3.5**	**4**	**2.5**

Traditional-Active listening ($p < 0.04$) and Traditional-Silent ($p < 0.03$). There are no significant differences between Coaching and Traditional, and conversely there are no differences between Active listening and Silent conditions. Thus, users in the Coaching and Traditional conditions express higher satisfaction with the website compared to users in the two other conditions.

The questionnaire also included a question on perceived friendliness of the test moderator that was rated on a 7 point scale from unfriendly to friendly. Users in the Coaching condition gave higher ratings than users in all other conditions, but a one-way ANOVA test reveals no significant differences in these ratings.

4.4 Usability Problems

Through the detailed video analysis, we identified a total of 165 usability problems across all four conditions.

Number of Identified Problems. Table 6 shows the mean number of problems identified per user in the four conditions. We identified most problems in Coaching. This is closely followed by Active listening and Traditional, while the Silent condition revealed only around half of the problems identified in the other conditions.

Table 6. Mean number of problems identified using the different TA protocols.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Mean (SD)	39.7 (13.8)	37.6 (14.6)	36.7 (8.9)	18.7 (6.1)

A one-way ANOVA test reveals highly significant differences between one or more of the conditions ($F(3,39) = 1.4$, 5 % level, $p < 0.001$). A Tukey's pairwise test for significance reveals highly significant differences between Silent and Coaching ($p < 0.001$), Silent and Active listening ($p < 0.002$) and Silent and Traditional ($p < 0.004$). We found no significant differences between Coaching, Active listening and Traditional. In other words, the Silent condition revealed significantly fewer usability problems per user compared to Coaching, Active listening and Traditional. This also means that we found no significant difference between any of the think-aloud

protocols. In order to test for type II errors in our statistical analysis, we calculated the β value (β = 0.02) which led to a statistical power of 1 - 0.02 = 0.98. This is above the threshold of 0.8 [4], which indicates that sample size, reliability of measurement and strength of treatment are acceptable.

Agreement Between Users Within Each Condition. Table 7 shows the mean any-two agreement between users in the different conditions. This describes the extent to which users in a condition experienced the same or different usability problems; a high number reflects a large overlap in experienced problems. We found most overlap in the Traditional condition, followed by Coaching and Active listening. The least overlap was found in the Silent condition. Table 8 shows the test statistics comparing each of the conditions with respect to the agreement between user.

Table 7. Mean any-two agreement between users, n = number of all pairs of users.

	Coaching (n = 45 pairs)	Active listening (n = 55 pairs)	Traditional (n = 55 pairs)	Silent (n = 55 pairs)
Mean (SD)	0.18 (0.07)	0.16 (0.09)	0.21 (0.08)	0.14 (0.09)

Table 8. Pairwise comparisons with respect to any-two agreements between users, ▫ = almost significant difference, * = significant difference, ** = highly significant difference.

	Coaching	Active listening	Traditional	Silent
Coaching		p>0.8	p>0.4	P<0.02 *
Active listening	p>0.8		p>0.09	P>0.1 ▫
Traditional	p>0.4	p>0.09		P<0.001 **
Silent	P<0.02 *	P>0.1 ▫	P<0.001 **	

A one-way ANOVA shows significant differences between one or more of the conditions ($F(3,206) = 7.32$, 5 % level, $p < 0.001$). Table 8 shows the results obtained from a Tukeys pairwise comparison test. The Tukey test reveals significant differences between all pairs of conditions and shows significant differences between Silent-Coaching and Silent-Traditional. The differences between Silent-Active listening are close to significant. These findings indicate that users in the Silent condition have a lower overlap in experienced problems compared to the three other protocols that performed similarly.

Agreement Between Conditions. The Venn diagram in Fig. 2 shows the overlap between the three TA protocols where 67 problems out of the 155 (43 %) are common between all. Additionally Coaching revealed 16 unique problems, Active listening 23

and Traditional 28. Thus, the Traditional protocol almost reveals twice as many unique problems compared to the Coaching condition. Note that, for simplicity the Silent condition was intentionally left out in Fig. 2. The Silent condition revealed 64 (39 %) of which 54 were identified through TA protocols.

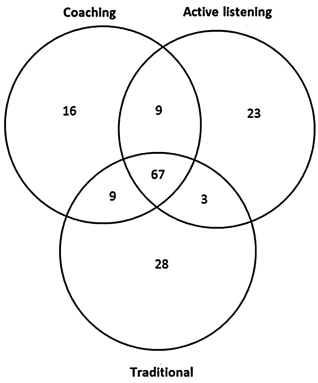

Fig. 2. Venn diagram showing overlap in problems between think-aloud protocols.

4.5 Types of Problems Identified

In this section we provide a qualitative account of the types of usability problems identified through each of the protocols. To cover individual differences between the conditions this qualitative analysis is based on the problems uniquely identified by each protocol.

In general we found that the problems descriptions of the Silent condition did not extend beyond the notes made in log files. We would observe a problem when users made an error, e.g. choosing a wrong link. A typical example is *"Wants to find the number of unemployed persons in 2007. Selects the wrong link, but returns to the previous page"*. This is very similar to the entries made in the log files for each user, i.e. they are pure observations, which did not provide enough details to extend problem descriptions with interpretations of the cause of the problems.

In contrast, we found that the Coaching condition in general led to identification of several types of problems. This protocol primarily led to the identification of problems regarding unclear information, e.g. *"Does not understand the meaning of the rows in the table"*. The coaching condition also led to the identification of problems regarding affordance, e.g. *"Wrongfully believes that more details can be obtained by clicking on a table row"*. Finally, Coaching also revealed problems related to visibility such as *"List of tables is immense, difficult to locate the correct one"*.

In case of the Active Listening condition, we found that most problems concerned unclear information. An example of such a problem is *"Difference between menu options is unclear"*. Otherwise there were relatively few problems in other categories such as affordance and visibility.

The Traditional condition primarily led to the identification of problems related to visibility, e.g. *"Does not notice heading of the graph and is therefore looking up numbers in the wrong location"*. Similar to Coaching, the Traditional condition also revealed problems from the other categories concerning unclear information and affordance.

As described above, we found that the three TA conditions led to data from which we could enrich problem descriptions to include notions of why a particular problem was observed. The opposite was the case for the Silent condition where we could observe a problem but not interpret why it occurred. Furthermore we found that the Coaching and Traditional conditions were similar in terms of the breadth of problem types identified. The Active Listening protocol primarily revealed problems related to unclear information.

Problem Severity. We categorized all identified problems according to critical, serious and cosmetic [16]. We found that the think-aloud protocols reveal about twice as many critical (mean = 23.7, SD = 2.3), serious (mean = 122.7, SD = 5.6) and cosmetic (mean = 248.7, SD = 8.6) problem instances as the Silent condition. This follows the tendency of the mean number of problems identified per user (see Table 6). There is a comparable distribution of the severity of problems in each of the conditions, e.g. 9 % of problem instances in the Coaching condition are critical, which is comparable to the 7 % in the Silent condition. Thus, we found no significant differences in this respect.

4.6 Categories of Problem Observations

We also categorized the identified usability problems according to the five ways of detecting a usability problem: (A) Problem identified due to time delay, (B) Shortcomings in a user's understanding of system functionality, (C) User frustration, (D) Test moderator assistance and (E) User makes an error.

As an example, we found that 29 % of all problems identified through the Coaching condition were observed based on time delays (category A). For this category of observations, we found no significant differences between conditions. This was also the case for category B observations.

As another example, we found that 8 % of the problems identified in the Traditional condition were observed on the basis of user frustration (category C). In this case a one-way ANOVA revealed significant differences between conditions ($F(3,39) = 3.2$, 5 % level, $p < 0.04$). A Tukey's pairwise comparison test reveals significant differences between the Silent and Traditional conditions ($p < 0.03$), but no differences between other pairs.

The Coaching condition is characteristic in allowing the test moderator to help test subjects. This is also reflected by the fact that 10 % of the problems identified in the Coaching condition were observed by test moderator assistance (category D).

Finally, it can be seen that 56 % of all problems found in the Silent condition have been observed due to a user making an error (category E). Here we found significant differences between one or more conditions ($F(3,39) = 15.7$, 5 % level, $p < 0.001$).

A Tukey's pairwise comparison test shows significant differences between the Coaching condition and all other conditions with $p < 0.001$ in all cases.

These results show that there are no differences between the four conditions on the proportion of usability problems identified because users were (A) slowed down relative to normal work speed or (B) demonstrated an inadequate understanding of the system. With problems identified because users (C) expressed frustration, significantly more problems were identified in the Traditional condition compared with Silent. For problems identified because of (D) test moderator intervention, there were significantly more problems identified with the Coaching condition compared to all the other. Finally, for problems identified because (E) users made errors compared to correct approach, the Silent condition identified significantly more compared to Coaching.

4.7 Limitations

The empirical results presented above have a number of basic limitations. The study involves three specific think-aloud protocols, and there are many alternatives. We chose the three because they are either classical or commonly used by usability practitioners [5, 19, 21]. Our aim of replicating the previous study made us consider the same protocols as they used, cf. [19].

In each condition, there was one test moderator, and this person was different between the conditions. This was chosen to ensure consistency across all users in each condition. We could have used the same moderator in all conditions, but that could potentially cause confusion and make the conditions overlap. We tried to reduce a possible effect of this by training and assessing the moderators in accordance with the protocol before the tests were conducted.

The empirical study was only based on a single data-dissemination website where users solved 8 pre-defined tasks. To enable replication, we chose the same type of website as the previous study and used exactly the same tasks.

We had 43 users participating in the study. They are not a random or representative sample of the entire population of Denmark. The number of users is limited, but this is comparable to the empirical studies we have in our list of references, where the number of users is between 8 and 40, except for one study based on 80 users [19].

5 Discussion

In this section, we compare our results to some of the related work and discuss implications for practitioners.

5.1 Comparison with Related Work

Our empirical study has focused on the consequences of using different think-aloud protocols both on the usability measures and on the usability problems identified.

For usability measures, we have found that users have the highest effectiveness, i.e. they complete more tasks in the coaching condition compared to all other conditions

and that active listening, traditional and silent perform similarly. This is in line with the study presented in [25] where it was found that the coaching protocol improved the test subjects' performance. The study in [19] compared the traditional, active listening and coaching protocols and results from that experiment shows a significantly higher level of effectiveness in the coaching condition than in the other conditions, while there were no significant differences between the traditional and active listening protocols [19]. Thus, we found similar tendencies in favor of the coaching condition in our study but the differences we found were not significant. This is supported in [26] where the traditional and coaching protocols are compared. Furthermore, the study in [14] found a discrepancy compared to our study. In that study the traditional and active listening protocols are compared and it is found that more tasks were completed with active listening. In our study we found almost no difference in effectiveness between these two protocols. However, the result in [14] is not entirely reliable as the active listening protocol included some elements of coaching.

In terms of efficiency we found similar performance between all the think-aloud protocols, while users in the silent condition had considerably lower task completion times. This is similar to [8], who showed that the traditional protocol, due to the requirement of thinking aloud while working, had a negative effect on task performance. Similarly, in [10] the traditional and coaching protocols are compared and findings from that study show that the coaching protocol resulted in higher task completion times. Thus, we do witness similar tendencies as reported in the above literature. However, like [19] we found no significant differences in terms of efficiency.

For the satisfaction ratings, we found that users in the coaching and traditional conditions were significantly more satisfied with the website compared to users in active listening and silent. In a similar study [19] partially agrees they found users in the coaching condition to be most satisfied [19].

For usability problems, we found that the silent condition revealed significantly fewer usability problems than any of the other conditions; and even when we were able to observe problems in that condition, we were often unable to explain why they occurred. This is also supported by the finding that most problems were observed simply by users making errors (category E observations). The study in [8] compared the traditional with a retrospective think-aloud protocol, which revealed comparable sets of usability problems. In contrast, another study has shown that different protocols can reveal different types of problems. The study presented in [26] compared the traditional and coaching protocols and found that the coaching protocol led to a larger number of usability problems related to dialogue, navigation, layout and functionality. Additionally, they found that the problems which were unique to the coaching condition were mainly at a low level of severity [26]. This is partly supported by our findings that the proportion of critical, serious and cosmetic problems is distributed similarly within conditions, also in case of the unique problems identified in each condition. However, we did find the types of identified problems to be similar between the coaching and traditional conditions while the active listening condition mainly led to problems concerning unclear information. In terms of problem severity, we found that the all think-aloud protocols revealed twice as many critical, serious and cosmetic problems as the silent condition.

5.2 Implications for Usability Practitioners

The most interesting result for usability practitioners is that the think-aloud protocols had only limited influence on user performance and satisfaction, but compared to the silent condition, they facilitated identification of the double number of usability problems with the same number of test subjects. Recruiting test subjects is a major task in usability evaluation, thus it is important to know how test subjects can be used most effectively. Furthermore, we found it difficult to interpret the causes of problems, which were observed through a lesser richness in problem descriptions.

Our results cannot be used to suggest practitioners to use a specific protocol. For example, the proponents of the Active listening protocol have argued that the Traditional protocol is unnatural. However, our results do not support this. The two protocols facilitate identification of a similar number of usability problems and richness in problem descriptions.

There were some interesting differences in the usability problems that were identified. The Traditional protocol revealed more usability problems in the frustration category, the Coaching protocol revealed more problems identified through test monitor intervention, and the Silent conditions mainly found usability problems identified when users made errors.

The results demonstrate that no single protocol version is superior overall; each has strengths and weaknesses. Yet in the silent condition, we could not capture verbalised problems, because the test subjects were silent, so the comparison is limited by this.

6 Conclusion

We have conducted an empirical study comparing three think-aloud protocols and a silent control condition. The study assessed the usability problems identified in each condition based on a systematic video analysis. We found that the differences between the three think-aloud protocols were limited. Contrary to the general emphasis on the Coaching protocol, we have no indication that it is superior for identifying usability problems. In fact, there were some aspects where the Traditional protocol performed surprisingly well. Overall, the think-aloud protocols are clearly superior to the silent condition; with the same number of test subjects, they revealed the double number of usability problems compared to the silent condition, and this applied across all levels of severity. Part of the study replicated a previous study on usability measures [19]. Here, we found only a few differences between the protocols. There was an indication that efficiency was higher in the Silent condition and that satisfaction with the website was higher for Coaching and Traditional. These limited effects do not contradict the literature as previous studies point in various directions.

Our study has a number of limitations. First, identification of usability problems is not as mechanical as determining usability measuring. Second, the specific experimental design impose limitations on our results, especially with our role as analysts, the moderator/analyst separation, and the physical moderator/user separation, but these were necessary to reduce potential sources of bias.

We deal with usability evaluation and the process of identifying usability problems. This is a controversial topic in HCI where some question the whole approach and its relevance to practice. Contrastingly, others maintain that within certain limits such studies produce relevant results. In order to achieve that, it is not possible to directly copy approaches from practice in the laboratory.

Our study points to interesting directions for future work. Most importantly, the consequences for and approaches used in practice should be studied extensively. It is also highly relevant to replicate the study with other types of systems.

Acknowledgments. We are grateful to Lise Tordrup Heeager, Rasmus Hummersgaard, Rune Thaarup Høegh Mikael B. Skov, Henrik Sørensen and the 43 users who helped us in the study.

References

1. Andreasen, M.S., Nielsen, H.V., Schrøder, S.O., Stage, J.: What happened to remote usability testing? an empirical study of three methods. In: Proceedings of Conference on Human Factors in Computing Systems 2007 (CHI 2007), pp. 1405–1414. ACM Press, New York (2007)
2. Boren, T., Ramey, J.: Thinking aloud: Reconciling theory and practice. IEEE Trans. Prof. Commun. **43**(3), 261–278 (2000)
3. Bruun, A., Gull, P., Hofmeister, L., Stage, J.: Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In: Proceedings of Conference on Human Factors in Computing Systems 2009 (CHI 2009), pp. 1619–1628. ACM Press, New York (2009)
4. Cohen, J.: Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale (1988)
5. Dumas, J., Redish, J.: A Practical Guide to Usability Testing. Intellect Press, Portland (1999)
8. Ericsson, K.A., Simon, H.A.: Protocol Analysis: Verbal Reports as Data, revised edn. MIT Press, Cambridge (1996)
7. Gray, W.D., Salzman, M.C.: Damaged Merchandise? A review of experiments that compare usability evaluation methods. Hum. Comput. Interact. **13**(3), 203–261 (1998)
8. van den Haak, M.J., de Jong, M.D.T., Schellens, P.J.: Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav. Inf. Technol. **22**(5), 339–351 (2003)
9. Henderson, R.D., Smith, M.C., Podd, J., Varela-Alvarez, H.: A comparison of the four prominent user-based methods for evaluating the usability of computer software. Ergonomics **38**(10), 2030–2044 (1995)
10. Hertzum, M., Hansen, K., Anderson, H.: Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? Behav. Inf. Technol. **28**(2), 165–181 (2009)
11. Hertzum, M., Holmegaard, K.D.: Thinking aloud in the presence of interruptions and time constraints. Int. J. Hum.-Comput. Interact. **29**(5), 351–364 (2013)
12. Hertzum, M., Jacobsen, N.E.: The evaluator effect: a chilling fact about usability evaluation methods. Int. J. Hum. Comput. Interact. **15**, 183–204 (2003). Taylor & Francis
13. ISO 9241-11 (1998) Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability. ISO (1998)
14. Krahmer, E., Ummelen, N.: Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. IEEE Trans. Prof. Commun. **47**(2), 105–117 (2004)

15. McDonald, S., Edwards, H., Zhao, T.: Exploring think-alouds in usability testing: an international survey. IEEE Trans. Prof. Commun. **55**(1), 2–19 (2012)
16. Molich, R.: User-Friendly Web Design (in Danish). Ingeniøren Books, Copenhagen (2000)
17. Nielsen, J.: Usability Engineering. Academic Press, Cambridge (1993)
18. Nørgaard, M., Hornbæk, K.: What do usability evaluators do in Practice? An explorative study of think-aloud testing. In: Proceedings of DIS 2006, pp. 209–219. ACM Press, New York
19. Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S., Ashenfelter, K.T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: Proceedings of CHI 2010, pp. 2381–2390. ACM, New York
20. Rhenius, D., Deffner, G.: Evaluation of concurrent thinking aloud using eye-tracking data. In: Proceedings of Human Factors Society 34th Annual Meeting, pp. 1265–1269 (1990)
21. Rubin, J., Chisnell, D.: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, Hoboken (2008)
22. Skov, M.B., Stage, J.: A conceptual tool for usability problem identification in website development. Int. J. Inf. Technol. Web. Eng. **4**(4), 22–35 (2009)
23. Wixon, D.: Evaluating usability methods: why the current literature fails the practitioner. Interactions **10**(4), 29–34 (2003)
24. Woolrych, A., Hornbæk, K., Frøkjær, E., Cockton, G.: Ingredients rather than recipes: A proposal for research that does not treat usability evaluation methods as indivisible wholes. Int. J. Hum.-Comput. Interact. **27**(10), 940–970 (2011)
25. Wright, R., Converse, S.: Method bias and concurrent verbal protocol in software usability testing. In: Proceedings of Human Factors Society 36th Annual Meeting, pp. 1220–1224 (1992)
26. Zhao, T., McDonald, S., Edwards, H.M.: The impact of two different think-aloud instructions in a usability test: a case of just following orders? Behav. Inf. Technol. (2012)