ACM Transactions on Computer-Human Interaction, vol. 31, no. 3 (2024), article 37
Author version. Published version at: https://doi.org/10.1145/3665327

Concurrent or Retrospective Thinking Aloud in Usability Tests? A Meta-Analytic Review

Morten Hertzum

Roskilde University, Roskilde, Denmark, mhz@ruc.dk

Abstract. In usability tests, the users are commonly asked to think aloud to let the evaluator listen in on their thoughts. Two variants of this procedure involve that the users either think aloud while using the tested product (concurrent thinking aloud, CTA) or after using it (retrospective thinking aloud, RTA). This study reviews the studies that compare CTA and RTA to investigate what is gained and lost by using one or the other variant in a usability test. A total of 29 studies, reporting from 42 comparisons of CTA and RTA, matched the inclusion criteria and were included in the meta-analyses. The main differences are that for CTA task time is longer, but total time shorter, whereas for RTA the users verbalize more explanations, problem formulations, and design recommendations. In addition, CTA users probably experience the evaluator's presence as less disturbing than RTA users do.

CCS concepts: Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods → Usability testing

Keywords: Thinking aloud, Concurrent thinking aloud, Retrospective thinking aloud, Usability test, Usability, User experience

1 Introduction

To work systematically with usability, designers must be able to examine whether a proposed design is usable or will cause problems for its users. A widely used method for identifying usability problems is the usability test. In a usability test, the users are commonly asked to think aloud in order for the evaluator to be able to listen in on their thoughts as well as observe their behavior [19,37]. The users may be asked to think aloud while using the tested product or after using it, that is, either concurrently or retrospectively. These two variants of thinking aloud represent different tradeoffs in the planning of a usability test. Many studies have investigated these tradeoffs for either one or the other variant. Fewer studies have compared concurrent and retrospective thinking aloud. The present study reviews these comparisons to investigate what is gained and lost by using concurrent thinking aloud (CTA) or retrospective thinking aloud (RTA) in a usability test.

While thinking aloud is informative for the evaluator, it is extra work for the users. It has been a topic of considerable debate under what conditions this extra work provides an accurate account of the users' thought process without altering their behavior and performance [e.g., 7,24,29]. In CTA, the extra work is performed along with the use of the tested product. The concurrency provides good conditions for the users to account for their thought process. However, this account may alter their behavior and performance, thereby jeopardizing the usability test. In RTA, the users first use the tested product without thinking out loud and only then provide an account of their thought process. That is, the extra work of thinking aloud cannot alter their behavior and performance. However, the users may not recall their thought process fully, thereby providing less informative and less accurate information about it. Surveys of usability practitioners show that they use CTA much more than RTA [26,51]. A practical argument raised against RTA is that it takes too long [51].

This review provides a meta-analysis of the studies that compare CTA and RTA. Meta-analyses use statistical techniques to combine the results from existing studies into a quantitative estimate of the overall effect of a variable. A meta-analysis is considered appropriate because the individual studies comparing CTA and RTA often produce results that conflict with those from some of the other studies. Thus, a systematic approach is needed to determine the accumulated results. The present review covers the results that have been accumulated about the following aspects of using CTA or RTA in a usability test:

- *Task performance*, that is, task success, task time, and total time. The issue in this part of the review is test reactivity – do CTA and RTA differentially affect how the users perform the test tasks?
- *Usability problems*, that is, the number of problems, detection rate, problem severity, problem types, and the source of information about the problems. The issue in this part of the review is test effectiveness – are CTA and RTA equally effective at identifying usability problems?
- *User verbalizations*, that is, the number and content of the verbalizations. This part of the review is about the data obtained with CTA and RTA – do the users talk to the same extent about the same kinds of subjects?
- *User experience*, that is, the users' experience of thinking aloud, their working procedure, and the evaluator's presence. This part of the review is about the acceptability of the test format to the users – how do they experience taking part in a CTA or RTA session?

The next section provides background about CTA and RTA and presents a taxonomy of thinking-aloud variants. The variants make thinking aloud applicable under different conditions and differ in how they balance validity against value to usability testing. Section 3 accounts for the review method. It describes the inclusion criteria, paper-selection process, and data analysis. Section 4 is the review. It provides the results of the meta-analysis and covers the four aspects listed above. Section 5 starts with a summary of the review results and then discusses how they relate to the maxims of making usability tests valid, robust, complete, low-cost, and impactful. The review ends with a discussion of its implications and limitations.

2 Background

Thinking aloud has been part of usability testing since the early 1980s [47] but dates back much further in psychology [69]. To understand thinking aloud, Ericsson and Simon [24] introduced a distinction among three types of verbalization: the verbalization of information that is already in a person's present focus of attention in verbal form (Level 1), the verbalization of information that is already in the person's present focus of attention but in nonverbal form (Level 2), and the verbalization of information that is not in the person's present focus of attention (Level 3). According to Ericsson and Simon [24], Level 1 verbalization does not introduce additional mental processing and can, thus, be made without altering the thought process that goes into performing a task. Similarly, Level 2 verbalization merely involves recoding the information into verbal form. This recoding involves mental processing but does not bring new information into the person's focus of attention. Thus, it does not alter the thought process. In contrast, Level 3 verbalization introduces mental processing that brings new information into the person's present focus of attention. The new information may for example be explanations, generalizations, assumptions, reasons, and summaries. Attending to this additional information is an alteration of the user's thought process compared to performing the same task without thinking aloud. The altered thought process may, in turn, alter the user's behavior and performance. On this basis, Ericsson and Simon [24] contend that Level 3 verbalization should be avoided and thinking aloud restricted to verbalizations at Levels 1 and 2.

In usability testing, this contention has been debated for at least three reasons. First, it has proven difficult to instruct and train users to restrict their verbalizations to Levels 1 and 2 (aka classic thinking aloud). Several studies that aim for classic thinking aloud report that users also verbalize explanations, user experiences, and redesign proposals [38]. Such verbalizations are at Level 3 and, thus, extend the thinking aloud to Levels 1 to 3 (aka relaxed thinking aloud). Willis and McDonald [70] suggest that relaxed thinking aloud is difficult to avoid in usability tests because the users are aware that the aim of the test is to assess the product; they may therefore deem their reflections important even when they are not directly solicited. That is, the users may approach the test tasks in a more self-reflective manner than if they were performing the tasks outside of a usability test. These self-reflections may be verbalized on the users' own initiative or in response to prompts from the evaluator. In both cases, they constitute Level 3 verbalization because they bring additional information into the user's focus of attention compared to the information involved in performing the tasks in non-test situations.

Second, some studies question whether classic thinking aloud leaves behavior and performance unaltered. These studies for example find that classic thinking aloud impairs users' performance on spatial tasks [31], influences their perception of time [40], and alters their breaking and acceleration behavior during driving tasks [64]. That said, most studies find that classic thinking aloud does not alter performance, except by prolonging it [24,29]. In contrast, many studies document that relaxed thinking aloud alters behavior and performance [e.g., 1,39,57].

Third, relaxed thinking aloud allows for relevant and informative verbalizations that are excluded from classic thinking aloud. To complement the users' observable behavior, usability evaluators are interested in why the users behave the way they do. To understand the user experience, evaluators also need information about the users' affective response to the product. Thus, many usability evaluators explicitly ask the users to verbalize reasons, reflections, and experiences, using prompts such as "John, could you tell us why you pressed the enter key?" [19], "Did you notice this column?" [54], and "Do you think this was easy or difficult to find?" [41]. These prompts directly solicit relaxed thinking aloud and stand in clear contrast to the neutral "keep talking" prompt recommended to elicit classic thinking aloud.

The risk that thinking aloud alters behavior and performance is specific to CTA. If the users instead think aloud after performing a task, then their behavior and performance are shielded from their verbalizations. By having users think aloud retrospectively, it becomes possible to get the additional content offered by relaxed thinking aloud without jeopardizing behavior and performance. However, relaxed thinking aloud will still alter RTA users' thought process because it is extended with reflections that were not in the users' focus of attention when they performed the task. For classic thinking aloud, studies have investigated the accuracy of the users' verbalizations by comparing them with an independent record of their thought process, typically obtained by eye-tracking the users while they solve the tasks. These studies find a good match between verbalizations and fixation sequences during both CTA [e.g., 16,23] and RTA [e.g., 32]. For example, Guan et al. [32] found an 88% overlap between the screen elements referenced in the users' verbalizations and the screen elements on which the users fixated. The remaining verbalizations were misstatements (9%), in which the users' verbalizations included screen elements in between those appearing in their fixation sequences, and fabrications (3%), in which the users' verbalizations included screen elements not in their fixation sequences. Relatedly, McDonald et al. [52] found just 2.4% inaccuracies (e.g., instances of forgetting) in the verbalizations from RTA compared to those in CTA.

Figure 1 provides a taxonomy of the main variants of thinking aloud. The present study will use this taxonomy to classify how thinking aloud is approached in the reviewed papers. For both CTA and RTA, thinking aloud can be either classic or relaxed. The choice between these two variants is largely about whether the main goal is to avoid altering the thought process or obtain additional information about it [69]. For RTA, the temporal separation between task performance and thinking aloud introduces additional variants.

First, the users can perform RTA by recalling their thought process without support from external cues or with such cues. The most common cues are to show the users a video recording of their task performance, possibly overlaid with a recording of their eye movements. Olsen et al. [58] report more verbalization and the identification of more usability problems with video cues and with gaze cues than without cues. It has also been reported that video cues make the prospect of RTA less daunting: "I wasn't daunted because you have the replay; without that it would have been a different prospect" [70]. Comparing video cues with gaze cues, Elbabour et al. [21] report that gaze cues help users recall details in their behavior and help evaluators identify more usability problems, especially minor navigation and comprehension problems. In contrast Elling et al. [22] found no difference between video cues and gaze cues on the number and types of problems identified.

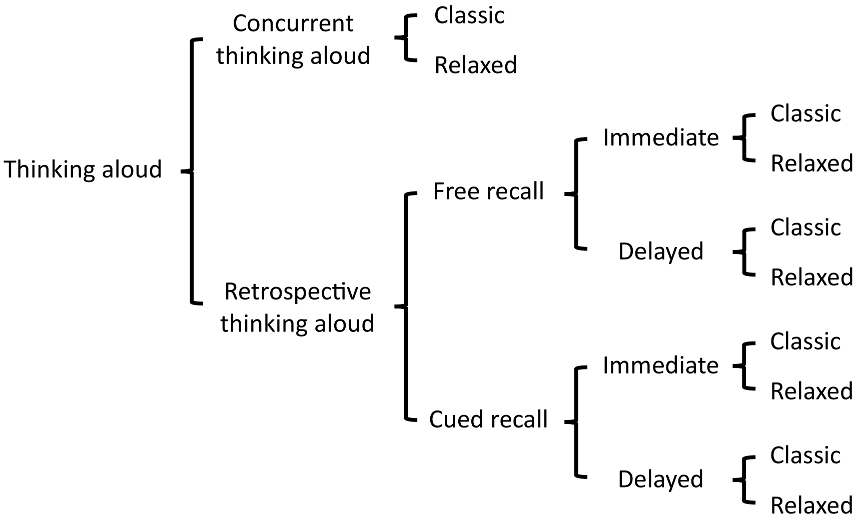


Figure 1. Taxonomy of thinking-aloud variants, adapted from Bruun et al. [10]

Second, the users can perform RTA immediately after task performance or with some delay between task performance and thinking aloud. To avoid memory decay, the users normally think aloud immediately after task performance [e.g., 21,32,36]. However, Ohnemus and Biers [55] compared thinking aloud immediately after task performance with thinking aloud 24 hours after task performance and found no difference in the time users spent verbalizing and no difference in the value of the verbalizations to designers. Willis and McDonald [70] compared thinking aloud after each task with thinking aloud at the end of the session, that is, after all tasks had been performed. When the users thought aloud after each task, they performed tasks slower, made more errors, and verbalized more explanations and expectations.

3 Method

The 29 papers included in this review were selected and analyzed through a process that involved formulating inclusion criteria, inspecting a total of 5382 papers for inclusion or exclusion, and analyzing the contents of the included papers.

3.1 Inclusion criteria

The selection of papers for inclusion in the review was governed by five criteria, formulated prior to the selection process. To be included, a paper had to meet all five criteria. First, papers had to compare CTA and RTA with each other. Second, papers had to report on empirical studies. Third, papers had to study thinking aloud in the context of usability testing. Fourth, papers had to be peer-reviewed research published in journals, at conferences, or as book chapters. Fifth, papers had to be in English.

3.2 Selection procedure

The process of selecting the papers for inclusion in this review involved multiple steps, see Figure 2. First, five databases were searched for papers containing the terms “concurrent”, “retrospective”, “thinking aloud” (or “think aloud”), and “usability test” (or “usability testing” or “thinking-aloud test”, or “think-aloud test” or “usability study” or “thinking-aloud study”, or “think-aloud study” or “usability evaluation”) anywhere in the paper. The five databases were ACM Digital Library, Google Scholar, Sage Pub, Scopus, and Web of Science. They were searched on April 25, 2023. These databases were chosen for their coverage and because they were known to include at least some studies on usability testing. Second, duplicates were removed from the set of 3306 papers that matched the queries, and then the unique papers were screened. The papers were screened by matching their title against the inclusion criteria and, if they passed this screening, by matching their abstract against the inclusion criteria. After the two screenings, 37 papers remained. Third, all papers that referenced these 37 papers were screened for inclusion. This step, formally known as forward chaining, was performed to capture papers that employed a terminology different from the query in the first step and to strengthen the inclusion of recently published papers. There were 2076 references to the 37 papers, according to the 'cited by' feature in Google Scholar on May 3, 2023. After duplicate removal, title screening, and abstract screening, 40 of these 2076 papers remained. Fourth, the 37 papers from the second step (database search) and the 40 papers from the third step (forward chaining) were added together. Because 30 papers were in both sets, the combined set contained 47 papers. Fifth, the full text of these 47 papers was looked up. All of them could be obtained. Sixth, the full text of the papers was matched against the inclusion criteria. After this final matching, 29 papers remained. They were included in the meta-analysis.

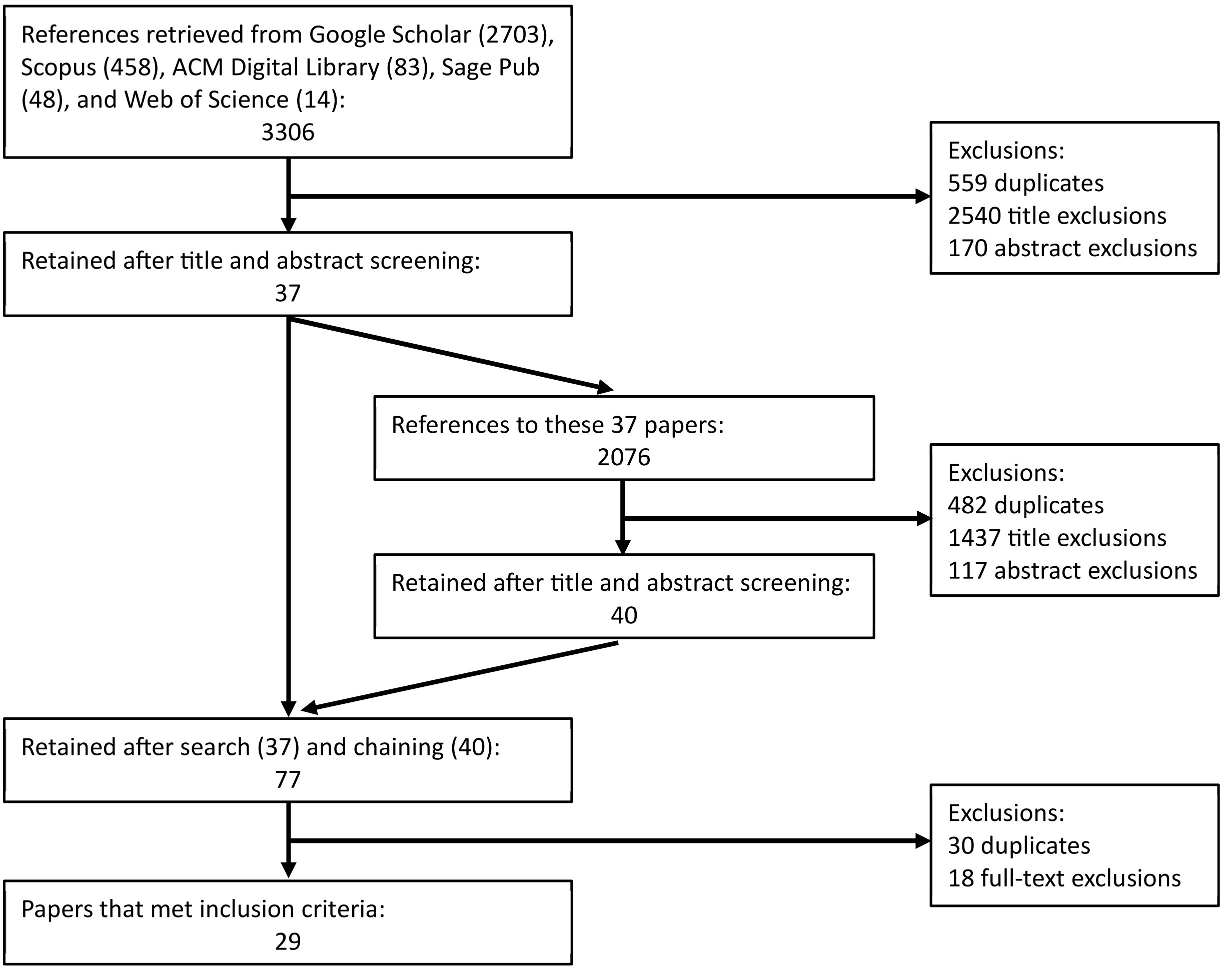

References retrieved from Google Scholar (2703), Scopus (458), ACM Digital Library (83), Sage Pub (48), and Web of Science (14):
3306
Exclusions:
559 duplicates
2540 title exclusions
170 abstract exclusions
Retained after title and abstract screening:
37
References to these 37 papers:
2076
Exclusions:
482 duplicates
1437 title exclusions
117 abstract exclusions
Retained after title and abstract screening:
40
Retained after search (37) and chaining (40):
77
Exclusions:
30 duplicates
18 full-text exclusions
Papers that met inclusion criteria:
29

Figure 2. Paper-selection process

3.3 Data analysis

The analysis of the 29 included papers proceeded in four steps. First, several papers reported results for more than one comparison between CTA and RTA. Some contained parallel comparisons of multiple systems, others tested the same system with different user groups, and still others compared CTA with multiple variants of RTA. These studies were split into one case for each comparison of CTA and RTA. In total, the 29 papers contained 42 cases.

Second, general information about each case was extracted from the papers. This information included the kind of system that was tested, the number of users in the test, and the classification of the compared CTA and RTA variants according to the taxonomy in Figure 1. In some cases [3,20,30,43,44,55,56,72], the methodological description of CTA or RTA left it unclear whether the users were instructed to do classic or relaxed thinking aloud. To err on the side of caution, these cases were classified as relaxed thinking aloud. For example, the thinking aloud in the RTA part of the study by Hyrskykari et al. [43] was classified as relaxed because their methodological description merely stated that the RTA users were "asked to think aloud or comment what they were thinking about during the test."

Third, data for the individual parts of the meta-analysis were extracted from the papers. To safeguard against errors, this was done in multiple rounds, each restricted to one part of the meta-analysis. In the first round, the papers were inspected for data about task performance. During this first round, the authors of two papers were contacted to confirm details about their study. In subsequent rounds, data were extracted for the other parts of the meta-analysis. The extraction process involved copying data that were directly available and, in some cases, calculating needed data from the data available. For example, data about the number of verbalizations in different content categories were converted to percentages. In addition, the direction of the user-experience ratings from several studies was reversed to obtain a consistent direction across all available ratings.

Fourth, the extracted data were analyzed statistically. The statistical analyses followed standard meta-analytic procedures [49,63] and were made with SPSS version 28.0.1.0. They involved determining the effect size of each study and estimating the overall effect size. For dichotomous variables, the effect size of each study was measured by the logarithm of the risk ratio. For example, the effect size for task success was measured by log(RTA success rate/CTA success rate). Such effect sizes are symmetric around zero, which represents equal performance with CTA and RTA. For continuous variables such as task time and user experience, the effect size of each study was measured by the standardized mean difference. The standardized mean difference was calculated as Hedges' *g*, that is, as the difference between the RTA and CTA means divided by the pooled standard deviation. These effect sizes are also symmetric around zero. For both dichotomous and continuous variables, the overall effect size was estimated by weighing each study with its inverse variance weight. This way, studies with lower variance (i.e., with more precise results) received higher weight. In addition, the estimation of the overall effect size involved the Hedges adjustment to compensate for small sample size and the Knapp-Hartung adjustment of the standard error.

4 Results

The 29 reviewed studies are listed in the appendix. They were conducted in Europe [2,3,43,44,46,52,60–62,65,66,4,6,18,20,33–36], North America [9,11,27,30,55,56,59], and Asia [13,45,72] during the 1990s [9,55,59], 2000s [11,18,20,33–36,43,44,46], 2010s [2,3,65,66,72,4,6,30,45,52,56,60,61], and 2020-2023 [13,27,62]. The evaluated systems were websites [2,3,56,60,65,66,4,20,33–36,43,52], games [13,18,27,46,62], office applications [9,11,30,72], healthcare systems [6,44,61], and other systems [45,55,59].

4.1 Task performance

The users in the reviewed studies performed specified tasks with the tested systems. Ideally, the users' task performance should be unaffected by the usability test, but their task performance may, inadvertently, be affected by CTA and RTA. In addition to task success and task time, the total time for CTA and RTA sessions was compared.

4.1.1 Task success rate

The users' task success rate (i.e., the percentage of correctly solved tasks) was reported for CTA and RTA in 16 cases, see Figure 3. Success rates varied between 32% and 92% (CTA) and between 34% and 100% (RTA), thereby indicating large cross-study differences in task difficulty or system usability. The studies that involved classic thinking aloud during both CTA and RTA tended to have smaller effect sizes (i.e., less difference in success rate between CTA and RTA) than the studies that involved relaxed thinking aloud during both CTA and

RTA. For all but one study, the 95% confidence intervals included 0, thereby indicating a significant difference in task success rate for this one study only. In this study, users achieved a 100% success rate during video-cued RTA compared to a 42% success rate during CTA [20]. The overall effect size across the studies in Figure 3 was 0.04 with a 95% confidence interval of [-0.06, 0.14]. The 95% confidence interval included zero and, thereby, indicated that the task success rate did not differ significantly between CTA and RTA at a *p*-level of .05. In addition to the studies in Figure 3, Ohnemus and Biers [55] also found no significant difference in task success rate between CTA and RTA (their study could not be included in the above analysis because they did not report the success rates).

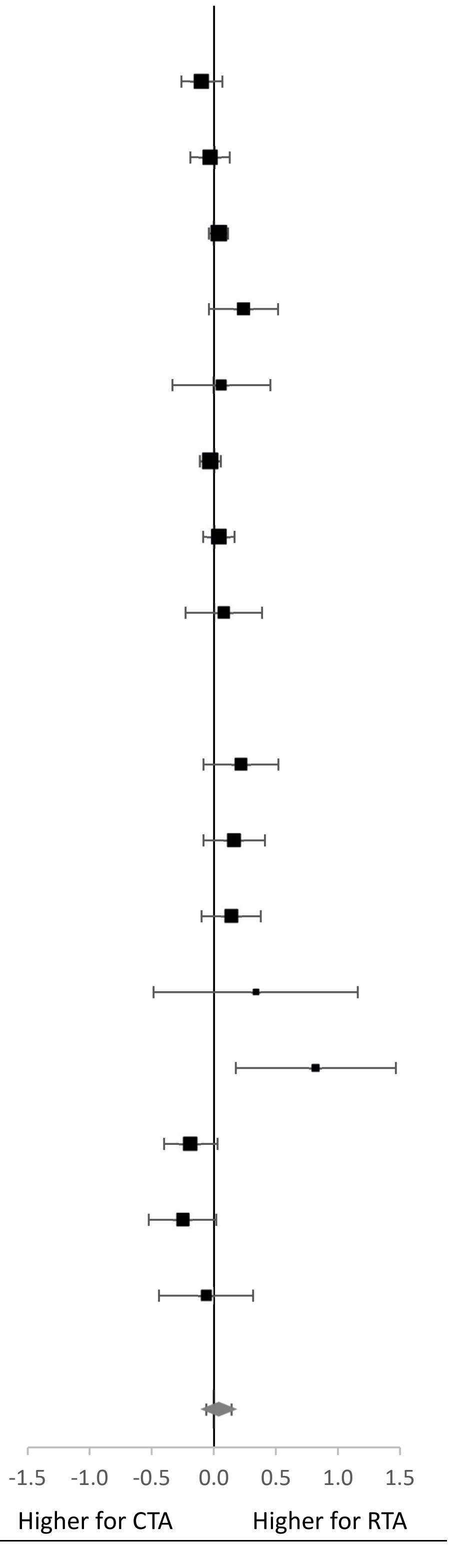
Study	Success rate (%)		Effect size [a]	Forest plot [b]
	CTA	RTA		
Group: classic CTA, classic RTA				
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	70	64	-0.10 [-0.26, 0.07]	
Alhadreti and Mayhew [2] – CTA vs RTA after first doing CTA	70	68	-0.03 [-0.19, 0.13]	
Alshammari et al. [4]	92	96	0.04 [-0.04, 0.11]	
van den Haak et al. [33]	37	47	0.24 [-0.04, 0.52]	
van den Haak et al. [34]	32	34	0.06 [-0.33, 0.46]	
van den Haak et al. [35]	83	81	-0.03 [-0.11, 0.05]	
van den Haak et al. [36]	67	70	0.04 [-0.09, 0.17]	
Prokop et al. [62]	67	72	0.08 [-0.23, 0.39]	
Group: relaxed CTA, relaxed RTA				
AlRoobaea et al. [3] – BBC website	46	58	0.22 [-0.08, 0.52]	
AlRoobaea et al. [3] – Skool website	56	66	0.16 [-0.08, 0.41]	
AlRoobaea et al. [3] – Academic Earth website	59	68	0.14 [-0.10, 0.38]	
Eger et al. [20] – CTA vs gaze-cued RTA	42	60	0.34 [-0.49, 1.16]	
Eger et al. [20] – CTA vs screen-cued RTA	42	100	0.82 [0.18, 1.47]	
Olmsted-Hawala and Bergstrom [56] – young adults	75	62	-0.19 [-0.40, 0.03]	
Olmsted-Hawala and Bergstrom [56] – adults	65	50	-0.25 [-0.52, 0.02]	
Olmsted-Hawala and Bergstrom [56] – old adults	44	41	-0.06 [-0.44, 0.31]	
Overall			0.04 [-0.06, 0.14]	

Figure 3. Task success rate, overall *N* = 520 users

Note: [a] The logarithm of the risk ratio and the 95% confidence interval. [b] The squares in the forest plot show the effect size of each study with the size of the squares indicating the weight of the study in the estimate of the overall effect size. The diamond at the bottom shows the overall effect size. The error bars show the 95%

confidence interval; when it crosses zero (the vertical line), there is no significant difference between CTA and RTA at a *p*-level of .05.

4.1.2 Task time

The task time (i.e., the time the users spent performing the test tasks) was reported in nine cases, see Figure 4. For RTA, task time included the time spent performing the tasks but excluded the time spent retrospectively thinking aloud. For CTA, task time was the time spent performing the tasks while thinking aloud. The task times varied across studies as a result of differences in the number and extent of the tasks. In two studies, task time was significantly longer during CTA than RTA [34,61]. The other studies tended toward a difference in the same direction but did not reach significance. On this basis, the overall effect size across the nine studies was -0.43 with a 95% confidence interval of [-0.64, -0.21]. That is, CTA significantly prolonged tasks compared to performing them without verbalizing the thought process. The difference in task time between CTA and RTA was 43% of the standard deviation.

Study	Task time (min) [a]		Effect size [b]	Forest plot
	CTA	RTA		
Group: classic CTA, classic RTA				
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	20.67 (4.07)	18.90 (3.76)	-0.44 [-1.06, 0.17]	
Alhadreti and Mayhew [2] – CTA vs RTA after first doing CTA	20.67 (4.07)	19.95 (3.50)	-0.19 [-0.79, 0.42]	
Alshammari et al. [4]	7.97 (1.45)	6.90 (2.65)	-0.49 [-1.19, 0.22]	
van den Haak et al. [33]	21.10 (5.70)	19.60 (5.00)	-0.27 [-0.88, 0.34]	
van den Haak et al. [34]	26.10 (6.90)	21.50 (6.50)	-0.67 [-1.30, -0.05]	
van den Haak et al. [35]	25.10 (7.30)	22.20 (6.50)	-0.41 [-1.03, 0.20]	
van den Haak et al. [36]	25.50 (6.70)	24.60 (5.30)	-0.15 [-0.75, 0.46]	
Peute et al. [61]	50.16 (7.62)	41.07 (7.49)	-1.14 [-2.14, -0.14]	
Group: relaxed CTA, relaxed RTA				
Hyrskykari et al. [43]	13.04 (2.51)	10.67 (1.16)	-1.05 [-2.34, 0.23]	
Overall			-0.43 [-0.64, -0.21]	

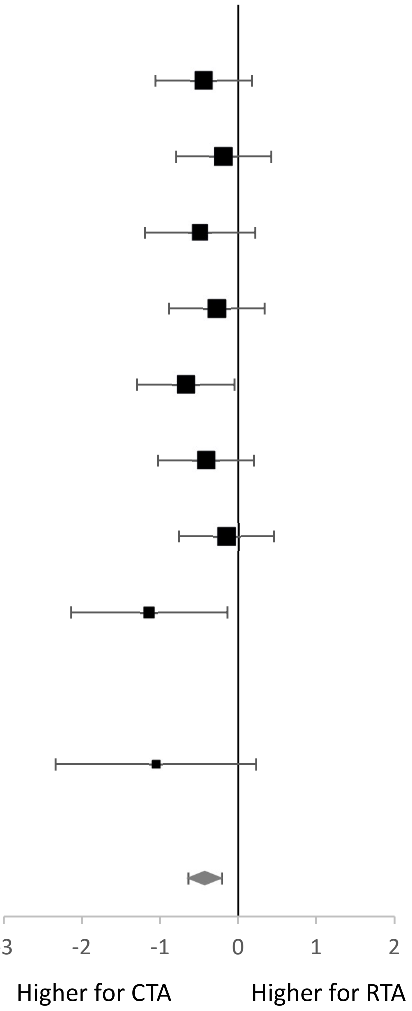


Figure 4. Task time, overall *N* = 270 users

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.1.3 Total time

The total time (i.e., the time the users spent performing the tasks and thinking aloud) was reported in six cases, see Figure 5. For RTA, total time was the time spent performing the tasks plus the time spent retrospectively thinking aloud. For CTA, total time equaled task time. Total time was significantly longer for RTA than CTA in four cases. In these cases, RTA was cued by a video recording of the task performance [13 (Case 1),61] or by a video recording overlaid with the user's eye movements [43]. In the remaining cases, Ji and Rau [45] cued RTA with the chat history of the tested chatbot. They explicitly chose this cue over a video recording to shorten the time spent thinking aloud during RTA. Charoenpruksachat and Longani [13 (Case 2)] provided no cue during RTA; the users were instead prompted with questions. The overall effect size across all six cases was 2.09 with a 95% confidence interval of [-0.48, 4.67]. That is, total time was 2.09 standard deviations longer during RTA than during CTA, but due to the wide confidence interval this large effect was not significant.

Study	Total time (min) [a]		Effect size [b]
	CTA	RTA	
Group: classic CTA, classic RTA			
Ji and Rau [45]	19.34 (3.24)	19.04 (3.58)	-0.09 [-0.59, 0.41]
Peute et al. [61]	50.16 (7.62)	87.12 (6.62)	4.89 [3.04, 6.75]
Group: relaxed CTA, relaxed RTA			
Charoenpruksachat and Longani [13] – CTA vs screen-cued RTA	9.35 (3.12)	11.26 (3.37)	0.58 [0.07, 1.09]
Charoenpruksachat and Longani [13] – CTA vs uncued RTA	9.35 (3.12)	9.28 (2.69)	-0.02 [-0.52, 0.48]
Hyrskykari et al. [43] – CTA vs RTA without first doing CTA	13.04 (2.51)	31.88 (3.43)	5.45 [2.84, 8.06]
Hyrskykari et al. [43] – CTA vs RTA after first doing CTA	13.04 (2.51)	38.67 (10.30)	2.97 [1.22, 4.71]
Overall			2.09 [-0.48, 4.67]

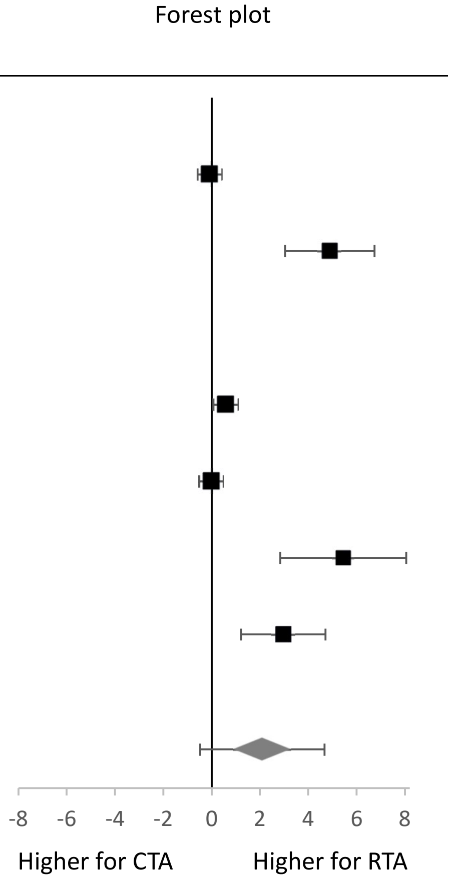


Figure 5. Total time, overall *N* = 174 users

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.2 Usability problems

The output from a usability test is a list of usability problems. For example, the user verbalization "The problem was finding the login page, so I kind of went all over the place looking for the log in page" [52] directly indicated a usability problem. In the reviewed studies, a usability test was considered more effective if it identified more problems, especially more severe problems, than another test. CTA and RTA were compared on several problem-related measures.

4.2.1 Number of problems

The number of usability problems identified per user session was reported in 14 studies, see Figure 6. In three of the seven studies that compared classic CTA with classic RTA, significantly more problems were identified with CTA. For example, Alshammari et al. [4] identified a mean of 16.80 usability problems per CTA user compared to 9.00 per RTA user in their 30-user evaluation of a university website. The seven other cases involved users who did relaxed thinking aloud during CTA and RTA. In all but one of these cases, there was no difference between CTA and RTA in the number of usability problems. As a result, the overall effect size across all 14 studies was -0.07 with a 95% confidence interval of [-0.52, 0.37]. That is, the use of either CTA or RTA did not cause a significant difference in the number of usability problems identified.

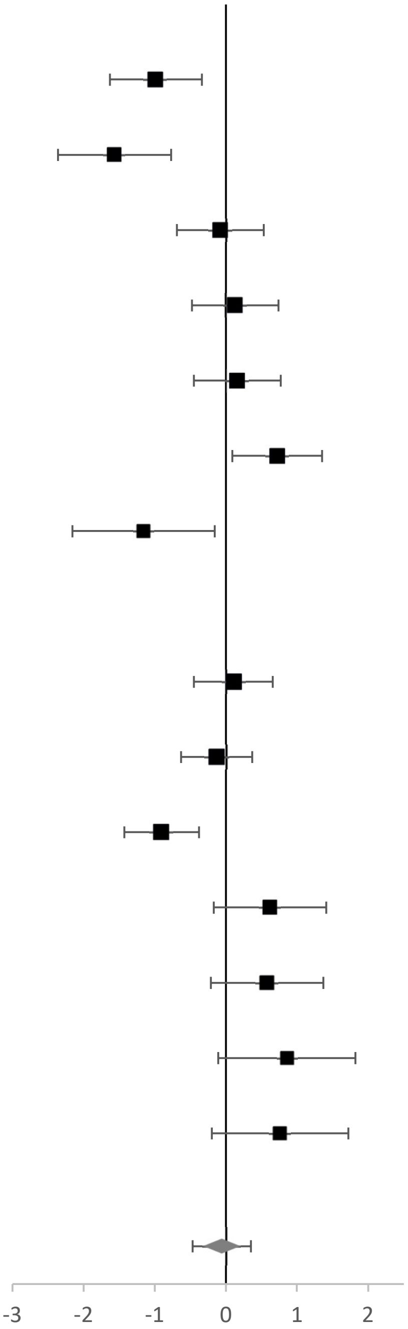
Study	Usability problems [a]		Effect size [b]	Forest plot
	CTA	RTA		
Group: classic CTA, classic RTA				
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	9.55 (3.26)	6.35 (3.09)	-0.99 [-1.63, -0.34]	
Alshammari et al. [4]	16.80 (6.00)	9.00 (3.30)	-1.57 [-2.36, -0.77]	
van den Haak et al. [33]	13.90 (3.30)	13.60 (4.10)	-0.08 [-0.69, 0.53]	
van den Haak et al. [34]	9.70 (2.00)	10.00 (2.50)	0.13 [-0.48, 0.74]	
van den Haak et al. [35]	7.70 (3.70)	8.30 (3.80)	0.16 [-0.45, 0.77]	
van den Haak et al. [36]	13.10 (3.00)	16.00 (4.70)	0.72 [0.09, 1.35]	
Peute et al. [61]	26.23 (3.01)	22.79 (2.58)	-1.16 [-2.16, -0.16]	
Group: relaxed CTA, relaxed RTA				
Capra [11]	3.80 (2.00)	4.00 (1.70)	0.11 [-0.45, 0.66]	
Charoenpruksachat and Longani [13] – CTA vs screen-cued RTA	6.30 (2.62)	6.00 (1.94)	-0.13 [-0.63, 0.37]	
Charoenpruksachat and Longani [13] – CTA vs uncued RTA	6.30 (2.62)	4.33 (1.53)	-0.91 [-1.43, -0.38]	
Eger et al. [20] – CTA vs gaze-cued RTA	8.70 (4.90)	12.50 (6.80)	0.62 [-0.17, 1.41]	
Eger et al. [20] – CTA vs screen-cued RTA	8.80 (4.77)	11.30 (3.39)	0.58 [-0.21, 1.37]	
Savva et al. [65]	6.56 (2.39)	9.69 (4.27)	0.86 [-0.11, 1.82]	
Savva et al. [66]	5.94 (2.02)	8.50 (4.00)	0.76 [-0.20, 1.72]	
Overall			-0.07 [-0.52, 0.37]	

Figure 6. Number of usability problems identified, overall *N* = 416 users

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.2.2 Detection rate

While a similar number of usability problems were identified with CTA and RTA, it could still be the case that CTA and RTA led to the identification of different problems. To investigate this possibility, 16 studies related the problems identified with CTA, or RTA, to the full set of problems identified with all the evaluation methods employed in the study. Figure 7 shows the resulting detection rate (i.e., the percentage of problems identified out of the full set of different problems). The detection rates were in the 47-85% (CTA) and 25-85% (RTA) range. That is, they varied substantially across studies and both CTA and RTA led to the identification of only a subset of the problems. Five studies found a significantly higher detection rate for CTA; three studies found a significantly higher detection rate for RTA (Figure 7). Within this mixed picture, the studies comparing relaxed CTA with relaxed RTA tended toward a larger difference in detection rate than those comparing classic CTA with classic RTA. The overall effect size across the 16 studies was small and not significant (Figure 7).

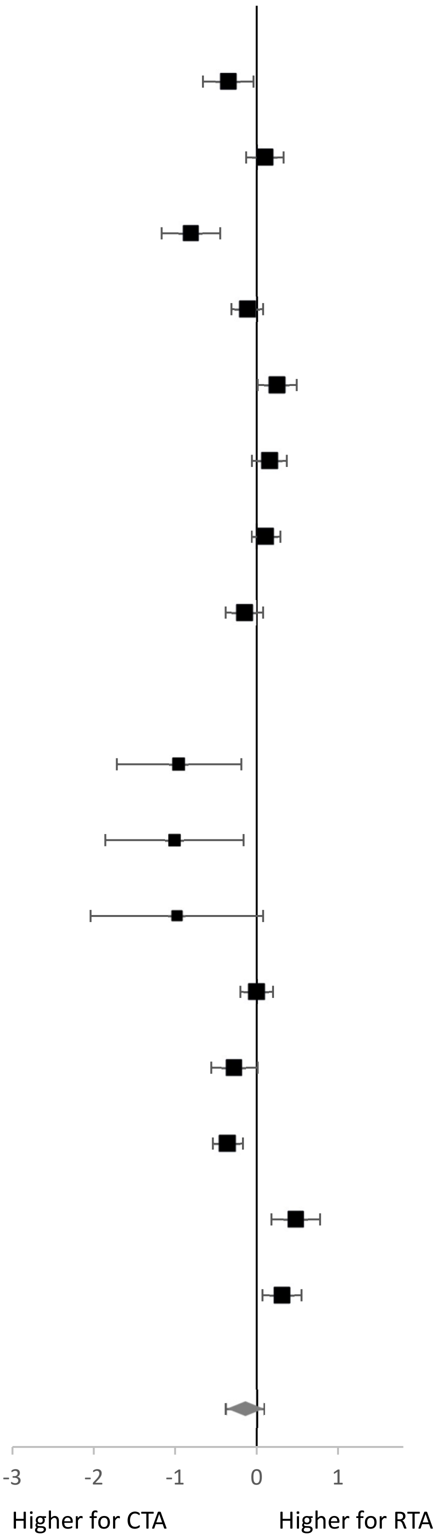
Study	Detection rate (%)		Effect size [a]	Forest plot
	CTA	RTA		
Group: classic CTA, classic RTA				
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	63	44	-0.35 [-0.66, -0.04]	
Alhadreti and Mayhew [2] – CTA vs RTA after first doing CTA	63	69	0.10 [-0.13, 0.33]	
Alshammari et al. [4]	85	38	-0.81 [-1.17, -0.45]	
van den Haak et al. [33]	78	69	-0.11 [-0.31, 0.08]	
van den Haak et al. [34]	54	69	0.25 [0.01, 0.49]	
van den Haak et al. [35]	54	63	0.16 [-0.06, 0.37]	
van den Haak et al. [36]	55	62	0.11 [-0.06, 0.29]	
Peute et al. [61]	84	72	-0.15 [-0.38, 0.08]	
Group: relaxed CTA, relaxed RTA				
AlRoobaea et al. [3] – BBC website	81	31	-0.96 [-1.72, -0.19]	
AlRoobaea et al. [3] – Skool website	85	31	-1.01 [-1.86, -0.16]	
AlRoobaea et al. [3] – Academic Earth website	67	25	-0.98 [-2.04, 0.08]	
Charoenpruksachat and Longani [13] – CTA vs screen-cued RTA	85	85	0.00 [-0.20, 0.20]	
Charoenpruksachat and Longani [13] – CTA vs uncued RTA	85	65	-0.28 [-0.56, 0.01]	
Jensen [44]	83	58	-0.36 [-0.54, -0.17]	
Savva et al. [65] – sigthed users	47	76	0.48 [0.18, 0.78]	
Savva et al. [65] – blind users	55	76	0.31 [0.07, 0.55]	
Overall			-0.14 [-0.38, 0.09]	

Figure 7. Detection rate, overall *N* = 507 users

Note: [a] The logarithm of the risk ratio and the 95% confidence interval.

4.2.3 Problem severity

In five studies [2,3,11,61,65], the authors rated the severity of the identified problems and reported numbers comparing CTA and RTA on how many problems they identified at each severity level. Three additional studies [34–36] provided ratings of the relevance of fixing the identified problems, a notion closely related to problem severity. The ratings in these eight studies were, however, based on dissimilar severity classifications and reported quite differently, thereby precluding meta-analysis. That said, none of the studies found significant differences between CTA and RTA for high-severity problems. Two of the studies indicated that more low-severity problems were identified with CTA than RTA [2,3]. For example, Alhadreti and Mayhew [2] identified a mean of 4.40 (*SD* = 3.74) minor usability problems per CTA user compared to 1.80 (*SD* = 1.63) per RTA user in their evaluation of a library website; this difference was statistically significant.

4.2.4 Problem types

In addition to problem severity, several studies investigated whether CTA and RTA differed in their sensitivity to different types of usability problems. Four problem types recurred in three or more studies: *terminology* (i.e., problems related to the terms and formulations in the user interface), *layout* (i.e., problems related to the structure and graphic design of the individual page), *navigation* (i.e., problems related to maintaining an overview while moving across pages), and *data entry* (i.e., problems related to specifying input for the system to record or process). Meta-analyses of these four problem types showed no significant overall difference between CTA and RTA in the number of problems identified for any of the four types, see Table 1. This result reiterated that the individual studies rarely found such differences. Only 3 of the 22 individual comparisons in Table 1 revealed a significant difference between CTA and RTA. For terminology problems, the meta-analysis showed that significantly more problems were found with CTA than RTA in the study by van den Haak et al. [34]. It may, however, be noted that van den Haak et al. [34] themselves reported this difference as non-significant (possibly because their Bonferroni-adjusted *F*-test was more conservative than the present meta-analysis). For layout problems, Alhadreti and Mayhew [2] found significantly more problems with CTA than RTA, whereas Eger et al. [20 (CTA vs screen-cued RTA)] found significantly fewer problems of this type with CTA than RTA. The other 19 comparisons showed no type-specific difference in the number of problems identified with CTA and RTA.

Table 1. Number of usability problems divided onto those concerning terminology, layout, navigation, and data entry

	Terminology problems [a]		Layout problems [a]		Navigation problems [a]		Data-entry problems [a]	
	CTA	RTA	CTA	RTA	CTA	RTA	CTA	RTA
Group: classic CTA, classic RTA								
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA			3.10 (2.22)	1.00 (0.85)	4.55 (3.42)	3.85 (3.34)		
van den Haak et al. [33]	4.10 (1.50)	4.10 (2.00)	2.90 (1.20)	2.60 (1.30)			4.90 (1.20)	4.90 (1.20)
van den Haak et al. [34]	3.80 (1.30)	2.80 (1.50)	1.10 (0.90)	1.30 (1.10)			4.80 (1.20)	5.00 (1.50)
van den Haak et al. [35]	0.00 (0.00)	0.10 (0.20)	0.90 (0.80)	0.70 (0.70)			0.00 (0.00)	0.00 (0.00)
van den Haak et al. [36]	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.20 (0.60)			0.00 (0.00)	0.00 (0.00)
Group: relaxed CTA, relaxed RTA								

Eger et al. [20] – CTA vs gaze-cued RTA	0.50 (0.60)	0.50 (0.80)	2.80 (2.80)	3.20 (2.00)	0.10 (0.10)	0.10 (0.20)	1.40 (2.30)	0.40 (0.60)
Eger et al. [20] – CTA vs screen-cued RTA	0.50 (0.80)	0.70 (0.80)	1.80 (1.30)	4.20 (2.60)	0.20 (0.36)	0.00 (0.00)	0.50 (0.70)	0.80 (1.10)
Overall effect size [b]	-0.15 [-0.81, 0.51]		-0.06 [-0.85, 0.73]		-0.13 [-1.38, 1.13]		-0.01 [-0.56, 0.55]	
Overall *N* (users)	104		184		52		104	

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.2.5 Source of information about usability problems

To investigate the contribution of thinking aloud to the test results, five studies analyzed whether the usability problems were identified by listening in on the users' verbalizations, observing their behavior, or both. All five studies found that with both CTA and RTA the users' verbalizations led to the identification of usability problems that were not identified by observation only, see Table 2. In addition, the users' verbalizations during both CTA and RTA helped emphasize or explain usability problems that were also observed. That is, thinking aloud contributed to the usability tests by leading to the identification and improved understanding of usability problems. In terms of differences between CTA and RTA, Alhadreti and Mayhew [2] found that verbalization led to the identification of significantly more problems during CTA than RTA. Conversely, van den Haak et al. [33] and van den Haak et al. [34] found that verbalization led to the identification of significantly fewer problems during CTA than RTA and that observation led to the identification of significantly more problems during CTA than RTA. To explain their findings, van den Haak et al. [33,34] proposed that the RTA users had more time to verbalize because they did not have to perform the test tasks concurrently and that the CTA users had more trouble with the test tasks because their workload was increased by the requirement to think aloud during the tasks. However, meta-analyses of the five studies showed moderate but non-significant overall effects for the three sources of evidence about usability problems (Table 2).

Table 2. Number of usability problems divided onto those identified through observation, verbalization, and both, overall *N* = 200 users

Study	Observed usability problems [a]		Verbalized usability problems [a]		Observed and verbalized usability problems [a]	
	CTA	RTA	CTA	RTA	CTA	RTA
Group: classic CTA, classic RTA						
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	1.35 (0.74)	1.30 (0.47)	2.65 (1.75)	1.00 (1.25)	5.55 (1.63)	4.05 (1.98)
van den Haak et al. [33]	6.70 (2.20)	4.00 (2.00)	0.50 (0.70)	4.50 (3.40)	6.70 (4.00)	5.10 (2.20)
van den Haak et al. [34]	5.50 (2.50)	3.10 (1.70)	1.70 (2.10)	3.40 (2.30)	2.50 (1.60)	3.40 (1.60)
van den Haak et al. [35]	1.80 (1.90)	2.80 (2.80)	2.00 (2.10)	3.40 (2.50)	3.90 (3.00)	2.10 (1.50)
van den Haak et al. [36]	6.30 (3.70)	7.20 (4.30)	1.90 (1.80)	2.50 (1.90)	4.90 (3.40)	6.60 (4.60)
Overall effect size [b]	-0.35 [-1.30, 0.60]		0.44 [-0.76, 1.63]		-0.21 [-1.02, 0.59]	

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.3 User verbalizations

The verbalizations produced by the users while thinking aloud are the primary data obtained with CTA and RTA. Differences in the number and content of the users' verbalizations were investigated in several of the reviewed studies.

4.3.1 Number of verbalizations

There were nine comparisons of the total number of verbalizations made by users during CTA and RTA. For seven of the comparisons, the numbers necessary for a meta-analysis were not reported. Instead, Table 3 summarizes the individual studies. They reported mixed results. In four comparisons, CTA users made significantly more verbalizations than RTA users [9,52,55 (both cases)]. In two comparisons, CTA users made significantly fewer verbalizations than RTA users [43 (both cases)]. In one comparison there was no significant difference between CTA and RTA in the number of user verbalizations [45]. And in the two last comparisons, the reported means were higher for RTA than CTA but no statistical tests were reported [72 (both cases)]. Across the nine comparisons, the number of verbalizations varied from five times more during CTA [52] to four time more during RTA [43].

Table 3. Total number of verbalizations made by users

Study	Verbalizations [a]		Statistical test
	CTA	RTA	
Group: classic CTA, classic RTA			
Bowers and Snyder [9]	-	-	Significantly more with CTA
Ji and Rau [45]	22.13 (12.07)	19.27 (11.63)	No significant difference
McDonald et al. [52]	156.40 (39.77)	31.20 (15.17)	Significantly more with CTA
Group: relaxed CTA, relaxed RTA			
Hyrskykari et al. [43] – CTA vs RTA without first doing CTA	66	267	Significantly more with RTA
Hyrskykari et al. [43] – CTA vs RTA after first doing CTA	66	214	Significantly more with RTA
Ohnemus and Biers [55] – CTA vs end-of-session RTA	45.52	-	Significantly more with CTA
Ohnemus and Biers [55] – CTA vs delayed RTA	45.52	-	Significantly more with CTA
Yang et al. [72] – CTA vs RTA	127	370	No statistics reported
Yang et al. [72] – CTA vs RTA after first doing CTA	127	311	No statistics reported

[a] Mean and, in parentheses, standard deviation

4.3.2 Content of verbalizations

To analyze the users' verbalizations further, several studies investigated whether the content of the verbalizations differed between CTA and RTA. Different content classifications were used but three content categories recurred in three or more studies: *procedural description* (i.e., verbalizations in which the users stated what they were doing), *explanation and problem formulation* (i.e., verbalizations in which users expressed why they did – or did not do – something or how the system caused them difficulty), and *design recommendation* (i.e., verbalizations in which the users made suggestions for improving the system). Four studies provided data about these categories and were included in the meta-analyses, see Figures 8-10. In addition, Bowers and Snyder [9] reported the results of statistical tests for the same categories but did not provide the data necessary for including their study in the meta-analyses. All five of these studies compared classic CTA with classic RTA. To make the analysis of the content of the verbalizations independent of differences in the total number of verbalizations, the following analyses were made on the percentage of verbalizations in each category.

Figure 8 shows the results for procedural verbalizations (e.g., "I write the name into this field"). CTA users made a significantly higher percentage of procedural verbalizations in two of the studies [52,59] and in the study by Bowers and Snyder [9]. In these studies, the percentage of procedural verbalizations was 3-5 times higher for CTA than RTA, thereby indicating that the users more consistently stated what they were doing when they thought aloud while they were doing it. This finding was, however, not confirmed by the last study in the meta-analysis. In that study, Fan et al. [27] did not find a significant difference in procedural verbalizations between CTA and RTA. On the basis of these data, the overall effect across the three studies in the meta-analysis was that the percentage of procedural verbalizations was 2.29 standard deviations higher during CTA than RTA, but due to the wide confidence interval this large effect was not significant (Figure 8).

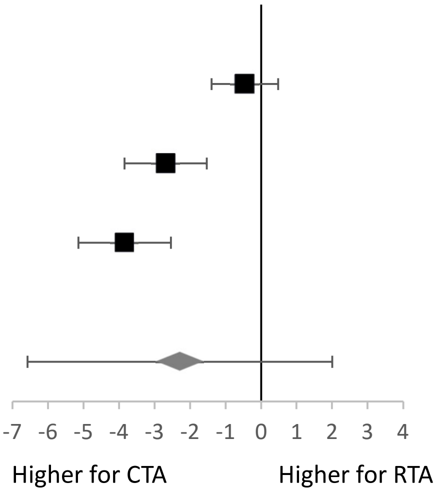
Study	Verbalizations (%) [a]		Effect size [b]	Forest plot
	CTA	RTA		
Group: classic CTA, classic RTA				
Fan et al. [27]	24.50 (11.20)	19.10 (10.90)	-0.46 [-1.40, 0.48]	
McDonald et al. [52]	53.07 (17.20)	10.58 (12.72)	-2.69 [-3.85, -1.53]	
Page and Rahimi [59]	49.10 (9.60)	16.60 (6.40)	-3.85 [-5.15, -2.54]	
Overall			-2.29 [-6.58, 2.00]	

Figure 8. Percentage of verbalizations in the category procedural description, overall N = 30 users

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

Figure 9 shows the percentage of verbalizations in the category explanation and problem formulation (e.g., "The order in which you wanted to do things seemed to be in a completely different order from what you had on the screen"). This percentage was significantly higher for RTA than CTA in two of the studies in the meta-analysis [52,59] but largely similar in the two other studies [27,45]. In all four studies, this category of verbalizations was rather common (13.75-50.10% of all verbalizations), thereby contradicting that the studies described themselves as restricted to classic thinking aloud, which should not include explanations. Overall, the percentage of verbalizations in this category was 1.27 standard deviations higher during RTA than CTA but this overall effect was not significant (Figure 9). Further support for the direction suggested by the overall effect was provided by Bowers and Snyder [9], who found that RTA users verbalized significantly more explanations than CTA users.

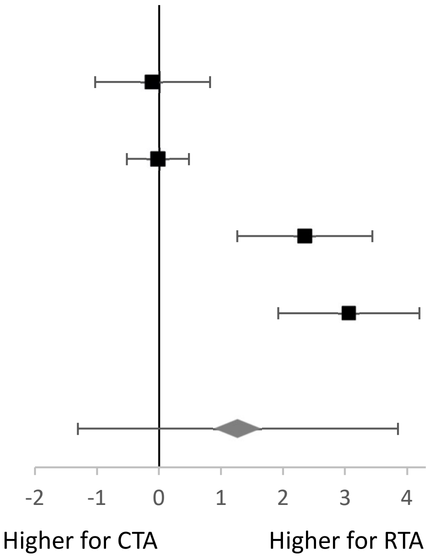
Study	Verbalizations (%) [a]		Effect size [b]	Forest plot
	CTA	RTA		
Group: classic CTA, classic RTA				
Fan et al. [27]	33.60 (16.70)	32.10 (8.90)	-0.11 [-1.03, 0.82]	
Ji and Rau [45]	32.45 (10.60)	32.21 (12.93)	-0.02 [-0.52, 0.48]	
McDonald et al. [52]	13.75 (4.02)	43.91 (16.92)	2.35 [1.26, 3.44]	
Page and Rahimi [59]	29.00 (6.70)	50.10 (6.60)	3.06 [1.92, 4.20]	
Overall			1.27 [-1.31, 3.85]	

Figure 9. Percentage of verbalizations in the category explanation and problem formulation, overall *N* = 90 users

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

Figure 10 shows the results for design recommendations (e.g., "I think it would be easier if they gave you a lot of drop-down lists"). The percentage of verbalizations that included design recommendations was 2.78 standard deviations higher during RTA than CTA, but due to a wide confidence interval this overall effect was not significant (Figure 10). The wide confidence interval was caused by mixed results: One study found no significant difference [27], another study found a significantly higher percentage of design recommendations during CTA [45], and two studies found a significantly higher percentage of design recommendations during RTA [52,59]. In addition, the percentage of verbalizations in this category also differed substantially from one study to another, for example from 0.19% [52] to 21.83% [45] for CTA. The study by Bowers and Snyder [9] supported the direction suggested by the overall effect; they found that RTA users made significantly more design recommendations than CTA users.

Study	Verbalizations (%) [a]		Effect size [b]	Forest plot
	CTA	RTA		
Group: classic CTA, classic RTA				
Fan et al. [27]	3.30 (0.40)	5.00 (2.90)	0.78 [-0.18, 1.74]	
Ji and Rau [45]	21.83 (12.01)	14.53 (12.02)	-0.60 [-1.11, -0.09]	
McDonald et al. [52]	0.19 (0.43)	2.24 (2.63)	1.04 [0.15, 1.94]	
Page and Rahimi [59]	0.40 (0.70)	5.80 (0.04)	10.52 [7.54, 13.49]	
Overall			2.78 [-5.18, 10.75]	

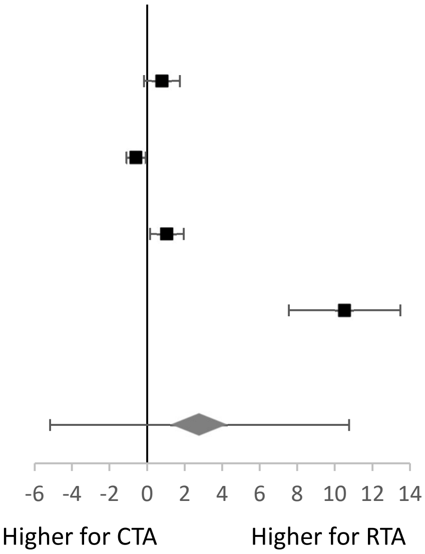


Figure 10. Percentage of verbalizations in the category design recommendation, overall *N* = 90 users

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.4 User experience

The users may experience CTA and RTA differently, irrespective of whether it affects the output from the usability test. The reviewed studies compared CTA and RTA with respect to the users' experience of thinking aloud, their experience of their working procedure, and their experience of the evaluator's presence.

4.4.1 Experience of thinking aloud

The users in several studies rated their experience of thinking aloud. Four scales recurred in three or more studies, see Table 4. For both CTA and RTA, the users tended toward experiencing thinking aloud more positively than negatively, that is, the means in most of the individual studies were on the half of the scale indicating that thinking aloud was easy, natural, not tiring, and pleasant to a larger extent than difficult, unnatural, tiring, and unpleasant. None of the individual studies found a significant difference between CTA and RTA regarding the users' experience of how difficult/easy, unnatural/natural, and tiring/not tiring it was to think aloud. For the unpleasant/pleasant scale, Eger et al. [20] found that RTA users experienced thinking aloud as more pleasant than CTA users. This difference was significant when RTA was gaze-cued as well as when it was screen-cued. The other individual studies found no significant difference in the experience of thinking-aloud pleasantness between CTA and RTA. For all four scales, the overall effect was small or moderate and not significant (Table 4).

In addition to the studies in Table 4, Franz et al. [30] provided qualitative data about the pleasantness of thinking aloud. Their results contradicted those of Eger et al. [20]. For three of the five users, who were frail and elderly, Franz et al. [30] discontinued the RTA sessions because the user found it stressful to watch their own mistakes on the video of their task performance: "I don't want to go through this again. I found it very stressful." In contrast, the four CTA sessions were completed but the users stopped thinking aloud as soon as they got stuck in the tested email client.

Table 4. Users' experience of verbalizing their thoughts, rated on five-point rating scales (higher numbers equal better experience)

	Difficult - Easy [a]		Unnatural - Natural [a]		Tiring – Not tiring [a]		Unpleasant - Pleasant [a]	
	CTA	RTA	CTA	RTA	CTA	RTA	CTA	RTA
Group: classic CTA, classic RTA								
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	3.40 (0.88)	3.65 (1.26)	2.95 (0.94)	3.25 (0.85)	3.50 (1.19)	4.00 (0.85)	3.35 (1.38)	3.60 (1.56)
Alhadreti and Mayhew [2] – CTA vs RTA after first doing CTA	3.40 (0.88)	3.80 (1.32)	2.95 (0.94)	3.10 (1.16)	3.50 (1.19)	3.20 (1.36)	3.35 (1.38)	3.00 (1.37)
Fan et al. [27]	4.50 (0.80)	4.10 (0.60)	4.10 (1.10)	4.10 (0.60)	4.10 (0.80)	4.50 (0.50)	4.40 (0.70)	4.50 (0.50)
van den Haak et al. [33]	2.40 (0.80)	2.70 (1.20)	3.40 (0.90)	3.00 (1.50)	3.40 (1.00)	3.80 (1.40)	2.70 (0.80)	2.90 (1.00)
van den Haak et al. [35]	3.40 (0.80)	3.20 (1.00)						
van den Haak et al. [36]	2.80 (0.80)	2.60 (0.90)						
Group: relaxed CTA, relaxed RTA								
Eger et al. [20] – CTA vs gaze-cued RTA	3.80 (1.10)	4.40 (0.80)	4.30 (0.90)	3.90 (1.20)	4.50 (0.80)	4.80 (0.50)	4.60 (0.60)	5.00 (0.10)
Eger et al. [20] – CTA vs screen-cued RTA	3.80 (1.10)	4.30 (1.20)	4.30 (0.90)	4.30 (1.20)	4.50 (0.80)	4.80 (0.50)	4.60 (0.60)	5.00 (0.01)
Overall effect size [b]	0.12 [-0.17, 0.42]		-0.02 [-0.31, 0.28]		0.29 [-0.04, 0.61]		0.29 [-0.18, 0.77]	
Overall *N* (users)	212		132		132		132	

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.4.2 Experience of working procedure

In several studies, the users rated how their working procedure during the usability test compared to their normal working procedures in terms of speed and focus, see Table 5. The ratings were provided on five-point rating scales with "3" indicating a speed/focus equal to the user's normal procedures. For both CTA and RTA, the users' experience varied across the individual studies from a mean slightly slower and less focused than normal to a mean that was somewhat faster and more focused. With respect to the effect of CTA or RTA, two studies found that RTA users experienced their work speed as significantly higher than CTA users [20 (gaze-cued RTA),35] and three studies found that RTA users experienced their working procedures as significantly more focused than CTA users [20 (both cases),33]. The other studies did not find significant differences between CTA and RTA. As a result, the overall effect was 0.31 (speed) and 0.35 (focus) standard deviations higher for RTA than CTA, but not significant (Table 5).

Table 5. Users' experience of their working procedure, five-point rating scale (higher number equals higher speed/more focused), overall *N* = 244 users

	Slower – higher speed [a]		Less – more focused [a]	
	CTA	RTA	CTA	RTA
Group: classic CTA, classic RTA				
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	3.60 (1.09)	3.85 (1.30)	3.05 (1.14)	2.80 (1.36)
Alhadreti and Mayhew [2] – CTA vs RTA after first doing CTA	3.60 (1.09)	3.35 (1.22)	3.05 (1.14)	3.20 (1.70)
van den Haak et al. [33]	3.30 (0.70)	3.70 (0.80)	3.40 (0.60)	3.90 (0.90)
van den Haak et al. [34]	2.50 (0.50)	2.70 (0.50)	3.50 (0.60)	3.20 (0.50)
van den Haak et al. [35]	2.50 (0.50)	3.00 (0.60)	3.30 (0.70)	3.60 (0.80)
van den Haak et al. [36]	2.60 (0.80)	2.40 (0.60)	3.00 (0.90)	3.60 (1.00)
Group: relaxed CTA, relaxed RTA				
Eger et al. [20] – CTA vs gaze-cued RTA	2.50 (1.00)	3.40 (0.90)	2.50 (0.90)	3.40 (0.80)
Eger et al. [20] – CTA vs screen-cued RTA	2.50 (1.00)	2.70 (0.50)	2.50 (0.90)	3.50 (0.90)
Overall effect size [b]	0.31 [-0.06, 0.68]		0.35 [-0.11, 0.82]	

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

4.4.3 Experience of the evaluator's presence

Finally, the users in five cases rated their experience of the evaluator's presence, see Table 6. For both CTA and RTA, the means in the individual studies were, with the exception of the study by van den Haak et al. [33], on the half of the scale indicating that the evaluator's presence was experienced as natural, not disturbing, and pleasant. The naturalness of the evaluator's presence did not differ significantly between CTA and RTA in any of the individual studies. As a result, the overall effect was small and not significant. In contrast, the evaluator's presence was significantly less disturbing during CTA than RTA in all but Alhadreti and Mayhew's [2] comparison of CTA with RTA after first doing CTA. The overall effect was, however, not significant (Table 6). Because the standard deviation in several of the individual studies was very small, the overall effect size (calculated by dividing with the pooled standard deviation) should be treated as unstable [5]. It was very large, possibly artificially large. The evaluator's presence was experienced as significantly more pleasant by the RTA users than the CTA users in both comparisons by Eger et al. [20]. The three other comparisons found no significant difference. The overall effect was small and not significant.

Table 6. Users' experience of the evaluator's presence, rated on five-point rating scales (higher numbers equal better experience), overall *N* = 124 users

	Unnatural - Natural [a]		Disturbing – Not disturbing [a]		Unpleasant - Pleasant [a]	
	CTA	RTA	CTA	RTA	CTA	RTA
Group: classic CTA, classic RTA						
Alhadreti and Mayhew [2] – CTA vs RTA without first doing CTA	4.65 (0.81)	4.20 (1.21)	4.80 (0.44)	4.40 (0.50)	4.90 (0.30)	4.70 (0.57)
Alhadreti and Mayhew [2] – CTA vs RTA after first doing CTA	4.65 (0.81)	4.50 (0.88)	4.80 (0.44)	4.60 (0.51)	4.90 (0.30)	4.75 (0.44)
van den Haak et al. [33]	2.90 (0.70)	3.10 (1.30)	4.30 (0.60)	3.70 (0.90)	2.80 (0.30)	2.70 (0.80)
Group: relaxed CTA, relaxed RTA						
Eger et al. [20] – CTA vs gaze-cued RTA	4.80 (1.10)	4.50 (0.70)	5.00 (0.10)	3.30 (0.10)	4.50 (0.80)	5.00 (0.10)
Eger et al. [20] – CTA vs screen-cued RTA	4.80 (1.10)	4.80 (0.40)	5.00 (0.10)	4.10 (0.10)	4.50 (0.80)	5.00 (0.10)
Overall effect size [b]	-0.14 [-0.46, 0.18]		-5.18 [-13.78, 3.42]		0.10 [-0.70, 0.89]	

[a] Mean and, in parentheses, standard deviation, [b] Hedges' *g* and 95% confidence interval

5 Discussion

The use of either CTA or RTA in usability tests has been investigated and debated for over three decades. This meta-analytic review combines the results from the existing comparisons of CTA and RTA into an overall analysis.

5.1 Summary of results

Table 7 summarizes the effect of the variables included in the meta-analyses. To conclude that a variable has no effect, its overall effect must be very small and not significant [63]. Conversely, a variable with a large overall effect size is likely to be of practical importance even if it does not reach the level of significance, especially when the number of studies in the meta-analysis is modest and, thereby, introduces a high risk that insufficient power masks a real effect [63]. Cohen [15] proposed that standardized mean difference effect sizes, such as Hedges' *g*, are small when they are at most 0.20 and large when they are at least 0.80. By analyzing over 300 meta-analyses, Lipsey and Wilson [49] proposed adjusting these thresholds to 0.30 and 0.67. On that basis, we will consider an effect very small when it is at most 0.15 and large when it is at least 0.80. Effects in between these values are considered moderate if they are significant and inconclusive if they are not significant. The rightmost column in Table 7 gives the resulting probable conclusion for each variable. These conclusions are as follows:

- For task performance, the task success rate is similar for CTA and RTA, task time is moderately higher for CTA, and total time is probably higher for RTA.
- For usability problems, the total number of problems is similar for CTA and RTA and so is the detection rate and the number of problems in specific problem categories. It is inconclusive whether the number of observed as opposed to verbalized problems differs between CTA and RTA.

- For user verbalizations, the probable conclusion is that CTA users make more procedural-description verbalizations, whereas RTA users make more explanations, problem formulations, and design recommendations.
- For user experience, CTA and RTA are similar with respect to how easy and natural it is to think aloud and how natural and pleasant the evaluator's presence is, but the evaluator's presence is probably less disturbing during CTA. The remaining user-experience variables are inconclusive.

These results span variants of CTA and RTA, including whether the users do classic or relaxed thinking aloud. All the meta-analyses show which studies involve classic and relaxed thinking aloud, but this distinction merely accounts for a modest part of the cross-study variation.

Table 7. Summary of results

Variable	Overall effect [a]	95% confidence interval	Number of studies	Number of users	Probable conclusion
Task performance					
Task success rate	0.04	[-0.06, 0.14]	16	520	No effect
Task time	-0.43	[-0.64, -0.21]	9	270	Moderate effect
Total time	2.09	[-0.48, 4.67]	6	174	(Large effect)
Usability problems					
Number of usability problems	-0.07	[-0.52, 0.37]	14	416	No effect
Detection rate	-0.14	[-0.38, 0.09]	16	507	No effect
Number of terminology problems	-0.15	[-0.81, 0.51]	6	104	No effect
Number of layout problems	-0.06	[-0.85, 0.73]	7	184	No effect
Number of navigation problems	-0.13	[-1.38, 1.13]	3	52	No effect
Number of data-entry problems	-0.01	[-0.56, 0.55]	6	104	No effect
Number of observed problems	-0.35	[-1.30, 0.60]	5	200	Inconclusive
Number of verbalized problems	0.44	[-0.76, 1.63]	5	200	Inconclusive
Number of observed-and-verbalized problems	-0.21	[-1.02, 0.59]	5	200	Inconclusive
User verbalizations					
Percentage of procedural-description verbalizations	-2.29	[-6.58, 2.00]	3	30	(Large effect)
Percentage of explanation-and-problem-formulation verbalizations	1.27	[-1.31, 3.85]	4	90	(Large effect)
Percentage of design-recommendation verbalizations	2.78	[-5.18, 10.75]	4	90	(Large effect)
User experience					
Difficult/easy to think aloud	0.12	[-0.17, 0.42]	8	212	No effect
Unnatural/natural to think aloud	-0.02	[-0.31, 0.28]	6	132	No effect
Tiring/not tiring to think aloud	0.29	[-0.04, 0.61]	6	132	Inconclusive
Unpleasant/pleasant to think aloud	0.29	[-0.18, 0.77]	6	132	Inconclusive
Slower/higher speed	0.31	[-0.06, 0.68]	8	244	Inconclusive
Less/more focused	0.35	[-0.11, 0.82]	8	244	Inconclusive
Evaluator presence is unnatural/natural	-0.14	[-0.46, 0.18]	5	124	No effect

Evaluator presence is disturbing/not disturbing	-5.18	[-13.78, 3.42]	5	124	(Large effect)
Evaluator presence is unpleasant/pleasant	0.10	[-0.70, 0.89]	5	124	No effect

[a] When the effect is negative, the variable (e.g., task success rate) is higher for CTA; when the effect is positive, the variable is higher for RTA.

5.2 Thinking aloud in usability tests

The effectiveness of a usability test is about the extent to which it attains the maxims of validity, robustness, completeness, cost, and impact [37]. Usability tests employ CTA or RTA to help attain these five maxims. In the following, the review results are discussed in relation to each maxim.

The *validity* of a usability test is about whether the problems that surface during the test also hamper real use and whether the problems that hamper real use also surface during the test. It is widely held that classic thinking aloud does not alter behavior, except by prolonging it, whereas relaxed thinking aloud poses a threat to validity because it may alter behavior [24,29]. This contention is the main motivation for RTA, which lets the users solve the test tasks without the interference of thinking aloud and is less dependent on restricting the users to classic thinking aloud when they retrospectively verbalize their thoughts. However, the meta-analysis shows that it has no effect on the task success rate whether the usability test employs CTA or RTA. That is, thinking aloud while solving the test tasks does not alter behavior to the extent of producing different task success rates than those for RTA. This finding is based on studies that span both classic and relaxed thinking aloud (Figure 3), thereby partly moderating the contention that relaxed thinking aloud alters behavior. As expected, Table 7 shows an increase in task time for CTA. In addition, CTA and RTA probably lead to differences in the content of the users' verbalizations and their experience of how much they are disturbed by the evaluator's presence. The users produce more explanations, problem formulations, and design recommendations during RTA than CTA. While these verbalization categories appear important to the identification of usability problems, it should be noted that they do not lead to the identification of more usability problems with RTA. The finding that CTA users are disturbed less by the evaluator's presence probably indicates that the evaluator remains in the background during these sessions to let the users interact with the tested system. In contrast, the evaluator may feel free to assume a more active role during the thinking-aloud part of RTA sessions because the users are no longer interacting with the system.

The *robustness* of a usability test is its ability to produce stable results across variations in the test situation. It is well-documented that test results are sensitive to variation in, for example, evaluators [42], tasks [17], and users [8]. Several of the reviewed studies compare CTA and RTA across variation in users or across variants of RTA. Regarding variation in users, Savva et al. [65,66] compared CTA and RTA across blind and sighted users. Neither of these two studies found significant interaction effects between thinking-aloud condition and user group on the number of usability problems identified. That is, the results for CTA and RTA were robust across the variation in user group. Olmsted-Hawala and Bergstrom [56] compared CTA and RTA across young (18-28 years), middle-aged (40-50 years), and older (64-76 years) adults. They found significant age-group differences in task success rate, task time, and some user-experience ratings but no significant differences between CTA and RTA. While they did not test for interaction effects between thinking-aloud condition and age group, Table 2 in their paper suggests that any interaction effects were in task success rate. That is, the results for CTA and RTA were robust across the variation in user group, with the possible exception that task success rate might be increasingly influenced by thinking-aloud condition with decreasing user age. Regarding variants of RTA, the reviewed studies compare the standard variant (screen-cued RTA at the end of the session without first having done CTA) with uncued RTA [13], gaze-cued RTA [20], delayed RTA [55], and RTA after first doing CTA [2,43,72]. The alternative variants are investigated in too few studies to enable firm conclusions about whether usability test results are robust across the variants of RTA. However, three potential inferences are worth mentioning. First, uncued RTA appears to identify fewer usability problems than cued RTA [13]. Second, the different variants of cued RTA appear to result in roughly the same number of user verbalizations with a similar distribution across content categories [43,55,72]. Third, the number and types of usability problems identified with RTA after first doing CTA appear more similar to those identified with CTA than to those identified with RTA without first having done CTA [2].

The *completeness* of a usability test is about whether it reveals the full set of usability problems or only part of it. In the absence of a definitive method for determining the full set of problems, it is commonly defined as the combined list of problems identified with the different usability tests in a study. With detection rates in the 47-85% (CTA) and 25-85% (RTA) range, the reviewed studies clearly show that neither CTA nor RTA leads to the identification of the full set of usability problems. These detection rates resemble those in other studies of usability tests based on thinking aloud as well as those in studies of other usability evaluation methods [e.g., 8,53]. In addition, the meta-analysis leads to the conclusion that it has no effect on the detection rate whether a usability test employs CTA or RTA. That is, the completeness of usability tests depends largely on factors other than the choice of either CTA or RTA. It provides further evidence in support of this conclusion that the severity of problems and the number of problems in specific problem categories are also similar for CTA and RTA.

The *cost* of a usability test is the base cost of equipment and evaluator competences and the variable cost of compensating users and running test sessions. The reviewed studies mainly address the variable costs, which depend on the number of users and the length of the sessions. Regarding the number of users, the meta-analysis shows no difference in the number of usability problems identified with CTA and RTA, irrespective of whether the users do classic or relaxed thinking aloud. Regarding the length of the test session, the total session time for RTA is about two standard deviations longer than for CTA (Figure 5). There are two reasons why the total time for RTA is not twice that for CTA. First, task time is longer for CTA than RTA because thinking aloud prolongs task completion. Therefore, the thinking-aloud part of an RTA session re-views a shorter task-completion process, even when cued by a video. Second, several of the reviewed studies use cues other than a video in an effort to shorten the thinking-aloud part of RTA sessions. For example, Ji and Rau [45] cued RTA with the chat history of the tested chatbot. With cues other than video, the time required for the thinking-aloud part of RTA is not tied directly to the length of the task-completion process but instead to the extent of the user's verbalizations. In addition to total session time, Charoenpruksachat and Longani [13] also investigated the time needed by the evaluator to analyze the test sessions and found that video-cued RTA sessions took 24% less time to analyze than CTA sessions, while uncued RTA sessions took 56% less time to analyze than CTA sessions. They attribute the former time saving to more audible verbalizations, which could be analyzed without replaying the video multiple times, and the latter time saving to more directed verbalizations, which required less analysis because they were more self-contained. A shorter analysis process for RTA may, to some extent, compensate for the longer session time. The base costs appear to be similar for CTA and RTA, with the possible exception that extra recording equipment may be needed for RTA to cue the users' thinking aloud. A finding common to CTA and RTA is that some of the studies self-describe as employing classic thinking aloud but report a fairly high percentage of verbalizations in the category of explanations and problem formulations (Figure 9). Such verbalizations are at Level 3 (see Section 2) and formally specific to relaxed thinking aloud, yet they occur, in practice, during both classic and relaxed thinking aloud. This finding has also been noted in previous studies [25,38]. It suggests that the evaluators may need better competence in instructing users about how to think aloud.

The *impact* of a usability test concerns whether the identified problems are fixed. It is notable that none of the reviewed studies investigate this issue. Despite considerable debate about the pros and cons of CTA and RTA, the reviewed studies are restricted to the effects of CTA and RTA on the usability test. They do not investigate downstream effects, such as the extent to which the tests have the persuasive power necessary to bring about changes in the tested system. Previous studies find that early usability tests tend to have a higher impact than later tests [37] and that usability inspections without users may struggle with persuasiveness because their results are perceived as opinion [71]. Relatedly, the different qualities of CTA and RTA could influence how test results are received by those who decide which problems to fix.

5.3 Implications

This review has multiple implications. The following list starts with implications for practice and then proceeds with implications for research.

First, practitioners who employ classic thinking aloud should choose CTA to identify about the same number of usability problems within a shorter total time, but they may also consider RTA to shield the performance of the test tasks from reactivity introduced by thinking aloud. Such reactivity should, however, be a minor issue provided that the evaluator instructs the users adequately in classic thinking aloud and the users comply with these instructions [24,29].

Second, practitioners who prefer relaxed thinking aloud should choose CTA if their main concern is total time, and RTA if their main concern is the content of the user verbalizations. The choice of CTA or RTA has little effect on the number and types of problems identified, also when the users do relaxed thinking aloud. That is, the richer verbalizations obtained with relaxed RTA do not appear to make usability tests more effective at identifying usability problems than CTA.

Third, factors other than the choice of either CTA or RTA are more important to the output of usability tests. These factors include the number, diversity, and representativeness of the test users [12]. Additional factors that influence the test results are the tasks solved by the users [17] and the number of evaluators who analyze the test sessions [42]. In complex domains, the quality of the test results also hinges on having domain experts on the usability team [14].

Fourth, RTA allows for relaxed thinking aloud without the risk of influencing the users' task performance. However, the thinking-aloud part of RTA may produce inaccurate verbalizations that mislead the evaluator and result in the reporting of erroneous usability problems. Thus, it appears risky to conduct RTA without cuing the users during the thinking-aloud part of the session, especially if the users are instead prompted with questions in a format that borders on an interview [e.g., 13].

Fifth, CTA taxes users with the added activity of thinking aloud during task performance but does not require that they remember what they are doing to be able subsequently to verbalize it. Only one of the reviewed studies investigates the net effect on the users' workload. This study [65] finds that users experience significantly higher workload during RTA than CTA. Thereby, it reinforces a comment from an RTA user: "While I was searching I had in the back of my mind that I needed to remember what I was doing" [70]. Further research is needed on the workload experienced during CTA and RTA.

Sixth, the reviewed studies make no use of technology to support CTA sessions, while RTA sessions involve screen or gaze recordings to cue the users' verbalizations. Future research should investigate how technology-enhanced CTA sessions compare with RTA. Possible enhancements of CTA include gaze tracking to support the evaluator in analyzing the user's focus of attention [68] and automated sentiment analysis to support the evaluator in assessing the user experience [67].

Seventh, one variant of RTA dominates: screen-cued RTA at the end of the session. Future research should investigate ways of strengthening the thinking-aloud part of such sessions by exploiting its separation from task performance. It may, for example, be possible to have AI pre-analyze the user's behavior and point out the video segments that warrant special attention during thinking aloud [28]. It may even be possible to shorten sessions by skipping over tasks, or subtasks, that did not cause the user any problems.

Eighth, usability tests should also exploit that RTA makes it possible to have users think aloud in settings where thinking aloud is inconvenient or impossible during task performance. Such tests are outside the scope of this review because they do not allow for comparing RTA with CTA. In a usability test that employs RTA, the users may think aloud in the lab after performing tasks in the field, in collaboration with others, in safety-critical domains, in high-workload settings, or in other situations that preclude CTA [e.g., 48,50].

Ninth, the results of the reviewed studies differ to the extent of being inconclusive for several of the variables included in this meta-analysis (Table 7). These variables should be investigated further in future research. It is, for example, important to establish whether users find it more unpleasant to think aloud during RTA than CTA, possibly because it is stressful for them to re-experience their mistakes when they watch the video of their task performance [30].

Tenth, the reviewed studies primarily involve able-bodied adults using websites or other simple systems. The few studies that involve children, blind users, and complex systems should be supplemented with additional studies. In addition, all the reviewed studies concern usability tests performed with the user and evaluator co-present in the lab. Future studies should investigate whether the review results extend to remote and unmoderated usability tests, which are increasingly common.

5.4 Limitations

Four limitations should be remembered in interpreting the results of this review. First, the quality of the meta-analysis hinges on the quality of the 29 reviewed studies. To bolster their quality, only peer-reviewed studies were included. It is, however, acknowledged that the classification of the studies into classic and relaxed thinking aloud on the basis of the methodological description in the studies was in some cases contradicted by the fairly high percentage of verbalizations that included explanation and problem formulation (Section 4.3.2). Second, the reviewed studies compare CTA and RTA on variables that differ across the studies. The difference in variables adds breadth and richness to this review but it also means that the number of studies that compare CTA and RTA is modest for most of the variables. As a result, the meta-analysis is inconclusive for seven of the variables (Table 7). Third, the reviewed studies were classified according to the taxonomy in Figure 1 but there may be additional factors that moderate how task performance, usability problems, user verbalizations, and user experience are influenced by CTA and RTA. The influence of such factors remains hidden in this review, but they may explain some of the variation in the data and this variation may partially mask the real effect of some of the analyzed variables. Moderating factors that could be considered if they were investigated in enough studies include user characteristics (e.g., children vs adults vs elderly) and system characteristics (e.g., website vs game vs office application). Fourth, this review is not an examination of the validity of thinking aloud. The review compares CTA and RTA and, in doing so, covers whether they differentially influence task performance. The validity, or reactivity, of thinking aloud is thoroughly investigated in studies restricted to either CTA or RTA. In the context of usability tests, the validity of thinking aloud has, for example, been investigated by Elling et al. [23] for CTA and by Guan et al. [32] for RTA.

6 Conclusion

This review provides a meta-analysis of the studies that compare the use of CTA or RTA in usability tests. The meta-analyses span 24 variables about task performance, usability problems, user verbalizations, and user experience. For practitioners, the main implication of this review depends on whether they employ classic or relaxed thinking aloud. In the former case, they should choose CTA to identify about the same number of usability problems within a shorter total time. In the latter case, they should choose CTA if their main concern is total time, and RTA if their main concern is the content of the users' verbalizations. For researchers, the main implication is the identification of multiple areas for future research. In particular, future research should untangle the inconclusive meta-analyses for seven of the analyzed variables and compare CTA and RTA in a wider variety of settings than usability tests with the user and evaluator co-present in the lab.

Appendix

The 29 studies included in the review are listed below. Each study is classified according to the taxonomy in Figure 1. The studies that contain multiple comparisons between CTA and RTA may have more than one classification in some of the columns. The list also gives the total number of users in each study.

Study	CTA levels	RTA levels	RTA cues	RTA delay	Users
Alhadreti and Mayhew [2]	Classic	Classic	Video	End of session	60
AlRoobaea et al. [3]	Relaxed	Relaxed	Video	End of session	120
Alshammari et al. [4]	Classic	Classic	Video	End of session	30
Balatsoukas et al. [6]	Classic	Relaxed	Video	End of session	35

Bowers and Snyder [9]	Classic	Classic	Video	End of session	48
Capra [11]	Relaxed	Relaxed	Video	End of session	24
Charoenpruksachat and Longani [13]	Relaxed	Relaxed	Video/Uncued	End of session	90
Donker and Markopoulos [18]	Relaxed	Relaxed	Uncued	Post task	30
Eger et al. [20]	Relaxed	Relaxed	Gaze+video/Video	End of session	24
Fan et al. [27]	Classic	Classic	Video	End of session	8
Franz et al. [30]	Relaxed	Relaxed	Video	End of session	8
van den Haak et al. [33]	Classic	Classic	Video	End of session	40
van den Haak et al. [34]	Classic	Classic	Video	End of session	40
van den Haak et al. [35]	Classic	Classic	Video	End of session	40
van den Haak et al. [36]	Classic	Classic	Video	End of session	40
Hyrskykari et al. [43]	Relaxed	Relaxed	Gaze+video	End of session	8
Jensen [44]	Relaxed	Relaxed	Video	End of session	15
Ji and Rau [45]	Classic	Classic	Chat history	End of session	60
van Kesteren et al. [46]	Classic/Relaxed	Relaxed	Video	End of session	6
McDonald et al. [52]	Classic	Classic	Uncued	End of session	10
Ohnemus and Biers [55]	Relaxed	Relaxed	Video	End of session/24h delay	30
Olmsted-Hawala and Bergstrom [56]	Relaxed	Relaxed	Not reported	End of session	95
Page and Rahimi [59]	Classic	Classic	Video	End of session	12
Petrie and Precious [60]	Relaxed	Relaxed	Not reported	End of session	16
Peute et al. [61]	Classic	Classic	Video	End of session	16
Prokop et al. [62]	Classic	Classic	Uncued	Post task	31
Savva et al. [65]	Relaxed	Relaxed	Video/Audio	End of session	16
Savva et al. [66]	Relaxed	Relaxed	Video/Audio	End of session	16
Yang et al. [72]	Relaxed	Relaxed	Gaze+video	End of session	20

References

[1] Obead Alhadreti and Pam Mayhew. 2017. To intervene or not to intervene: An investigation of three think-aloud protocols in usability testing. *J. Usability Stud.* 12, 3 (2017), 111–132. Retrieved from https://uxpajournal.org/intervene-think-aloud-protocols-usability-testing/

[2] Obead Alhadreti and Pam Mayhew. 2018. Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the CHI2018 Conference on Human Factors in Computing Systems*. ACM, New York, paper 44. DOI:https://doi.org/10.1145/3173574.3173618

[3] Roobaea AlRoobaea, Ali H. Al-Badi, and Pam J. Mayhew. 2013. The impact of the combination between task designs and think-aloud approaches on website evaluation. *J. Softw. Syst. Dev.* 2013, (2013), article 172572. DOI:https://doi.org/10.5171/2013. 172572

[4] Thamer Alshammari, Obead Alhadreti, and Pam J. Mayhew. 2015. When to ask participants to think aloud: A comparative study of concurrent and retrospective think-aloud methods. *Int. J. Recent Trends Human-Computer Interact.* 6, 3 (2015), 48–64. Retrieved from https://www.cscjournals.org/library/manuscriptinfo.php?mc=IJHCI-118

[5] Arthur Bakker, Jinfa Cai, Lyn English, Gabriele Kaiser, Vilma Mesa, and Wim Van Dooren. 2019. Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educ. Stud. Math.* 102, 1 (2019), 1–8. DOI:https://doi.org/10.1007/s10649-019-09908-4

[6] Panagiotis Balatsoukas, John Ainsworth, Richard Williams, Emma Carruthers, Colin Davies, James McGrath, Artur Akbarov, Claudia Soiland-Reyes, Saurin Badiyani, and Iain Buchan. 2013. Verbal

protocols for assessing the usability of clinical decision support: The retrospective sense making protocol. In *Proceedings of the MEDINFO2013 Conference*. IOS Press, Amsterdam, 283–287. DOI:https://doi.org/10.3233/978-1-61499-289-9-283

[7] Ted Boren and Judith Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE Trans. Prof. Commun.* 43, 3 (2000), 261–278. DOI:https://doi.org/10.1109/47.867942

[8] Simone Borsci, Robert D. Macredie, Julie Barnett, Jennifer Martin, Jasna Kuljis, and Terry Young. 2013. Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. *ACM Trans. Comput. Interact.* 20, 5 (2013), article 29. DOI:https://doi.org/10.1145/2506210

[9] Victoria A. Bowers and Harry L. Snyder. 1990. Concurrent versus retrospective verbal protocols for comparing window usability. *Proc. Hum. Factors Soc. Annu. Meet.* 34, 17 (1990), 1270–1274. DOI:https://doi.org/10.1177/154193129003401720

[10] Anders Bruun, Effie L.-C. Law, Thomas D. Nielsen, and Matthias Heintz. 2021. Do you feel the same? On the robustness of cued-recall debriefing for user experience evaluation. *ACM Trans. Comput. Interact.* 28, 4 (2021), article 25. DOI:https://doi.org/10.1145/3453479

[11] Miranda G. Capra. 2002. Contemporaneous versus retrospective user-reported critical incidents in usability evaluation. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 46, 24 (2002), 1973–1977. DOI:https://doi.org/10.1177/154193120204602408

[12] David A. Caulton. 2001. Relaxing the homogeneity assumption in usability testing. *Behav. Inf. Technol.* 20, 1 (2001), 1–7. DOI:https://doi.org/10.1080/01449290010020648

[13] Alongkorn Charoenpruksachat and Pattama Longani. 2021. Comparative study of usability evaluation methods on a hyper casual game. In *Proceedings of the International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*. IEEE, New York, 153–156. DOI:https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425717

[14] Parmit K. Chilana, Jacob O. Wobbrock, and Andrew J. Ko. 2010. Understanding usability practices in complex domains. In *Proceedings of the CHI2010 Conference on Human Factors in Computing Systems*. ACM, New York, 2337–2346. DOI:https://doi.org/10.1145/1753326.1753678

[15] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences, 2nd edition*. Erlbaum, Hillsdale, NJ.

[16] Lynne Cooke. 2010. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Trans. Prof. Commun.* 53, 3 (2010), 202–215. DOI:https://doi.org/10.1109/TPC.2010.2052859

[17] Richard E. Cordes. 2001. Task-selection bias: A case for user-defined tasks. *Int. J. Hum. Comput. Interact.* 13, 4 (2001), 411–419. DOI:https://doi.org/10.1207/S15327590IJHC1304_04

[18] Afke Donker and Panos Markopoulos. 2002. A comparison of think-aloud, questionnaires and interviews for testing usability with children. In *People and Computers XVI: Proceedings of HCI2002*. Springer, London, 305–316. DOI:https://doi.org/10.1007/978-1-4471-0105-5_18

[19] Joseph S. Dumas and Janice C. Redish. 1999. *A practical guide to usability testing. Revised edition*. Intellect Books, Exeter, UK.

[20] Nicola Eger, Linden J. Ball, Robert Stevens, and Jon Dodd. 2007. Cueing retrospective verbal reports in usability testing through eye-movement replay. In *People and Computers XXI: Proceedings of HCI2007*. British Computer Society, 129–137. DOI:https://doi.org/10.14236/ewic/HCI2007.13

[21] Fatma Elbabour, Obead Alhadreti, and Pam Mayhew. 2017. Eye tracking in retrospective think-aloud usability testing: Is there added value? *J. Usability Stud.* 12, 3 (2017), 95–110.

[22] Sanne Elling, Leo Lentz, and Menno de Jong. 2011. Retrospective think-aloud method: Using eye

movements as an extra cue for participants' verbalizations. In *Proceedings of the CHI2011 Conference on Human Factors in Computing Systems*. ACM, New York, 1161–1170. DOI:https://doi.org/10.1145/1978942.1979116

[23] Sanne Elling, Leo Lentz, and Menno de Jong. 2012. Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Trans. Prof. Commun.* 55, 3 (2012), 206–220. DOI:https://doi.org/10.1109/TPC.2012.2206190

[24] K. Anders Ericksson and Herbert A. Simon. 1993. *Protocol analysis: Verbal reports as data. Revised edition*. MIT Press, Cambridge, MA.

[25] Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Troung. 2019. Concurrent think-aloud verbalizations and usability problems. *ACM Trans. Comput. Interact.* 26, 5 (2019), article 28. DOI:https://doi.org/10.1145/3325281

[26] Mingming Fan, Serina Shi, and Khai N. Troung. 2020. Practices and challenges of using think-aloud protocols in industry: An international survey. *J. Usability Stud.* 15, 2 (2020), 85–102. Retrieved from https://uxpajournal.org/practices-challenges-think-aloud-protocols-survey/

[27] Mingming Fan, Vinita Tibdewal, Qiwen Zhao, Lizhou Cao, Chao Peng, Runxuan Shu, and Yujia Shan. 2022. Older adults' concurrent and retrospective think-aloud verbalizations for identifying user experience problems of VR games. *Interact. Comput.* 34, 4 (2022), 99–115. DOI:https://doi.org/10.1093/iwc/iwac039

[28] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N. Truong. 2020. VisTA: Integrating machine intelligence with visualization to support the investigation of think-aloud sessions. *IEEE Trans. Vis. Comput. Graph.* 26, 1 (2020), 343–352. DOI:https://doi.org/10.1109/TVCG.2019.2934797

[29] Mark C. Fox, K. Anders Ericsson, and Ryan Best. 2011. Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychol. Bull.* 137, 2 (2011), 316–344. DOI:https://doi.org/10.1037/a0021663

[30] Rachel L. Franz, Barbara B. Neves, Carrie D. Epp, Ronald Baecker, and Jacob O. Wobbrock. 2019. Why and how think-alouds with older adults fail: Recommendations from a study and expert interviews. In *Perspectives on Human-Computer Interaction Research with Older People*. Springer, Cham, 217–235. DOI:https://doi.org/10.1007/978-3-030-06076-3_14

[31] K. J. Gilhooly, E. Fioratou, and N. Henretty. 2010. Verbalization and problem solving: Insight and spatial factors. *Br. J. Psychol.* 101, 1 (2010), 81–93. DOI:https://doi.org/10.1348/000712609X422656

[32] Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. 2006. The validity of the stimulated retrospective think-aloud method as measured by eye-tracking. In *Proceedings og the CHI2006 Conference on Human Factors in Computing Systems*. ACM, New York, 1253–1262. DOI:https://doi.org/10.1145/1124772.1124961

[33] Maaike J. van den Haak, Menno D. T. de Jong, and Peter J. Schellens. 2003. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behav. Inf. Technol.* 22, 5 (2003), 339–351. DOI:https://doi.org/10.1080/0044929031000

[34] Maaike J. van den Haak, Menno D. T. de Jong, and Peter J. Schellens. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interact. Comput.* 16, 6 (2004), 1153–1170. DOI:https://doi.org/10.1016/j.intcom.2004.07.007

[35] Maaike J. van den Haak, Menno D. T. de Jong, and Peter J. Schellens. 2007. Evaluation of an informational web site: Three variants of the think-aloud method compared. *Tech. Commun.* 54, 1 (2007), 58–71.

[36] Maaike J. van den Haak, Menno D. T. de Jong, and Peter J. Schellens. 2009. Evaluating municipal

websites: A methodological comparison of three think-aloud variants. *Gov. Inf. Q.* 26, 1 (2009), 193–202. DOI:https://doi.org/10.1016/j.giq.2007.11.003

[37] Morten Hertzum. 2020. *Usability testing: A practitioner's guide to evaluating the user experience*. Springer, Cham. DOI:https://doi.org/10.2200/S00987ED1V01Y202001HCI045

[38] Morten Hertzum, Pia Borlund, and Kristina B. Kristoffersen. 2015. What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *Int. J. Hum. Comput. Interact.* 31, 9 (2015), 557–570. DOI:https://doi.org/10.1080/10447318.2015.1065691

[39] Morten Hertzum, Kristin D. Hansen, and Hans H. K. Andersen. 2009. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behav. Inf. Technol.* 28, 2 (2009), 165–181. DOI:https://doi.org/10.1080/01449290701773842

[40] Morten Hertzum and Kristin D. Holmegaard. 2015. Thinking aloud influences perceived time. *Hum. Factors* 57, 1 (2015), 101–109. DOI:https://doi.org/10.1177/0018720814540208

[41] Morten Hertzum and Kristina B. Kristoffersen. 2018. What do usability test moderators say? "Mm hm", "uh-huh", and beyond. In *NordiCHI2018: Proceedings of the Tenth Nordic Conference on Human-Computer Interaction*. ACM, New York, 364–375. DOI:https://doi.org/10.1145/3240167.3240181

[42] Morten Hertzum, Rolf Molich, and Niels E. Jacobsen. 2014. What you get is what you see: Revisiting the evaluator effect in usability tests. *Behav. Inf. Technol.* 33, 2 (2014), 144–162. DOI:https://doi.org/10.1080/0144929X.2013.783114

[43] Aulikki Hyrskykari, Saila Ovaska, Pävi Majaranta, Kari-Jouko Räihä, and Merja Lehtinen. 2008. Gaze path stimulation in retrospective think-aloud. *J. Eye Mov. Res.* 2, 4 (2008), 1–18. DOI:https://doi.org/10.16910/jemr.2.4.5

[44] Janne J. Jensen. 2007. Evaluating in a healthcare setting: A comparison between concurrent and retrospective verbalisation. In *Proceedings of HCI International 2007*. Springer, Berlin, 508–516. DOI:https://doi.org/10.1007/978-3-540-73105-4_56

[45] Xiang Ji and Pei-Luen P. Rau. 2019. A comparison of three think-aloud protocols used to evaluate a voice intelligent agent that expresses emotions. *Behav. Inf. Technol.* 38, 4 (2019), 375–383. DOI:https://doi.org/10.1080/0144929X.2018.1535621

[46] Ilse E. H. van Kesteren, Mathilde M. Bekker, Arnold P. O. S. Vermeeren, and Peter A. Lloyd. 2003. Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In *Proceedings of the IDC2003 Conference on Interaction Design and Children*. ACM, New York, 41–49. DOI:https://doi.org/10.1145/953536.953544

[47] Clayton Lewis. 1982. *Using the "thinking-aloud" method in cognitive interface design, RC 9265 (#40713)*. IBM Thomas Watson Research Center, Yorktown Heights, NY.

[48] Cindy H. Lio, C. Melody Carswell, Stephen E. Strup, John S. Roth, and Russell Grant. 2010. The operating room as classroom: Understanding cognitive challenges facing surgical trainees. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 54, 19 (2010), 1571–1575. DOI:https://doi.org/10.1177/154193121005401945

[49] Mark W. Lipsey and David B. Wilson. 2001. *Practical meta-analysis*. Sage, Thousans Oaks, CA.

[50] Cindy Mayas, Stephan Hörold, Christina Rosenmöller, and Heidi Krömker. 2014. Evaluating methods and equipment for usability field tests in public transport. In *Proceedings of HCI International 2014*. Springer, Cham, 545–553. DOI:https://doi.org/10.1007/978-3-319-07233-3_50

[51] Sharon McDonald, Helen M. Edwards, and Tingting Zhao. 2012. Exploring think-alouds in usability testing: An international survey. *IEEE Trans. Prof. Commun.* 55, 1 (2012), 2–19. DOI:https://doi.org/10.1109/TPC.2011.2182569

[52] Sharon McDonald, Tingting Zhao, and Helen M. Edwards. 2013. Dual verbal elicitation: The complementary use of concurrent and retrospective reporting within a usability test. *Int. J. Hum. Comput. Interact.* 29, 10 (2013), 647–660. DOI:https://doi.org/10.1080/10447318.2012.758529

[53] Jakob Nielsen and Thomas K. Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERCHI1993 Conference on Human Factors in Computing Systems*. ACM, New York, 206–213. DOI:https://doi.org/10.1145/169059.169166

[54] Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the DIS2006 Conference on Designing Interactive Systems*. ACM, New York, 209–218. DOI:https://doi.org/10.1145/1142405.1142439

[55] Kenneth R. Ohnemus and Dawid W. Biers. 1993. Retrospective versus concurrent thinking-out-loud in usability testing. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 37, 17 (1993), 1127–1131. DOI:https://doi.org/10.1177/154193129303701701

[56] Erica Olmsted-Hawala and Jennifer R. Bergstrom. 2012. Think-aloud protocols: Does age make a difference? In *Proceedings of the STC Technical Communitation Summit 2012*. Society for Technical Communication, Fairfax, VA, 86–95.

[57] Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the CHI2010 Conference on Human Factors in Computing Systems*. ACM, New York, 2381–2390. DOI:https://doi.org/10.1145/1753326.1753685

[58] Anneli Olsen, Linnea Smolentzov, and Tommy Strandvall. 2010. Comparing different eye tracking cues when using the retrospective think aloud method in usability testing. In *Proceedings of the 24th BCS Interaction Specialist Group Conference*. British Computer Society, 45–53.

[59] Colleen Page and Mansour Rahimi. 1995. Concurrent and retrospective verbal protocols in usability testing: Is there value added in collecting both? *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 39, 4 (1995), 223–227. DOI:https://doi.org/10.1177/154193129503900401

[60] Helen Petrie and John Precious. 2010. Measuring user experience of websites: Think aloud protocols and an emotion word prompt list. In *CHI2010 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, 3673–3678. DOI:https://doi.org/10.1145/1753846.1754037

[61] Linda W. P. Peute, Nicolette F. de Keizer, and Monique W. M. Jaspers. 2015. The value of retrospective and concurrent think aloud in formative usability testing of a physician data query tool. *J. Biomed. Inform.* 55, (2015), 1–10. DOI:https://doi.org/10.1016/j.jbi.2015.02.006

[62] Michal Prokop, Ladislav Pilař, and Ivana Tichá. 2020. Impact of think-aloud on eye-tracking: A comparison of concurrent and retrospective think-aloud for research on decision-making in the game environment. *Sensors* 20, 10 (2020), article 2750. DOI:https://doi.org/10.3390/s20102750

[63] Robert Rosenthal. 1991. *Meta-analytic procedures for social research. Revised edition*. Sage, Newbury Park, CA.

[64] Paul M. Salmon, Natassia Goode, Antje Spiertz, Miles Thomas, Eryn Grant, and Amanda Clacy. 2017. Is it really good to talk? Testing the impact of providing concurrent verbal protocols on driving performance. *Ergonomics* 60, 6 (2017), 770–779. DOI:https://doi.org/10.1080/00140139.2016.1214752

[65] Andreas Savva, Helen Petrie, and Christopher Power. 2015. Comparing concurrent and retrospective verbal protocols for blind and sighted users. In *Proceedings of the INTERACT2015 Conference on Human-Computer Interaction*. Springer, Cham, 55–71. DOI:https://doi.org/10.1007/978-3-319-22701-6_5

[66] Andreas Savva, Helen Petrie, and Christopher Power. 2016. Types of problems elicited by verbal

protocols for blind and sighted participants. In *Proceedings of the ICCHP2016 International Conference on Computers Helping People with Special Needs*. Springer, Cham, 560–567. DOI:https://doi.org/10.1007/978-3-319-41267-2_79

[67] Roberto Y. da Silva Franco, Rodrigo S. do Amor Divino Lima, Rafael do Monte Paixão, Carlos G. R. dos Santos, and Bianchi S. Meiguins. 2019. UXmood—A sentiment analysis and information visualization tool to support the evaluation of usability and user experience. *Information* 10, 12 (2019), article 366. DOI:https://doi.org/10.3390/info10120366

[68] Jiahui Wang, Pavlo Antonenko, Mehmet Celepkolu, Yerika Jimenez, Ethan Fieldman, and Ashley Fieldman. 2019. Exploring relationships between eye tracking and traditional usability testing data. *Int. J. Human–Computer Interact.* 35, 6 (2019), 483–494. DOI:https://doi.org/10.1080/10447318.2018.1464776

[69] Paul Ward, Kyle Wilson, Joel Suss, William D. Woody, and Robert R. Hoffman. 2019. A historical perspective on introspection: Guidelines for eliciting verbal and introspective-type reports. In *The Oxford Handbook of Expertise*. Oxford University Press, Oxford, 377–407. DOI:https://doi.org/10.1093/oxfordhb/9780198795872.013.17

[70] Leanne M. Willis and Sharon McDonald. 2016. Retrospective protocols in usability testing: A comparison of post-session RTA versus post-task RTA reports. *Behav. Inf. Technol.* 35, 8 (2016), 628–643. DOI:https://doi.org/10.1080/0144929X.2016.1175506

[71] Chauncey Wilson. 2014. *User interface inspection methods: A user-centered design method*. Morgan Kaufmann, Waltham, MA.

[72] Zengyao Yang, Yu Zhang, Meng Li, and Tianning Chen. 2018. The comparison study of usability test methodology based on eye-tracking technology. In *Proceedings of the MMESE2017 Conference on Man-Machine-Environment System Engineering*. Springer, Singapore, 763–772. DOI:https://doi.org/10.1007/978-981-10-6232-2_91