Concurrent Think-Aloud Verbalizations and Usability Problems

MINGMING FAN, JINGLAN LIN, CHRISTINA CHUNG, and KHAI N. TRUONG,
University of Toronto

The concurrent think-aloud protocol—in which participants verbalize their thoughts when performing tasks—is a widely employed approach in usability testing. Despite its value, analyzing think-aloud sessions can be onerous because it often entails assessing all of a user's verbalizations. This has motivated previous research on developing *categories* to segment verbalizations into manageable units of analysis. However, the way in which a category might relate to usability problems is currently unclear. In this research, we sought to address this gap in our understanding. We also studied how *speech features* might relate to usability problems. Through two studies, this research demonstrates that certain patterns of verbalizations are more telling of usability problems than others and that these patterns are robust to different types of test products (i.e., physical devices and digital systems), access to different types of information (i.e., video and audio modality), and the presence or absence of a visualization of verbalizations. The implication is that the verbalization and speech patterns can potentially reduce the time and effort required for analysis by enabling evaluators to focus more on the important aspects of a user's verbalizations. The patterns could also potentially be used to inform the design of systems to automatically detect when in the recorded think-aloud sessions users experience problems.

CCS Concepts: • **Human-centered computing** → *Usability testing*;

Additional Key Words and Phrases: Concurrent think-aloud, usability testing, verbalization, verbalization categories, speech features, silence, verbal fillers, sentiment, speech rate, loudness, pitch, usability problems

ACM Reference format:
Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Truong. 2019. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 28 (July 2019), 35 pages.
https://doi.org/10.1145/3325281

1 INTRODUCTION

A product's design can critically impact its user experience. If a device is poorly designed, people may stumble. If a website is difficult to navigate, people may seek alternatives. Thus, it is important that products are iteratively designed and tested prior to their release. The think-aloud protocol is a widely used and highly valued usability testing method that is often used in iterative design to help ensure that products work as intended [30]. While thinking aloud, participants are asked to verbalize their thoughts when working on a task; this enables evaluators to learn about

Authors' addresses: M. Fan, J. Lin, C. Chung, and K. N. Truong, Department of Computer Science, University of Toronto, 40 St. George Street, Toronto, Ontario, Canada M5S 2E4; emails: {mfan, cjlin}@cs.toronto.edu, ricecrispi@dgp.toronto.edu, khai@cs.toronto.edu.

1073-0516/2019/07-ART28 $15.00
https://doi.org/10.1145/3325281

problems encountered by potential users and gain insights that cannot be easily obtained from mere observations [9, 46].

Despite the value of conducting think-aloud sessions, analyzing them is an involved process that entails reviewing sessions, often repeatedly, to pinpoint when users are encountering problems. Because usability evaluators often face time pressure, it is not unusual for their analysis reports to be left incomplete [32]. This has motivated researchers to formulate classification schemes that decompose a user's verbalizations into more manageable segments for later analysis. For example, a systematic study by Cooke concludes that the participants' verbalizations (i.e., utterances) in think-aloud sessions could be characterized by four *categories*: reading (i.e., regurgitating instructions or texts on the user interface), procedure (i.e., describing one's actions), observation (i.e., making remarks), and explanation (i.e., motivating one's behavior) [9], which were further validated by Elling et al. [10]. In this work, we examine if the occurrence of certain categories of verbalizations can indicate when users experience problems during think-aloud sessions.

In addition to what people say (i.e., verbalizations), how people say it (e.g., pitch, speech rate) can also reveal their feelings, mood [36], signs of high cognitive load [15, 45], and levels of confidence (e.g., [15, 25, 36]). Thus, we also examine if the ways in which users speak during think-aloud sessions can be indicators of when they experience problems as well.

In this work, we explore the following overarching research question: does what users say (i.e., *verbalization*) and how they say it (i.e., *speech features*) during think-aloud sessions indicate when they have experienced a usability problem? As part of this research, we seek to identify *verbalizations* and *speech features* that are most indicative of when problems are encountered. The findings of our research could potentially be used to inform the future design of systems to automatically detect when in the recorded think-aloud sessions users experience problems. Furthermore, the findings of this research could also allow usability evaluators to pay special attention to parts of a think-aloud session, which contain the verbalization and speech patterns that tend to occur when users experience problems.

To answer the overarching research question, we designed, conducted, and analyzed two studies. In Study 1, we first conducted and audio recorded think-aloud sessions. We then recruited usability evaluators to identify usability problems from these sessions. The findings of Study 1 show that evaluators are likely to identify usability problems when users' verbalization and speech features exhibit certain patterns.

The second study, Study 2, assessed the generalizability of the findings. First, we examined whether the findings hold on different types of products (i.e., physical devices and digital systems). Second, we examined whether having access to additional information (i.e., video recordings of the think-aloud sessions and visualizations of the verbalizations) would affect when usability evaluators identify problems in the think-aloud session recordings.

In the rest of the article, we first describe the background and related work around think-aloud protocols and verbalizations during think-aloud sessions. We then describe the design and results of the two studies. To conclude, we discuss key findings that set the foundation for further research on automatically detecting usability problems based on the patterns in users' verbalizations and speech features in concurrent think-aloud sessions and future research opportunities.

2 BACKGROUND AND RELATED WORK

2.1 Think-Aloud Protocols

Ericsson and Simon [11] introduced and developed the theoretical framework for two types of think-aloud protocols: *concurrent think-aloud*, in which participants verbalize their thoughts during a task and *retrospective think-aloud*, in which participants verbalize their thoughts after a task.

Previous research compared concurrent think-aloud with retrospective think-aloud and found no difference in terms of task performance [33] or the total number of problem discovered [22]. In a recent survey study, McDonald et al. found that Ericsson and Simon's classic concurrent think-aloud protocol was the most frequently used one in usability testing [41]. Compared with retrospective think-aloud, classic concurrent think-aloud is considered to be more efficient, easier to perform and moderate [2], avoids biases arising from post-task rationalization [22, 41], and has been shown to have negligible influence on participants' behavior [17]. Relaxed concurrent think-aloud, in which evaluators actively probe or ask users questions [14, 41], is also used in practice, because it helps encourage users to verbalize their thoughts (e.g., [16, 50]). However, there has been much debate about whether relaxed concurrent think-aloud might impact user behavior (e.g., [2, 17, 28, 47]) and performance, in terms of accuracy and time. For example, some found that using relaxed thinking aloud protocol can consume less task time and commit fewer errors than working in silence [48]. Consequently, researchers are divided on the use of relaxed think-aloud protocols (e.g., [4, 17, 24, 38, 40, 47, 48]).

2.2 Verbalizations During Think-Aloud Sessions

In their work, Ericsson and Simon [11] categorized verbalizations during think-aloud sessions into three levels based on the amount of cognitive processing involved. Level-1 (L1) verbalizations occur when the thoughts being verbalized are in the person's present focus of attention and stored in verbal form. Level-2 (L2) verbalizations occur when the thoughts being verbalized are in the person's focus of attention but are stored in non-verbal form. Level-3 (L3) verbalizations require users to access their long-term memory, involving additional mental processing that may influence their focus of attention. For example, requesting explanation from participants will result in L3 verbalizations, which in turn could change the participants' task performance [6, 7]. Only L1 and L2 retrieve data from working memory and are considered to be valid verbalizations that have not been altered by external factors.

To encourage participants to make valid verbalizations, previous studies (e.g., [1, 9, 10, 17, 28, 49, 50]) have adhered to a set of guidelines that were proposed by Ericsson and Simon [11], i.e., use neutral instruction scripts to avoid bias, conduct practice trial sessions prior to data collection, and make neutral reminders to remind participants to think aloud when they fall silent. Breaching these guidelines may induce reactivity [13], increase mental workload (e.g., [28, 38]) or cause changes in participants' behavior (e.g., [17, 28]). Previous research also confirmed that without demonstration and practice before an actual think-loud session, participants may fail to report on their thought processes frequently or thoroughly [5]. We thus closely adhered to these guidelines when conducting concurrent think-aloud sessions in the two studies of this research.

2.3 Verbalization Categorization

Early work from Bowers and Snyder found that most verbalizations during classic concurrent think-aloud sessions were descriptions of participants' onscreen behavior [4]. Cooke later systematically studied the categories of verbalization produced during classic concurrent think-aloud sessions and identified four main categories as follows: reading, procedure, observation, and explanation [9]. Elling et al. later confirmed that these four categories covered the majority (over 80%) of verbalizations and added an additional category, which was specific to the tasks used [10]. Zhao and McDonald proposed a more detailed categorization that could be mapped to Cooke's four categories (e.g., "action description" corresponds to "procedure," "result evaluation," "user experience," and "recommendation" corresponds to "Observation") [49]. Later research examining verbalizations of think-aloud sessions (e.g., [16, 20, 29, 50]) often cited Cooke [9] and Zhao and McDonald's categorizations [49]. In this work, we use Cooke's four categories and extend the

current literature by exploring how these categories relate to usability problems in classic concurrent think-aloud sessions.

2.4 Speech Features

Speech features may be useful in identifying usability problems. For instance, the literature has shown that participants may still actively think even when they fall into silence [9] or use verbal fillers (e.g., "um," "ah") [8–10] during a classic think-aloud session. Speech features regarding how people speak can reveal their feelings, mood [36], and signs of high cognitive load [15, 45]. For example, sentiment relates to how a user feels, and hesitation in speech (e.g., using more verbal fillers) and a slower speech rate are associated to an increased cognitive load [15, 45] while interacting with a product. Users may compensate for the increased mental demand by directing more attention toward the task at hand, causing them to slow their speech, fall into complete silence [11, 37] or decrease the volume of their voice [9, 10]. Similarly, the pitch of the user's voice may become higher when users are excited or surprised. In this research, we explore how sentiment, silence, verbal fillers, speech rate, loudness, and pitch may be used by evaluators to identify potential usability problems.

3 STUDY 1: VERBALIZATION PATTERNS AND USABILITY PROBLEMS

Study 1 examined how *verbalizations* and *speech features* could be used to identify usability problems. The study consisted of two phases as follows: one to curate a dataset of think-aloud sessions and one to assess how verbalizations and speech features were used to analyze those sessions. In the first phase, we first conducted and recorded think-aloud sessions. In the second phase, we recruited usability evaluators to identify usability problems by reviewing these think-aloud sessions. Each evaluator was provided with a tool for reviewing a session's audio recording, for visualizing verbalizations and speech features, and for logging usability problems and the speech features used to identify the problems. As it is not uncommon for usability evaluators to transcribe recorded sessions in practice, our tool also visualized the transcripts for the sessions to explore how evaluators might leverage them. At the end of the study, we conducted semi-structured interviews to further understand how evaluators made use of the tool, verbalizations and speech features to identify usability problems.

3.1 Think-Aloud Data Collection

3.1.1 Participants. We recruited eight participants (five females, aged 19–24) from a student social group at a local university to participate in think-aloud sessions. To reduce any language issues that might interfere with their verbalization process, all participants were native English speakers. Participants had diverse background, including design, life science, cell biology, cognitive science, computer science, occupational therapy, psychology, and cinema studies. This diverse background was chosen to reduce the biases inherent to any discipline. Each participant was compensated with $20 for the hour-long study.

3.1.2 Procedure. We followed Ericsson and Simon's guidelines when conducting think-aloud sessions [11]. First, the moderator described the study details to the participant and played a short online video tutorial [31] on the think-aloud protocol. Afterwards, each participant was asked to perform three think-aloud sessions, using the primary functions of three devices, as follows: to set an alarm clock to *ring one hour from now*, program a multi-function coffee machine (De'Longhi BCO264B) to *prepare two cups of strong-flavored drip coffee for 7:30 in the morning*, and program a universal remote control (RCA RCRN03BR) to *operate a DVD player*. The alarm clock was given as a *practice trial* to help participants practice thinking aloud. The coffee machine and the universal

Table 1. Tasks for Two Devices Used in Think-Aloud Sessions

Device	Tasks
Coffee machine	Program the coffee machine to make two cups of strong flavor drip coffee at 7:30 in the morning.
Universal remote control	Program the universal remote control to operate a DVD player.

remote control were chosen particularly because they were representative of devices that people may use on a regular and occasional basis, respectively. All participants had not used these specific device models prior to the study. We counter-balanced the ordering of the two devices given to participants to ensure that there were an equal number of participants using each device first. For each think-aloud session, participants were given the device, a hard-copy of its instruction manual, and the task to perform. Table 1 shows the tasks, which involved using each device's primary functions. For the universal remote task, participants were also given a DVD player, and a TV that connects to the DVD player to carry out the task.

All think-aloud sessions were audio recorded. Because each participant performed two think-aloud sessions using two devices, there were a total of 16 think-aloud sessions, which formed the dataset for further analysis. The average duration of the sessions was 891 seconds ($SD = 222$) for the coffee machine and 649 seconds ($SD = 100$) for the universal remote control. The average duration of all sessions was 770 seconds ($SD = 208$).

3.2 Analysis of Think-Aloud Sessions

3.2.1 Participants. We recruited 16 participants (12 females) as usability evaluators to analyze the think-aloud audio recordings that were collected in the previous step. Ages ranged from 20 to 28 ($M = 24, SD = 2$). Two participants worked in the industry as UX evaluators and other participants were graduate students in UX programs or senior year undergraduate students who had previously taken UX courses at the university. All participants had previously conducted and analyzed think-aloud sessions and reported the number of projects, as part of a job, an internship or a course, for which they had employed the think-aloud method is as follows: 1–5 projects (12), 6–10 projects (2), and >10 projects (2). The overall experience of our evaluators was relatively less compared to usability evaluators who have worked in the industry for years. These participants are referred to as *usability evaluators* to distinguish them from those who participated in the think-aloud data collection.

3.2.2 Study Design. We counter-balanced the think-aloud sessions assigned to each usability evaluator so that (1) each usability evaluator analyzed two think-aloud sessions for distinct devices and users; (2) half of the usability evaluators began the study with a coffee machine think-aloud session. With this study design, each think-aloud session was analyzed by exactly two usability evaluators. Table 2 shows the counter-balancing scheme employed by the study.

3.2.3 Verbalization Categorization and Voice Features Extraction. In our pilot study, we experimented with automatic speech recognition (i.e., Web Speech API [43]) to generate transcripts of think-aloud sessions and found that it lacked accuracy and we had to devote substantial effort in correcting automatic transcription errors. Thus, in this study, we manually transcribed all think-aloud recordings.

Two coders followed a similar approach used in previous work [9, 10] to divide each audio recording into small audio segments and assign each audio segment with one of the four *verbalization categories as follows*: *reading, procedure, observation*, and *explanation*. The four categories

Table 2. The Counter-Balancing Scheme (p1–p8 Denote the Participants' ID in the Think-Aloud Data Collection)

Evaluator ID	1st session	2nd session	Evaluator ID	1st session	2nd session
	Coffee machine	Universal Remote		Universal Remote	Coffee machine
1	p1	p2	9	p1	p5
2	p3	p4	10	p3	p7
3	p5	p6	11	p2	p6
4	p7	p8	12	p4	p8
5	p2	p3	13	p5	p2
6	p4	p5	14	p7	p4
7	p6	p7	15	p6	p1
8	p8	p1	16	p8	p3

were based on the literature [9] and adjusted slightly to better fit our tasks: Reading (R): *read words, phrases, or sentences directly from the device or instructions*; Procedure (P): *describe his/her current/future actions*; Observation (O): *make remarks about the device, instructions, or themselves*; Explanation (E): *explain motivations for their behavior*. The beginning and end of an audio segment was determined by pauses between verbalizations and the content of these verbalizations, following the same procedure used in the literature [9, 10]. Each audio segment corresponded to a verbalization unit, which could include single words, but also clauses, phrases, and sentences.

We assessed the level of agreement between the two coders by computing the inter-rater reliability (IRR) for a single think-aloud session. The IRR score came out to be sufficiently high (Cohen's kappa: $k = 0.91$). For the audio segments that the coders labeled differently, they discussed and resolved disagreements. The remaining audio recordings were then labeled separately by the two coders.

We computed six voice features from each think-aloud audio recording: *sentiment, speech rate, loudness, pitch, silence,* and *verbal filler*. The sentiment of each audio segment was computed using VADER [21], a state-of-the-art sentiment analysis model. To compute the speech rate for each audio segment, we divided the number of words spoken in an audio segment by the segment's duration. Loudness (dB) and pitch (Hz) was computed using the speech processing toolkit Praat [51]. We manually labeled the start and end times of each period of silence and verbal filler. These voice features and the verbalization category labels were loaded and displayed in a tool that we created to assist usability evaluators to identify and log usability problems.

3.2.4 Tool for Analyzing Think-Aloud Sessions. We built a tool to assist usability evaluators to analyze and log the usability problems that they identified. The definition of a usability problem that we adopted was "anything that interfered with a user's ability to efficiently and effectively complete tasks" [23] and asked evaluators to consider any aspect of the products that might cause confusion, frustrations and/or hamper the user's ability to use them. Figure 1 shows the tool's interface. The left panel (a) visualizes the transcript of a think-aloud session. The right panel (b) visualizes the six voice features and the verbalization categories for all the audio segments in the think-aloud session recording. The right panel (c) provides functions to log the identified usability problems.

Clicking on any point of a feature panel brings the audio to the corresponding timestamp and subsequent pressing of the ESC key plays the audio from that timestamp. Dragging the cursor on any feature panel highlights a portion of the visualization. The background color of the selected

Fig. 1. The tool for evaluators to analyze a think-aloud session (center). It visualizes the transcript of the recording on the left panel (a), one line per audio segment labeled with a verbalization category. The seven audio features are represented as charts on the right panel (b), which is enlarged and shown in the top window. Highlighting any part of a chart will highlight the corresponding transcript on the left panel. The bottom of the tool (c) allows an evaluator to describe usability problems that they identified and the features (i.e., category, silence, verbal fillers, sentiment, speech rate, loudness, pitch, and transcript) that indicate these problems.

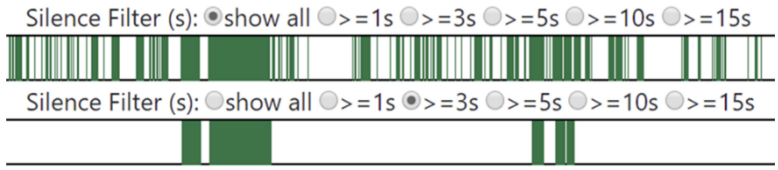

Fig. 2. The Visualization of the silence (the colored bars) before and after selecting a silence length filter.

Table 3. The Frequency and Percentage of Audio Segments Labeled with Each Verbalization Category

Device	Verbalization category			
	Reading	Procedure	Observation	Explanation
Coffee machine	276 (29%)	235 (25%)	371 (40%)	52 (6%)
Universal remote	185 (28%)	177 (27%)	256 (39%)	33 (5%)
All devices	461 (29%)	412 (26%)	627 (40%)	85 (5%)

portion in all features panels turns grey to indicate the highlight. After highlighting, pressing ESC plays the audio from the start of the highlighted portion. Because longer periods of silence may reveal different information about the verbalization than shorter ones, the tool also provides five length filters (>1s, 3s, 5s, 10s, and 15s) to allow usability evaluators to selectively focus on longer- or shorter- durations of silence (Figure 2).

The bottom of the tool provides functions for logging usability problems. To ease the logging of the start and end time of a usability problem, the tool automatically detects and fills these two timestamps whenever usability evaluators highlight a portion of any chart or the transcript. Inspired by previous work [26, 27], we used a structured problem report that included a description of the user problem, the problem's context and verbalization features that indicated the problem. Specifically, the text fields and checkboxes at the bottom of the UI allows evaluators to describe usability problems and select the verbalization features that they used to identify problems. To better visualize the temporal relationship between usability problems and all visualized features, a colored segment will be visualized on the "Problem" timeline (between the panel (b) and (c) in Figure 1) when a usability problem is added.

3.2.5 Procedure. Prior to the start of the study, the facilitator showed each evaluator the product that the think-aloud user used in each recording, and informed them the functions of the product, and the task that the think-aloud user was working on. Then, the facilitator asked the evaluator to identify and log the problems that users were experiencing in the audio recordings and the features that guided them to identify the problems using the tool (Figure 1). The facilitator introduced the tool's function, how to use it, and then gave each evaluator a few minutes to familiarize themselves with the tool. Each evaluator had a maximum of 30 minutes to analyze each of the two think-aloud audio recordings that were assigned to them. After analyzing the audio recordings, we conducted semi-structured interviews to understand how evaluators identified problems and made use of verbalization features. The entire study lasted for about 1.5 hours. Each evaluator was compensated with $20.

3.3 Analysis and Results

3.3.1 Number of Labels per Verbalization Category. We quantified the number of times that the four verbalization categories were used as labels in all 16 think-aloud audio recordings. Table 3

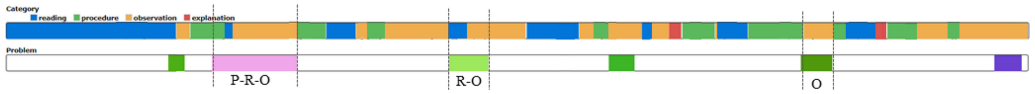

Fig. 3. The verbalization categories for a think-aloud recording and the problems identified by an evaluator.

displays this information for each device separately and in tandem. Notably, the four categories were used in similar proportions for both devices.

3.3.2 Problems Used in the Analysis. In total, 273 problems were identified by the usability evaluators, particularly 148 in the think-aloud sessions for coffee machine and 125 for universal remote. Two of the authors validated each problem that was logged, by checking the problem description and listening to the corresponding audio segment independently. Disagreements were resolved via discussion. Of these problems, seven were assessed to be invalid because they were either missing proper problem description or the problem descriptions did not match with the content of the associated audio segment. We considered the remaining total of 266 problems in analysis.

The average number of problems identified per evaluator for each device is as follow: coffee machine ($M = 9, SD = 3$) and universal remote ($M = 8, SD = 3$). A repeated-measures ANOVA test with Bonferroni correction found no significant difference in the number of problems identified between the two devices ($F(1, 15) = 3.78, p = .07, \eta_p^2 = .20$).

3.3.3 Verbalization Categories and the Identified Problems. To explore how verbalization categories are related to the problems that users experienced, we conducted an analysis of the problems that were logged by the usability evaluators. First, for each problem logged by an evaluator, we counted the number of different verbalization categories that fell into the problem's start and end times. Figure 3 shows the audio recording of a think-aloud session with labeled verbalization categories (top) and the problems identified by an evaluator (bottom). For example, the first problem shown in Figure 3 was associated with three categories (i.e., *Procedure, Reading*, and *Observation*), which occurred once each.

To better understand the correlation between verbalization categories and usability problems, we computed the *precision* and *recall* of each verbalization category in locating usability problems using the following equations:

$$precision = \frac{\text{the number of segments labeled as a particular category associated with an identified problem}}{\text{the total number of segments labeled as the same category in the entire session}}$$

$$recall = \frac{\text{the number of segments labeled as a particular category associated with an identified problem}}{\text{the total number of segments associated with an identified problem}}$$

We used precision and recall as measures because they account for the base rate of each verbalization category in a think-aloud session when considering their relationship with usability problems. Precision and recall can be used to answer the following two questions:

(1) If an evaluator randomly checks a segment labeled with a particular category, what is the chance of finding a problem?

(2) If an evaluator checks all segments labeled with a particular category, what percentage of problems could be found?

The greater precision of a verbalization category indicates that evaluators would have a higher chance of finding a problem by examining a segment labeled as the category and the greater recall of a verbalization category indicates that evaluators would be able to catch more problems if they examine segments labeled as the category. Furthermore, to assess the overall relevance of a verbalization category with usability problems, we further computed the *F*-measure, which combines precision and recall as a single measure using the following equation: $\frac{2*precision*recall}{precision+recall}$. Figure 4

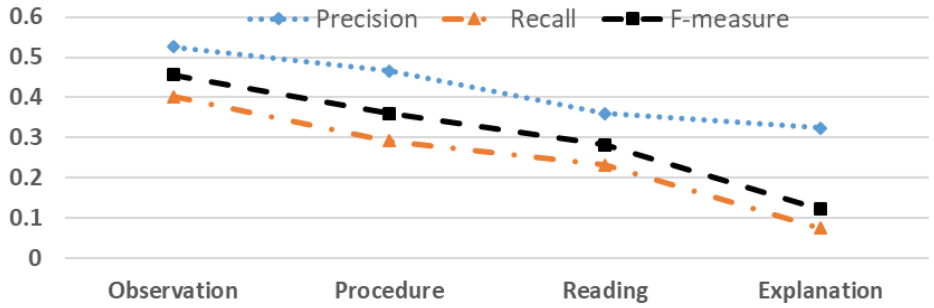

Fig. 4. Precision, recall, and *F*-measure of each verbalization category in identifying problems.

Table 4. Precision, Recall, and *F*-measure of each Category in Identifying Problems for Each Test Device

Category	Precision		Recall		*F*-measure	
	Coffee machine	Universal remote	Coffee machine	Universal remote	Coffee machine	Universal remote
Observation	0.54	0.52	0.40	0.40	0.46	0.45
Procedure	0.47	0.46	0.30	0.28	0.37	0.35
Reading	0.37	0.35	0.23	0.24	0.28	0.28
Explanation	0.29	0.37	0.07	0.08	0.11	0.13

shows the precision, recall, and *F*-measure of each verbalization category in identifying usability problems. It shows that while the segments labeled as *Observation* are the most relevant to usability problems, the segments labeled as *Explanation* are the least relevant to usability problems. The segments labeled as *Procedure* or *Reading* are also relevant to usability problems, but less so than the ones labeled as *Observation* and more so than the ones labeled as *Explanation*.

We computed the precision, recall, and *F*-measure of the four categories in identifying problems for each device separately to examine if the trend shown in Figure 4 still holds for different devices. Table 4 shows these measures for each device respectively. The numbers in each column under each one of the three measures in Table 4 generally decrease, which indicate that the same trend we observed in Figure 4 largely holds for each device separately.

To understand how evaluators used the combination of verbalization categories in identifying problems, we further computed the precision, recall, and *F*-measure of twelve pairs of verbalization categories in identifying problems. For example, the pair "R-O" refers to one *Reading* category segment or an uninterrupted sequence of the *Reading* category segments followed by one *Observation* category segment or an uninterrupted sequence of the *Observation* category segments. The category pairs are mutually exclusive. Table 5 shows the results, which suggest that the verbalization pairs that are most relevant to usability problems typically contain the *Observation* category and the verbalization pairs that are least relevant to usability problems contain the *Explanation* category. Comparing Tables 4 and 5, we can see that the *Observation* category was more relevant to usability problems than any verbalization pairs.

3.3.4 Speech Features. Usability evaluators also selected the features that guided them in finding usability problems using the tool (Figure 1). Figure 5 shows the number of times that each feature was used by all evaluators. Category and sentiment were among the most highly-used features, while pitch and loudness were among the least. A repeated-measures ANOVA with

Table 5. Precision, Recall, and *F*-measure of the Verbalization Category Pairs in Identifying Problems

Verbalization category pair	Precision	Recall	*F*-measure
R-O / O-R	0.27	0.46	0.34
P-O / O-P	0.26	0.33	0.29
P-R / R-P	0.21	0.11	0.15
O-E / E-O	0.30	0.05	0.08
P-E / E-P	0.18	0.04	0.06
R-E / E-R	0.16	0.01	0.02

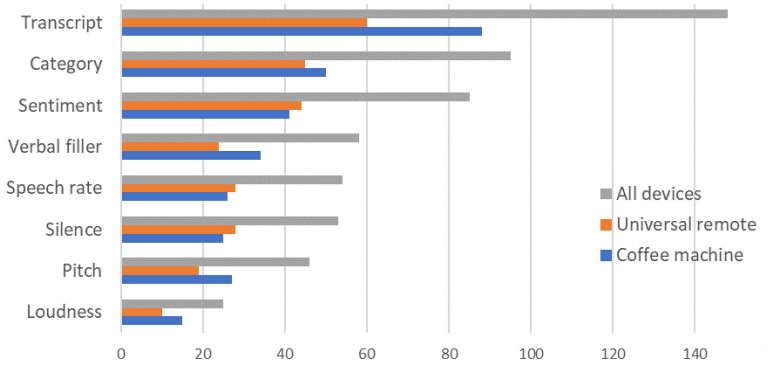

Fig. 5. The number of times that each feature was used by evaluators for finding usability problems.

Bonferroni correction found no significant difference between the features except the following: pitch and transcript ($p = .04$), loudness and transcript ($p = .008$). In addition to the speech features, evaluators often frequently used the transcript in their analysis.

3.3.5 Qualitative Feedback. Two researchers transcribed the interviews and coded the transcripts independently. They then discussed to consolidate their codes. In this section, we present the key findings to provide a deeper, more detailed understanding of how evaluators used *verbalization* and *speech features* to identify problems that users experienced.

Verbalization category. Evaluators underscored that the *Observation* category was most indicative of problems (*"Observation describes how the users were feeling and how they commented their confusions"-ev6).* Some evaluators relied on segments labeled *Observation* to help them focus on finding problems quicker (*"I know that most of the problems aren't going to be in Reading or Procedure. Instead, they would be in Observation."-ev15*). Moreover, some found that *Observation* audio segments with a long duration or sometimes contained some explanations signaled a problem (*"When users are confused, they do a lot more Observations and sometimes explanations. You'll see less of the Reading and Procedure."-ev7*). Evaluators also noted that *Observation* category contained a diverse amount of information, which is not necessarily related to problems (*"It could be users expressing a problem but could also be them commenting something worked"-ev1*). These feedback is consistent with the quantitative measures in Section 3.3.3.

Evaluators generally thought that the *Reading* category was tied less to problems, mainly because *"reading was just users reading instructions"-ev13.* On the other hand, some evaluators noted that the *Reading* category was still useful in indicating problems. For example, a user who is confused about a set of instructions may repeatedly regurgitate them (*"I noticed that if they were*

confused, they tended to read the instructions more. If they knew what they would do, they would intuitively work through it."-ev1; "Repetitions mean that they try to say or do the same thing over and over again in a short period of time. When someone is experience difficulties, you might see such repetitions."-ev7).

While the *Observation* category was the most useful one for identifying problems, evaluators emphasized that category combinations were helpful in providing context as to why users were encountering problems (*"When the problem presents itself, it is usually in the Observation category. But the real problem was usually already there for a while. You need to go back and read or listen to the segment before the Observation to understand the context."-ev13; "Reading becomes important to understand a comment when the user expressed an Observation after Reading, such as 'oh, I don't understand… '"-ev10*).

Sentiment. When using sentiment, evaluators mostly focused on audio segments with negative sentiment ("*I mostly checked the low part of the sentiment chart. When they are unsatisfied or confused, they naturally tend to say negative words, which would be the low part of the chart."-ev15*). Evaluators gauged user sentiments by examining the transcript as well (*"Sentences with negative sentiment, such as 'it sucked,' were the ones that I tried to find while reading the transcript. The sentiment is important but having to looking at the transcript and sentiment chart at two different places is a bit distracting."-ev2*).

Apart from using negative sentiment as a place to look for problems, evaluators also paid attention to sudden changes in sentiment (*"'I feel like they should do that unless… Oh, No, OK.' When there are two words back and forth that one is negative and the other one is neutral or positive. It means that they changed their tone immediately, which usually indicate their confusions"-ev14).* In essence, abrupt transitions in sentiment might be *Eureka moments* (or *Aha! moments*) for the user [3], i.e., the sense of suddenly coming to an understanding of a concept that was previously confusing. In usability testing, Eureka moments might imply that a product does not follow users' intuitions and is likely not easy to use.

Evaluators also noted that a shortcoming of the sentiment voice feature is that it is based solely on the contents of a verbalization (i.e., what was said) and does not give insight into how such verbalizations are made (*"It is possible that the same content can mean different things until I listen to it."-ev1).* As a result, the sentiment feature was sometimes inaccurate because it failed to consider one's tone of voice, which may at times be more telling of a user's emotions rather than what was verbalized, as in the case when users are being sarcastic. For example, sentences like "*oh, that's helpful*" may be negative in actuality, but be classified as positive using the text-based sentiment analysis approach. Thus, evaluators suggested that listening to the audio can be important in assessing the true sentiment of a sentence.

Verbal fillers. Some evaluators mentioned that users would use more verbal fillers right before and during the presence of problems. Rather than using the verbal filler chart, evaluators reported that they primarily used the transcript to look for the verbal fillers. Many evaluators had expressed a desired to have the verbal fillers be more visually salient in the transcript, such as by highlighting them. Evaluators also noted that they could not rely only on verbal fillers in making judgments about usability problems, as people's use of them can vary widely: some people may use verbal fillers sparingly, while some people may use them habitually. To gain a sense of users' manner of speech, evaluators suggested engaging in a conversation with them prior to a think-aloud session.

Words such as "*what?*" "*where?*" and "*how?*" were also considered to be verbal fillers and evaluators found them to be useful in identifying audio segments with problems (*"There are certain things that you can say to show your confusion without literally say 'I'm confused.' For example, you may*

use 'huh' or ask questions, like what? Where? These words mean that you are confused. Otherwise, you wouldn't be asking questions."-ev14).

Speech rate. Like verbal fillers, speech rate varies from individual to individual and is therefore difficult to use as a telltale sign of usability issues. In spite of this, some evaluators noted that a lower-than-normal speech rate may indicate that users were thinking, confused, or interpreting instructions (*"I looked at parts of the chart that were below the average, because when the user in the first session I analyzed had a problem, she spoke slower."-ev5*). Moreover, a higher-than-normal speech rate could also indicate problems. One evaluator mentioned that the user of one think-aloud session was reading the instructions very fast when she had trouble finding the right content. However, high speech rate can be unreliable, since users may speak quickly even though they are not encountering problems.

Silence. Evaluators also made use of periods of silence in verbalizations as a sign of usability issues. In particular, evaluators took advantage of the filtering function (Figure 2) to look for prolonged periods of silence (>3s) that may suggest user confusion (*"I felt that with 1 second filter, there are too many left. With 3 seconds, there are reasonable number of silences for me to analyze."-ev9*).

Similar to speech rate and verbal fillers, relying primarily on silence could result in making false conclusions. Users may fall silent for reasons other than usability issues, such as when they are operating a machine, thinking, or when quietly reading or comprehending instructions. To gain contextual information, evaluators reported examining audio segments occurring just before and after silent periods.

Pitch and loudness. Pitch and loudness were the least used features. Evaluators felt that it was difficult to detect patterns in the pitch and loudness charts since they did not have much meaningful variation (*"The chart was mostly the same kind of looking, so it's hard to tell exactly what's meaningful."-ev5).* They, however, still believe that pitch and loudness can be useful to assess a user's level of confidence or the state of confusion, such as when users speak more quietly or raise their pitch (*"When users are losing confidence in what they are doing, the loudness of their voice tends to be lower."-ev6; "Whenever a user ends a sentence with a higher pitch like asking questions, it has always been that he is confused."-ev14).*

Transcript. Evaluators reported that having access to audio transcripts saved time because it *"got more into users' head"-ev2* and allowed them to attend to important or interesting verbalizations without having to listen to the audio recording all the time. For example, they noted that they could easily skip irrelevant audio segments (*"I skipped [listening to] the parts that I knew were just them describing what they were doing."-ev1*) and focus on problematic segments (*"I highlighted the part in the transcript that seems to be a problem and then listened to the audio and analyzed the charts on the right."-ev5*). This feedback is consistent with the log data, which showed that on average, evaluators only listened to 70% of the think-aloud audios.

Evaluators also expressed that the transcript helped identify verbal fillers and other remarks made by participants that were signs of usability problems, such as *"I'm going to start this over again or I'm stuck"-ev9.*

3.4 Summary

We present and discuss the findings about how verbalization categories and speech features relate to usability problems in this subsection.

Verbalization categories. As evidenced by this study, usability evaluators benefitted from having access to an audio recording's verbalization categories. Firstly, our results revealed that audio segments with the *Observation* category were more indicative of usability problems than other categories, presumably because these audio segments often described a user's concerns about a product or their behavior. Additionally, the *Reading* category helped to pinpoint places where

users had difficulties in making sense of instructions, as they would spend long periods of time reading instructions; often repeating the same set of instructions over and over again. Segments categorized as *Procedure*, as in prior work [29], helped evaluators understand and assess the ease at which users could follow a set of instructions.

Verbalization categories were also helpful in finding contextual information to understand the problems faced by users, particularly segments categorized with *Reading* or *Procedure*, as these segments described actions that users attempted to perform.

The segments that were least associated with problems contained the *Explanation* category. One reason could be that the audio segments with this category were low in general (5%). This number is in line with those reported in previous studies (e.g., 5% in Cooke's study and 7% in Elling et al.'s study), perhaps implying that users tend to not explain or provide motivation for their behavior. One example of an *Explanation* category segment following a *Procedure* category segment from a universal remote control session was as follows: "*let's try Auto Code search, because it says it's the easiest method.*"

Notably, the pairs that were most closely associated with problems were the combinations of *Observation* (O) with either *Reading* (R) or *Procedure* (P). In particular, the accumulated recall of the top four pairs (*R-O, O-R, P-O, O-P*) was 0.79, which suggests that evaluators could find 79% of the problems when examining these pairs. This is perhaps because the context information provided by *Reading* or *Procedure* segments is needed to understand problems in *Observation* segments. In the ideal case where no problems are encountered, a user's verbalizations should alternate between *Reading* and *Procedure*. We posit that such pairs, in which users deviate from reading and performing procedures to make an observation, indicate that they may be facing difficulties. In addition, the likelihood that users are facing difficulty increases with the amount of deviation from reading and performing procedures to make an observation. However, further investigation is needed to confirm this speculation.

As shown in this study, the *Observation* category was the greatest telltale sign of problems, with around half of all audio segments containing the *Observation* category label being tied to a usability problem (see the precision values shown in Table 4). This result implies that with a roughly 50% rate of accuracy, usability evaluators can identify problems when randomly examining a segment labeled as the *Observation* category. As the recall values for the *Observation* category were also around 0.5, usability evaluators would find around half of the usability problems if they only focused on segments labeled as the *Observation* category. The implication is that although the *Observation* category is the greatest telltale sign of problems, usability evaluators should also leverage other information to increase the chance of identifying usability problems. For example, for greater reliability when examining *Observation* segments, many evaluators suggested combining *Observation* and *negative sentiment* information, on the grounds that if an *Observation* segment is about something working as expected, the corresponding sentiment would not be negative. However, as text-based sentiment analysis is inaccurate, this approach still requires evaluators to refer to the corresponding audio segments.

Voice features. Evaluators found that all the voice features were useful, especially sentiment. They often used *sentiment* together with *category* (e.g., the *Observation* category and *negative* sentiment) to quickly focus on interesting segments of the transcript or audio. Regarding the visual design of the tool, evaluators expressed a desire for sentiment and verbal filler information to be combined with the transcript, as opposed to being visualized in separate charts, as integrating these features may reduce the spread of their attention on the tool's user interface. Evaluators also proposed other useful parts of speech that may indicate problems, such as when users ask questions (i.e., What? Where? How? Huh?). *Repetitive patterns*, such as reading a set of instructions over and over again or performing actions repeatedly, also raised red flags.

Because verbal fillers, speech rate and silence tend to vary from individual to individual, evaluators felt that they would need to speak to the participants to get a sense of their normal speech patterns in order to use these features. The implication is that although these voice features are potentially useful to identify usability problems, knowing a user's colloquial speech habits (i.e., the baseline of the voice features) might help evaluators better leverage these features.

4 STUDY 2: GENERALIZATION STUDY

We found the relationship between users' verbalization, speech features and usability problems in Study 1. There are three factors in the study design that concern the generalizability of its findings.

1) Physical Devices vs. Digital Systems. In Study 1, physical devices were used for think-aloud sessions. Digital systems, such as websites, are another type of products that require extensive usability testing and have been used as test products for think-aloud related research (e.g., [2, 9, 10, 16, 17, 24, 28, 29, 35, 47, 49, 50]). People operate digital systems (e.g., websites) differently than physical devices. Physical devices have fixed interfaces with a limited number of controls that the user can interact with. The challenge with completing tasks on physical devices might be figuring out how to map steps and actions to features and controls. In contrast, digital systems have a different set of constraints. The challenge here might be finding specific interface features to satisfy the user's need. Additionally, limb motion is often required for operating physical devices, while digital systems require more eye motion and relatively small-scale hand movement (e.g., operating a mouse).

2) Verbalizations with Audio Recording vs. Verbalizations with Video Recording. We designed the Study 1 so that usability evaluators assessed think-aloud sessions from their audio recordings to avoid the potential influence of other modalities and to better assess the role of verbalizations. Although evaluators were able to identify problems to a proficient degree from just the audio recordings, it would be interesting to explore whether including other modalities, such as video, might have added benefit or change the verbalization patterns that indicate problems.

3) Visualization of Verbalization and Speech Features. In Study 1, usability evaluators had access to a visualization of verbalizations (e.g., verbalization categories and speech features). This might have influenced their analyses since the visualizations might have directed their attentions to certain parts of the sessions more often than others and subsequently led them to identify more or less problems.

Thus, in this study (Study 2), we sought to answer the following three research questions:

Research Question 1: Are verbalization patterns that signal usability problems different for *physical devices* and *digital systems*?

Research Question 2: Are verbalization patterns that signal usability problems different when a *video* recording of a think-aloud session is also provided?

Research Question 3: Are verbalization patterns that signal usability problems different when a *visualization* of verbalizations is not provided?

4.1 Think-Aloud Data Collection

4.1.1 Participants. We recruited a new set of participants ($N = 8$, 4 females, aged 19–26), all of whom were native English speakers, from student social groups at a local university. Like Study 1, native English speakers were chosen to reduce language barriers. Participants had diverse backgrounds, such as biology, creative writing, environmental science, neuroscience, and pharmacology.

4.1.2 Procedure. The study's procedure was the same as the first study. The products tested in this think-aloud data collection included *two websites* in addition to the two physical devices, the

Table 6. Tasks for the Two Websites Used in Think-Aloud Data Collection. STM Stands for the National Science and Technology Museum, and HM Stands for the National History Museum

Websites	Tasks
STM	Your friend is an 8th grade science teacher. She asks you to check if there are any available school programs in April at the Science museum. Your task is to find out whether there are any programs that may be suitable for 8th grade students in April.
STM	Your uncle has an 11-year-old child. One day, the child asks you a question, "what is it like to be a scientist or an engineer?" You've heard that the museum offers interactive presentations during which children can interact with speakers, who are scientists. Thus, your task is to find out if there is any such program in March for an 11-year-old child.
STM	You are a college student and are working on an assignment about early telescopes. Your task is to obtain a photo of an instruction manual, which is for an early telescope.
HM	Your friend is a 7th grade teacher. She is organizing a trip for 30 7th grade students to the history museum. Your task is to help your friend find an available program in March for 30 7th grade students.
HM	Your friend has a 4-year-old child and is planning to take him to the history museum. Please help your friend to find out the number of the activities that are appropriate for a 4-year-old child in March.
HM	You are a graduate student and are currently doing research on the topic of first peoples in Canada. Your task is to search for an essay on the topic.

coffee machine (*De'Longhi BCO264B*) and the universal remote control (*RCA RCRN03BR*): a national science and technology museum (*STM*) and a national history museum (*HM*) website. These two websites were chosen as they represented websites that our participants would potentially be users of, and they possessed certain number of usability problems, as was determined by a preliminary heuristic evaluation conducted by the first author. Similar to the tasks in Study 1, we identified three tasks that covered some of the target websites' main functions. Table 6 shows the tasks for each website.

All think-aloud sessions were video- and audio-recorded. Participants performed all website-based tasks on a 27" 4K monitor, which was connected to a laptop and placed on a desk. All website-based task sessions were screen captured with a picture-in-picture window of a participant's face using a Logitech HD Pro Webcam and the Open Broadcaster Software. The think-aloud sessions of participants using two physical devices (the coffee machine and the universal remote) were captured with two wall-mounted cameras, which monitored each participant's face and hand movements. For better quality audio, we used a clip-on voice recorder instead of the camera's embedded microphones and later manually synchronized audio and video streams. Each participant was compensated with $20 for the hour-long study.

In total, 64 think-aloud sessions were recorded (each participant performed 8 think-aloud sessions: one task for each physical device and three tasks for each digital website). All sessions ranged from 62 seconds to 1,255 seconds ($M = 360$, $SD = 279$). The average duration of the sessions for each device or website was as follows: coffee machine ($M = 854$, $SD = 251$), universal remote control ($M = 619$, $SD = 195$), STM($M = 222, SD = 131$), and HM($M = 247, SD = 153$).

Table 7. A Balanced Latin Square Design for Four Evaluators

Coffee machine	Remote	History museum	Science and technology museum
Audio + Visualization	Video + Visualization	Video	Audio
Video + Visualization	Audio	Audio + Visualization	Video
Audio	Video	Video + Visualization	Audio + Visualization
Video	Audio + Visualization	Audio	Video + Visualization

4.2 Analysis of Think-Aloud Sessions

4.2.1 Participants. We advertised the study in several local UX/HCI social groups via Facebook and Slack. We recruited 16 participants (11 females) as usability evaluators to analyze the think-aloud sessions. Their ages ranged from 22 to 50 ($M = 27, SD = 7$). Their self-reported professions were: usability specialist (2), UX designer (2), UX researcher (1), graduate students specialized in UX (11). All participants had experience with the think-aloud method from their jobs, internships, and/or graduate course projects. The number of projects for which they had used the think-aloud method to conduct usability tests was as follows: 1–5 (3 participants), 6–10 (11), and >10(2).

4.2.2 Study Design. We counter-balanced three factors—test products (i.e., the physical devices and digital websites), modality of think-aloud recordings, and visualization—through a balanced Latin-square design so that each evaluator analyzed all four of the test products. Evaluators analyzed two of their sessions with the audio recording. For the other two, evaluators were given the video-recording, which came with the audio as well. Additionally, each evaluator only had access to the visualization in two of the four sessions that they analyzed. An example of the assignment mechanism for four usability evaluators is shown in Table 7. To reduce potential carry-over effect between test products, we altered their order according to a 4×4 balanced Latin-square design for the rest evaluators. The sessions assigned to each evaluator were also conducted by different think-aloud participants to avoid potential biases that might occur if they were to analyze the same think-aloud participant's sessions more than once. Note that for each website, an evaluator analyzed three recordings, each one corresponding to a task.

4.2.3 Verbalization Categorization and Voice Feature Extraction. We followed the same process described in Study 1 to create verbalization categories and the speech features (i.e., silence, verbal fillers, sentiment, speech rate, loudness, and pitch) for each think-aloud recording. All these features were loaded and displayed on the analysis tool as described in the next section.

4.2.4 Tool's Interfaces for Different Study Conditions. We updated our think-aloud analysis tool with a different interface (Figure 6) for each experimental condition (Table 7): *Audio, Video, Audio + Visualization and Video + Visualization.* Figure 6(a) shows the interface when evaluators had access to the audio recording of a think-aloud session, which included an audio player and controls to play and pause audio. Figure 6(b) shows the interface when evaluators only had access to the video recording of a think-aloud session, which included a video player and controls to play and pause video. Figure 6(c) shows the interface when evaluators had access to both the audio recording of a think-aloud session and the visualization of verbalizations. The visualization of the verbalizations was the same as that of Study 1, which include a transcript, a verbalization category chart, and seven voice feature (i.e., silence, verbal fillers, sentiment, speech rate, loudness, and pitch) charts. Figure 6(d) shows the interface when evaluators had access to both the video recording of a think-session and the visualization of the verbalizations. Figure 7 shows a close-up view of the interface for *Video + Visualization.*

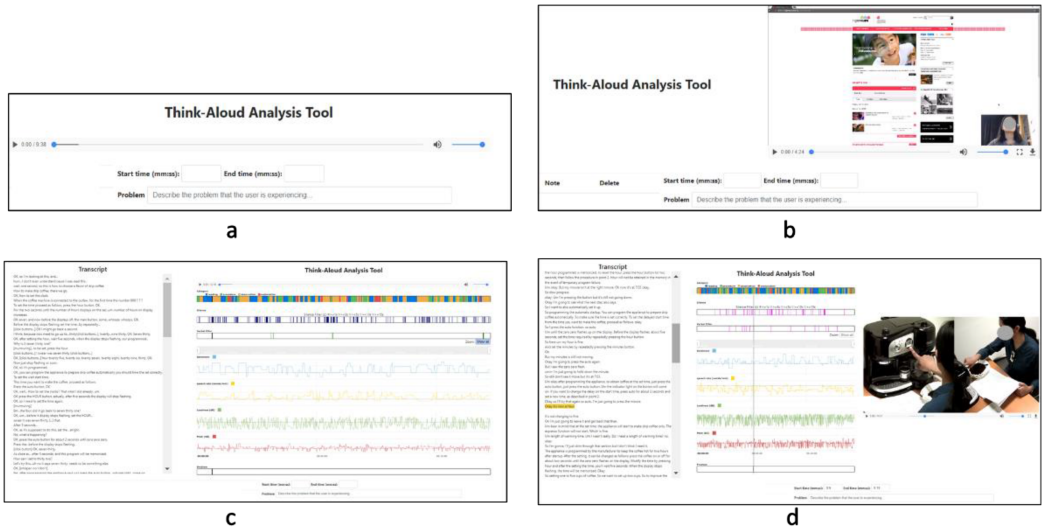

Fig. 6. Think-aloud analysis tool's four interfaces for four different conditions as follows: (a) audio only; (b) video only; (c) audio + visualization; (d) video + visualization.

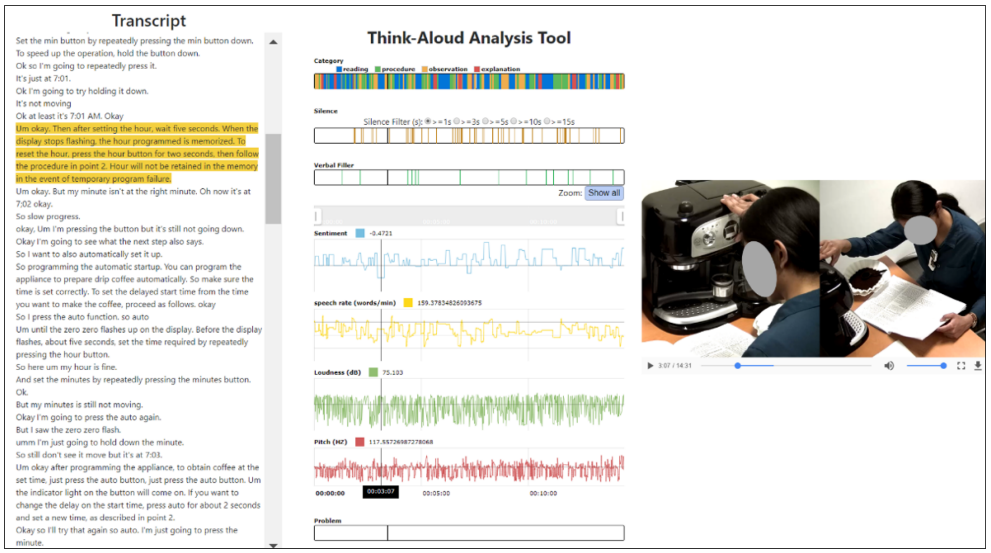

Fig. 7. The tool's interface for *Video + Visualization* condition. The left two columns are the same as the *Audio + Visualization* condition. The right column shows the video recording of a think-aloud session.

In all conditions, usability evaluators were asked to specify the time period during which the user in the think-aloud session encountered a problem and to describe the problem in plain text using the logging functions on the interface, which was the same as described in Figure 1. All the information was automatically saved into a log file.

4.2.5 Procedure. Prior to the start of the study, the facilitator informed each evaluator that the product that the think-aloud user used, its main functions, and the task that the user was working on in each recording that the evaluator would analyze. Then the facilitator informed evaluators

Table 8. The Percentage of Segments Labeled with Each Verbalization Category for Each Testing Object

Device or website	Verbalization category			
	Reading	Procedure	Observation	Explanation
Coffee machine	28.4%	28.4%	35.4%	7.8%
Universal remote	29.4%	29.6%	36.9%	4.1%
Science and tech museum	23.7%	30.7%	37.2%	8.4%
History museum	23.0%	35.7%	37.2%	4.1%
All together	26.1%	31.1%	36.7%	6.1%

that they would identify and log the problems that users were experiencing in the recordings using the tool (Figure 6). Then the facilitator introduced the tool's functions, explained how to use it, and gave each evaluator a few minutes to familiarize themselves with the tool. Because think-aloud sessions varied in length, evaluators were allocated 1.5 times the length of a session to spend on their analysis. When the evaluator finished analyzing a session or if allocated time was up, the evaluator was asked to proceed to their next session. The study lasted about 2 hours in total and each evaluator was compensated with $40.

4.3 Analysis and Results

4.3.1 Number of Labels per Verbalization Category. We quantified the number of times that the four verbalization categories were used as labels in all think-aloud recordings. Table 8 displays this information for each device and website separately and as a whole. The results appeared to be similar to that of Study 1 (see Table 3), despite including digital systems as a testing object and the new pool of participants who took part in the study. The labels used for the verbalizations appeared in similar proportions across the devices and websites as follows: (1) roughly 60% of the verbalizations were about users reading contents (*Reading*) or describing their actions (*Procedure*); (2) the *Observation* category was the most popular single category and around one third of verbalizations were given the *Observation* label; (3) the *Explanation* category was the least popular category and appeared significantly less than all the other three categories. A repeated-measures ANOVA test with Bonferroni correction found significant differences between four categories ($F(3, 21) = 46.96, p < .01,\ \eta_p^2 = .87$). Post-hoc results show that the following: (1) the *Observation* category appeared significantly more than the *Reading* category ($p = .04$) and the *Explanation* category ($p < .01$); (2) the *Explanation* category also appeared significantly less than the *Reading* or the *Procedure* category ($p < .01$).

4.3.2 Problems Identified by Usability Evaluators. In total, usability evaluators identified 418 problems. Two of the authors validated each problem that evaluators had identified, by checking the problem description and listening to (or watching) the corresponding audio (or video) segment. Any disagreements about the correctness of the logged problems were discussed and resolved. Of these problems, 33 were assessed to be invalid because these problems either (1) missed the starting or ending timestamp, which made it impossible to know when evaluators thought that users were encountering problems; or (2) the problem descriptions provided by evaluators did not match the content in the corresponding audio or video segments. With these problems removed, we considered a total of 385 problems in all subsequent analyses.

The average number of problems identified per evaluator for each physical device or digital website is as follows: coffee machine ($M = 5.9, SD = 2.3$), universal remote ($M = 5.1, SD = 3.8$), science and tech museum ($M = 6.6, SD = 5.3$), and history museum ($M = 6.5, SD = 3.4$). A

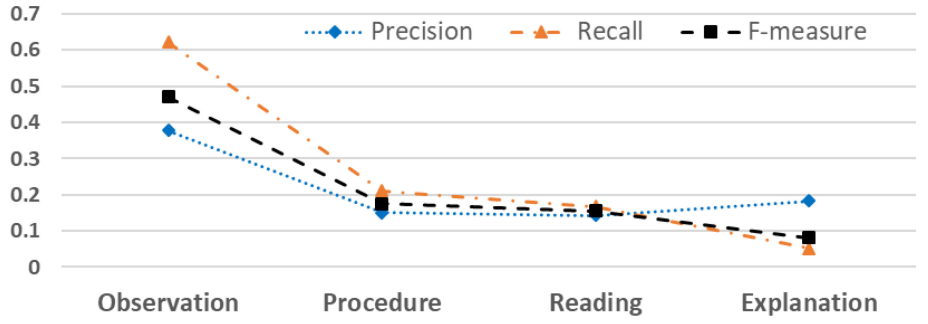

Fig. 8. Precision, recall, and *F*-measure of each verbalization category in identifying problems.

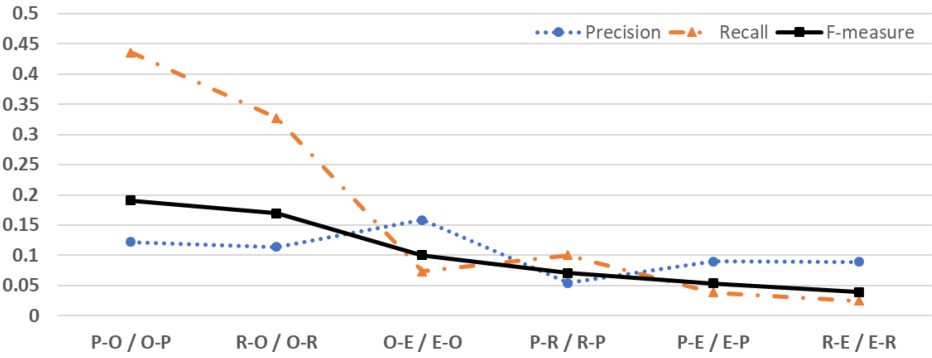

Fig. 9. Precision, recall, and *F*-measure of each verbalization category pair in identifying problems.

repeated-measures ANOVA with Bonferroni correction found no significant difference ($F(3, 45) = 1.42, p = .25, \eta_p^2 = .09$).

4.3.3 Verbalization Categories and the Identified Problems. We followed the same procedure as described in Study 1 to analyze the relationship between verbalization categories and the identified problems. We computed the *precision, recall*, and *F-measure* of each verbalization category in locating problems. Results (Figure 8) show that the *Observation* category was most likely associated with problems. The *Explanation* category was still least likely associated with problems. The general trend here is consistent with the trend shown in Study 1 (see Figure 4).

We further computed the precision, recall, and *F*-measure of each *verbalization pair* in identifying usability problems (Figure 9). The results reveal that the pairs with the highest precision and recall all contained the *Observation* category. Particularly, pairs of *Observation* and *Procedure* (P-O or O-P) and pairs of *Observation* and *Reading* (R-O or O-R) were most likely associated with problems. Pairs of *Observation* and *Explanation* (E-O or O-E) had a relatively high precision. The implication here is that a large amount of usability problems can be detected simply by focusing on the *Observation* category as the *Observation* category has a higher precision, recall, and *F*-measure than any given pair.

4.3.4 Physical Devices vs. Digital Systems. We analyzed whether there were differences between *physical devices* and *digital systems* on how verbalization patterns may be related to usability problems, by grouping the verbalizations for the two physical *devices* and two *websites* in our analysis. We observed similar trends to that in Study 1 (see Table 3) in each verbalization category's frequency of occurrence, as shown in Table 9. Particularly, when considering physical devices and

Table 9. The Percentage of Audio Segments Labeled with Each Verbalization Category

Device or website	Verbalization category			
	Reading	Procedure	Observation	Explanation
Physical devices	28.8%	28.9%	36.0%	6.3%
Digital websites	23.3%	33.2%	37.2%	6.2%
All together	26.1%	31.1%	36.7%	6.1%

Table 10. Precision, Recall, and *F*-measure of Each Verbalization Category in Identifying Problems for Physical Devices vs. Digital Websites

Category	Precision		Recall		*F*-measure	
	Physical devices	Digital websites	Physical devices	Digital websites	Physical devices	Digital websites
Observation	0.36	0.40	0.61	0.57	0.45	0.47
Procedure	0.11	0.19	0.14	0.25	0.12	0.22
Reading	0.14	0.15	0.19	0.13	0.16	0.14
Explanation	0.17	0.19	0.05	0.05	0.08	0.07

websites separately, *Observation* was still the most frequently occurring category, whereas *Explanation* was the least.

However, one difference between physical devices and websites is that verbalizations for physical devices contained slightly more *Reading* than those for websites. This suggests that users engaged in more reading when using physical devices than websites, most likely because users referred to instruction manuals, which provided a set of steps for completing tasks, when using physical devices. When using websites, they had less to read and instead verbalized more often about their actions, which resulted in higher amount of the *Procedure* category.

We computed the number of problems that usability evaluators identified for physical devices and digital websites respectively. The average number of problems identified per evaluator for the physical devices was 10.9 ($SD = 5.8$). For digital websites, it was 13.1 ($SD = 8.1$). A repeated-measures ANOVA test with Bonferroni correction found no significant differences ($F(1, 15) = 2.53, p = .13, \eta_p^2 = .14$).

We further examined whether physical devices and digital websites affect how verbalization categories relate to the problems by computing the precision, recall, and *F*-measure of each verbalization category in identifying problems for the physical devices and the digital websites separately (Table 10). Results show a similar trend that the *Observation* category was the most relevant category to usability problems while the *Explanation* category was the least relevant category to usability problems. However, there was a difference between the *Procedure* category and the *Reading* category. Compared to the physical devices, the *Procedure* category was more relevant to usability problems than the *Reading* category for the digital websites.

We also computed the precision, recall, and *F*-measure of each verbalization pair in identifying usability problems for the physical devices and the digital websites separately (Table 11). Results show a similar trend that pairs of *Observation* and *Procedure* (P-O or O-P) and pairs of *Observation* and *Reading* (R-O or O-R) were most likely associated with problems. One difference is that pairs of *Observation* and *Reading* (R-O or O-R) were more relevant to problems for physical devices while pairs of *Observation* and *Procedure* (P-O or O-P) were more relevant to problems for digital websites.

Table 11. Precision, Recall, and *F*-measure of Each Verbalization Category Pair in Identifying Problems for Physical Devices vs. Digital Websites

Category pair	Precision		Recall		*F*-measure	
	Physical devices	Digital websites	Physical devices	Digital websites	Physical devices	Digital websites
P-O / O-P	0.09	0.15	0.35	0.50	0.15	0.23
R-O / O-R	0.11	0.12	0.43	0.25	0.18	0.16
O-E / E-O	0.18	0.14	0.09	0.06	0.12	0.08
P-R / R-P	0.03	0.08	0.08	0.12	0.05	0.09
P-E / E-P	0.04	0.15	0.02	0.05	0.03	0.08
R-E / E-R	0.08	0.10	0.03	0.02	0.05	0.03

Table 12. Precision, Recall, and *F*-measure of Each Verbalization Category in Identifying Problems When Evaluators Had Access to the Audio or Video Modality of the Think-Aloud Sessions

Category	Precision		Recall		*F*-measure	
	Audio	Video	Audio	Video	Audio	Video
Observation	0.42	0.34	0.57	0.62	0.48	0.44
Procedure	0.18	0.13	0.20	0.20	0.19	0.16
Reading	0.17	0.11	0.19	0.13	0.18	0.12
Explanation	0.18	0.18	0.05	0.05	0.07	0.08

4.3.5 Audio vs. Video. To analyze the effect of the *modality* that evaluators had for analyzing the think-aloud sessions, we first computed the number of problems that usability evaluators identified when they were given the audio recording only and when they were given the video recording as well. Results show that evaluators found on average 12.2 ($SD = 8.0$) problems when they had access to only the audio recording and 11.9 ($SD = 5.9$) problems when they had access to the video recording also. This difference was not statistically significant ($F(1, 15) = 0.05, p = .82, \eta_p^2 = .003$). It is, however, worth noting that the effect size was small.

We then conducted the same analysis to examine if the modality affects the verbalization categories and category pairs associated to the problems identified by the evaluators by computing the precision, recall, and *F*-measure of each verbalization category in identifying problems for the physical devices and the digital websites respectively. Results are shown in Table 12. The three measures of how each verbalization category relates to problems is consistent when the evaluators had access to the audio or video modality of the think-aloud sessions. Regardless of the modality, the *Observation* category was again the most relevant to the usability problems in terms of the three measures while the *Explanation* category was the least relevant.

Table 13 shows the measures of each verbalization category pair in identifying problems when evaluators had access to the audio or video modality of the think-aloud sessions. The general trend for each modality is consistent with pairs of *Observation* and *Procedure* (P-O or O-P) and pairs of *Observation* and *Reading* (R-O or O-R) were most likely associated with problems.

4.3.6 Visualization vs. Without Visualization. To analyze the effect of visualization, we grouped problems based on whether evaluators had access to the visualization and computed the number of problems that they identified. The results show that evaluators, on average, identified 12.7

Table 13. Precision, Recall, and *F*-measure of Each Verbalization Category Pair in Identifying Problems When Evaluators Had Access to the Audio or Video Modality of the Think-Aloud Sessions

Category pair	Precision		Recall		*F*-measure	
	Audio	Video	Audio	Video	Audio	Video
P-O / O-P	0.16	0.10	0.46	0.41	0.23	0.16
R-O / O-R	0.10	0.13	0.32	0.33	0.16	0.18
O-E / E-O	0.13	0.19	0.07	0.08	0.09	0.11
P-R / R-P	0.05	0.06	0.10	0.10	0.06	0.08
P-E / E-P	0.11	0.08	0.04	0.04	0.06	0.05
R-E / E-R	0.04	0.18	0.01	0.04	0.02	0.06

Table 14. Precision, Recall, and *F*-measure of Each Verbalization Category in Identified Problems When Evaluators Worked with or without Visualization

Category	Precision		Recall		*F*-measure	
	With	Without	With	Without	With	Without
Observation	0.41	0.34	0.59	0.60	0.48	0.43
Procedure	0.18	0.12	0.22	0.18	0.20	0.14
Reading	0.16	0.12	0.15	0.17	0.16	0.14
Explanation	0.20	0.16	0.05	0.05	0.07	0.08

Table 15. Precision, Recall, and *F*-measure of Each Verbalization Category Pair in Identified Problems When Evaluators Worked with or without Visualization

Category pair	Precision		Recall		*F*-measure	
	With	Without	With	Without	With	Without
P-O / O-P	0.14	0.09	0.46	0.39	0.22	0.16
R-O / O-R	0.13	0.10	0.31	0.37	0.18	0.16
O-E / E-O	0.17	0.14	0.06	0.09	0.09	0.11
P-R / R-P	0.07	0.04	0.10	0.10	0.08	0.06
P-E / E-P	0.10	0.07	0.04	0.04	0.05	0.05
R-E / E-R	0.14	0.04	0.03	0.01	0.05	0.02

($SD = 7.6$) problems with the visualization, and 11.4 ($SD = 6.0$) problems without the visualization. Although slightly more problems were identified when evaluators had access to the visualization, the difference was not statistically significant ($F(1, 15) = 1.42, p = .25, \eta_p^2 = .086$).

We computed the precision, recall, and *F*-measure to examine if the visualization affects how verbalization categories and category pairs relate to the identified problems. Results for verbalization categories and category pairs are shown in Tables 14 and 15, respectively. The general trend of the three measures is consistent when evaluators worked with or without visualizations.

4.3.7 What Users Talked About When They Encountered Problems? Besides examining the relationship between verbalization categories and the usability problems, we further calculated the most frequently used words that users verbalized when encountering problems and plotted the

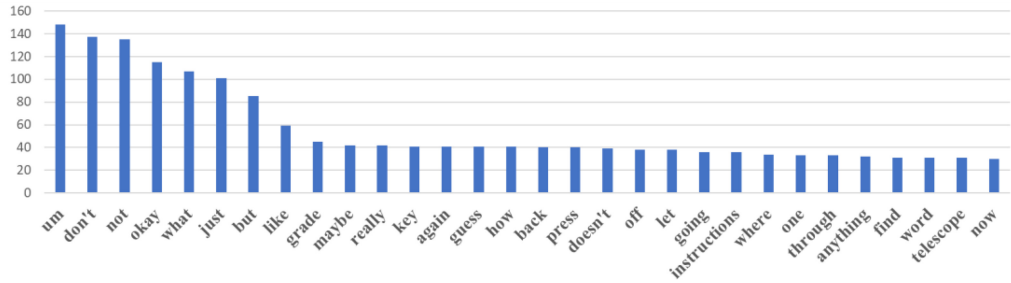

Fig. 10. Most frequently verbalized words when users encountered problems when thinking aloud.

top 30. Figure 10 shows the result. Stop words were removed from the analysis (e.g., pronouns, articles, common verbs such as be).

Based on the results, the most frequently verbalized words consisted of the following: (1) verbal fillers, such as *um*; (2) negations, such as *don't* and *not*; *doesn't*; (3) words expressing uncertainty, such as *maybe* and *guess;* (4) words signaling repetitive effort, such as *again* and *back;* (5) words used to raise questions, such as *what, how,* and *where;* (6) nouns related to the tasks or the test products, such as *grade, key,* and *telescope;* (7) verbs related to the tasks or the test products, such as *press,* and *find.*

We further analyzed the verbalizations that were associated with each problem to better understand the utility of these frequently occurring words as follows: (1) *how often did think-aloud users use verbal fillers (e.g., um)?* (2) *how often did think-aloud users use negation (e.g., not, don't, doesn't)?* (3) *how often did think-aloud users use uncertain words* (*e.g., maybe, guess*)? (4) *how often did think-aloud users use words suggesting repetition (e.g., again, back)?* (5) *how often did think-aloud users ask themselves questions (e.g., what, how)?* We note that *Okay* was not included as a verbal filler in our analysis, since it can also be used for confirmation.

The results show that out of the 385 problems, think-aloud users used the following: (1) *negation* words in 266 (69%) problems; (2) *filler* words in 148 (38%) problems; (3) words showing *uncertainty* in 95 (25%) problems; (4) words that *raised questions* in 94 (24%) problems (e.g., *"did I miss anything?"*); (5) words showing *repetitive effort* in 56 (15%) problems. It is worth mentioning that uncertainty was not always expressed through a single signaling word (e.g., *maybe, guess*). It was sometimes expressed through *their verbalized actions* (e.g., *"I'm just clicking some random links on this page"*). Furthermore, we also noticed that in 41 (11%) problems, users experienced *Aha! moments*, which were the moments when they suddenly came to an understanding of something that they had previously misunderstood or could not understand (e.g., "*oh, I thought they meant the power key.*"). This phenomenon was previously articulated by evaluators in Study 1 as well. Usability evaluators also identified five problems in which users articulated *suggestions* (e.g., *"it would be better if I could filter through them to choose grade"*).

4.3.8 How Users Verbalized Problems. We learned from the interviews in Study 1 that evaluators found all the verbalization and speech features useful as cues for identifying problems. Although the evaluators were mostly positive toward using the verbalization categories, silence, filler words, and sentiment for problem identification, their thoughts on speech rate, loudness, and pitch were mixed.

To understand whether and how the voice features were related to the usability problems by computing the same three measures (i.e., precision, recall, *F*-measure) for each feature. We considered a feature's value to be abnormal (i.e., high or low) if it was greater or less than two standard deviations away from the feature's average value in the whole audio recording. Figure 11 shows

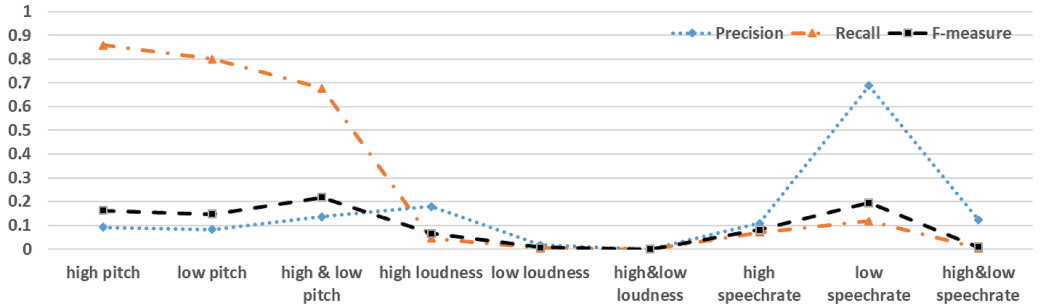

Fig. 11. Precision, recall, and *F*-measure of each speech feature in identifying problems.

the precision, recall, and *F*-measure of each speech feature in identifying problems. While high and low pitch has high recall values (i.e., 0.86 and 0.80 respectively), the low speech rate has a high precision value (0.70). The implication is that if evaluators examined all the verbalization segments with abnormal pitch values, they would have a high chance to locate a high percentage of all the usability problems due to the high recall values. If evaluators examined all the verbalization segments with a low speech rate value, they would have a high success rate in finding a usability problem due to the high precision value. However, at the same time, it is important to mention that there is no single voice feature that has both high precision and recall. The implication is that usability evaluators should not just rely on any single voice feature if they would like to identify as many usability problems as possible. These features should be used together with other features, such as the verbalization categories, sentiments, negations, filler words, and words for asking questions, expressing uncertainty, or signaling repetition.

4.4 Summary

In Study 2, we validated and extended the findings of the Study 1 with more testing products, a new pool of think-aloud participants and usability evaluators. We further investigated whether the relationship between verbalizations and usability problems found in Study 1 was generalizable to three factors as follows: (1) test products (i.e., physical objects vs. digital websites); (2) modality used to record the think-aloud sessions that evaluators were provided with (i.e., audio vs. video recording); (3) access to a visualization of the verbalizations. In general, the results suggest that the relationship between verbalizations and usability problems was not affected by the testing object, modality, or the presence of visualization.

In sum, the *Observation* category was the most useful category for identifying usability problems in terms of precision, recall, and *F*-measure. One plausible reason for this could be that by nature, audio segments are labeled with the *Observation* category when users express remarks about the test products. Since these remarks are likely to provide insight into the user's mental state and their current mood, such audio segments might help evaluators make judgments about whether users are experiencing problems.

Our analysis into the verbalizations and the speech features further indicated that segments that were labeled as the *Observation* category and had negative sentiment had a higher chance of revealing problems. Furthermore, these segments were even more likely associated with problems when users verbalized them with a high or low pitch.

On the other hand, evaluators also found problems in audio segments labeled with the *Procedure* category. Particularly, evaluators noted that users likely encountered problems when they were verbalizing repeated actions. This was evident by the frequently verbalized words used that signal repetitive attempts, such as "*back*" and "*again,*" when users encountered problems (Figure 10). For

example, "*let me just go back up to see if I missed anything*" and "*the device did not turn on, so I'll press it again.*" Procedure category with uncertain words, such as "*maybe*" or "*guess*" (Figure 10), can also be suggestive of problems (e.g., "*maybe I'll go to education programs*" and "*so I guess I'll just click the rest of the links that I haven't tried yet*"). Furthermore, our evaluators also noted that problems also seemed to occur when users engaged in long periods of reading. On the contrary, based on the recall and *F*-measure, the *Explanation* category was the least indicative of problems. One plausible reason for this could be that the proportion of segments labeled as the *Explanation* category was significantly lower than other categories.

5 DISCUSSION

5.1 Physical Devices vs. Digital Systems

Our findings show that the verbalization categories appeared in similar proportions for physical devices and digital systems. For example, regardless of the device, *Observation* was the most frequently occurring category while *Explanation* was the least. The frequency of the *Reading* and *Procedure* category were also very similar. Additionally, the verbalization patterns that cue to problems were also very similar. However, one subtle difference was that there was slightly more of the *Reading* and less of the *Procedure* category when using physical devices (Table 10), perhaps because users were not familiar with the devices and had a greater need to solicit help from instruction manuals that were readily available to them. In contrast, while navigating websites, users did not have a prescribed set of steps that they could refer to for completing tasks and thus less relied on reading from websites but more exploration.

5.2 Audio vs. Video Modality Available to Usability Evaluators

The analysis in Study 2 shows that usability evaluators identified roughly the same number of problems when they had access to either audio or video modality of think-aloud recordings. It is worth noting that the effect size was small and thus it is possible that there might be a difference if a large number of evaluators and think-aloud sessions were included in the analysis.

The verbalization patterns that are related to usability problems are similar when evaluators had access to either audio or video modality of think-aloud recordings. We further analyzed the problem descriptions that evaluators provided to understand the types of problems that they identified when having access to different modality of the think-aloud recordings. The result shows that evaluators found roughly the same types of usability problems when they had access to different modalities in our studies. During the interviews with the evaluators, many expressed that even when they did not have the video stream, the richness of the sounds in the audio stream helped evaluators imagine what users were experiencing ("*Yes, [without the video stream] I can't see their faces or their interactions with the interfaces. But I can still experience their emotions and struggles by listening to the audio."-ev15*). Evaluators also consistently agreed that the audio was useful not only because verbalized words provided insight into users' thought process and their feelings (based on the tone/pitch of their voice), but also because non-words uttered by users and noise from the surrounding environment provided valuable contextual information. For example, the sound of sighing could indicate that a user is frustrated. Frequent page flipping sounds could indicate that an instruction manual was poorly designed, subjecting users to constantly revisiting pages. Mechanical sounds generated by operations on devices, such as the clicking of a button, can help evaluators understand the fluency of a user's actions.

However, we also noticed that evaluators who had access to the video modality provided evidence from the visual channel that was unavailable from the audio channel. For example, while evaluators who had access to audio modality inferred the mismatch between instructions and the

actual device from what the users said for the universal remote control, evaluators who had access to the video modality pointed out that some labels on the device did not match with the instructions. Similarly, evaluators pointed out the issue of lack of images on the searching result page to illustrate the content presentation issue of the science and technology museum website.

Thus, this difference in the evidence provided to support their identified usability problems suggests that although evaluators found that having access to only audio stream in analyzing think-aloud sessions was sufficient most of the time, they still found the video stream to be useful. Being able to see a user's face could be helpful because facial expressions could reflect their mood. But because some users kept a neutral face throughout the entire think-aloud session, seeing the user's face was not always useful. How facial expressions could be used to identify usability problems remains to be explored. Moreover, evaluators pointed out that a user's body language could also signal problems. For example, one evaluator found that a user had a tendency to scratch his head whenever he encountered problems. However, it is unclear whether body language is a reliable and consistent cue to locate usability problems.

Evaluators also felt that there were particular times when having access to a video recording would have been useful. For example, a video stream would be valuable when users become silent, since evaluators sometimes had a hard time determining whether users were stuck or just waiting for something to happen. This finding suggests that it might be a good idea to draw attention to the video stream of a recording when the think-aloud users become silent or maybe slightly before they fall into silence. Additionally, evaluators pointed out that video could be important when think-aloud users verbalized their actions using demonstratives (i.e., this, that, these, those) or adverbs of place (e.g., here, there). For example, verbalizations, such as "I'm going to hold *this* button and *this* button" or "I'm clicking the link *here*," can be hard to understand without seeing what users are referring to. On the other hand, evaluators might not want to constantly monitor the video when having access to it ("*without video, I can concentrate better on listening. If needed, I'll look at the video*"-ev12). Thus, one interesting question would be to help evaluators figure out what users are implying when they verbalize vague statements and to highlight moments in the videos that should be given attention, to reduce the need to constantly monitor the video stream.

5.3 Visualization of Verbalization and Speech Features

Our tool provided visualizations of verbalization categories, six speech features, and the transcript to evaluators in the *Audio + Visualization* and *Video + Visualization* conditions, which was a novel feature that had not been explored in the literature. These two conditions were compared with two baseline conditions, which were the *Audio only* and the *Video only* conditions. The evaluators' feedback from the two studies suggested that all verbalization and speech features were useful for identifying usability problems. However, we also learned from Study 2 that the number of problems that evaluators identified with access to the visualizations was not significantly more than having no access to them, and the patterns in verbalizations relating to problems were also similar when evaluators had or did not have access to the visualizations. One possible reason might be that the way these features were presented in the visualization tool might have overwhelmed evaluators. This is evident in evaluators' feedback. Particularly, one evaluator reported that she almost completely ignored the visualizations because the interface was "too busy." This raises an interesting challenge for future exploration: *how to visualize the verbalization categories and speech features to maximize their utility to usability evaluators?* As the advancement of automatic speech recognition may make the automatic generation of transcript more accurate in the near future, it is also worth exploring *how to best leverage transcript information with verbalization categories and speech features together to facilitate the analysis of large amounts of recorded think-aloud sessions.*

Table 16. Verbalization Category Proportion

Studies	Verbalization category		
	Reading, procedure	Observation	Explanation
Our studies	56.3%	37.6%	5.9%
Classic instruction condition in Zhao et al. [50]	70.3%	20.1%	9.6%
Explicit instruction condition in Zhao et al. [50]	49.9%	33.8%	16.3%

5.4 Verbalization Category Proportions

When conducting think-aloud sessions, we followed Ericsson and Simon's three guidelines: use neutral instructions, allow participants to practice thinking aloud, and no probe or intervene during think-aloud sessions except to remind participants to keep talking if they fall into silence for a long time [11]. One study that examined users' verbalizations when following these guidelines was conducted by Zhao et al. [50]. In their study, authors analyzed users' verbalizations in think-aloud sessions that were conducted under the following two conditions: the *classic instruction* condition and the *explicit instruction* condition. The *classic instruction* condition strictly followed all three guidelines advocated by Ericsson and Simon. In contrast, the *explicit instruction* condition was the same as the *classic instruction* condition, except that it included an explicit instruction requesting participants to report both the explanations and verbalizations that are relevant to understanding the user experience.

In their study, users' verbalizations were categorized into the following five categories: *procedural description, positive experience, negative experience, expectation*, and *explanation*. Based on the definitions of these categories, the relationship between these categories and the four categories that were used in our study is as follows: *procedural description* is equivalent to the combination of the *Reading* and the *Procedure* categories; *positive experience, negative experience* and *expectation* together are equivalent to the *Observation* category; *explanation* was equivalent to the *Explanation* category. As a result, we combined the *Reading* and the *Procedure* categories into one category and computed the average proportion of the verbalization categories in our two studies. Table 16 shows the result.

Based on the result, it is evident from the result in Table 16 that verbalizations of the *Observation* and the *Explanation* categories exist are present even when following the guidelines proposed by Ericsson and Simon's guidelines. In other words, users do verbalize their comments, feelings and rationales (labeled as the *Observation* and the *Explanation* categories) even when users were not explicitly instructed to do so. Second, both our studies and the two conditions in Zhao et al.'s study found that the majority of the verbalizations fall into sequences of the *Reading* and the *Procedure* categories. Third, both our studies and the explicit instruction condition had less amount of the *Reading* and the *Procedure* categories compared to the classic instruction condition. For the explicit instruction condition, this was because the explicit instruction was given, which was evident from the significantly higher number of occurrences of the *Explanation* category. We reflected on how we conducted think-aloud sessions and how the process has differed from that of Ericsson and Simon to explain this marked increased in the amount of the *Observation* category (and similarly, the decreased in the amount of the *Reading* and the *Procedure* categories). We found that although we followed the three guidelines proposed by Ericsson and Simon [11], we also showed our participants a 1-minute demo video of a think-aloud session being carried out by an actor, which is offered online by Nielsen and Norman group [31]. In this 1-minute demo video, the participant verbalized her comments and feelings about a test website in addition to describing her actions.

Table 17. The Average Any-Two Agreement between Evaluators

All devices and websites together	Coffee machine	Remote	History museum	Science and technology museum
0.80	0.76	0.88	0.82	0.76

This demo video might have implicitly influenced our participants to verbalize their comments and feelings, in their attempts to mimic the actor in the demo.

5.5 Evaluator Effect

Previous studies reported that evaluators might find different sets of usability problems, even when they analyze the same usability test sessions (e.g., [18, 19]). We further analyzed Study 2's data to see if evaluators who analyzed the same think-aloud session would agree on the verbalization segments that were linked to problems. To measure the agreement between two evaluators, we computed the *any-two agreement* measure using the following equation: $\frac{P_i \cap P_j}{P_i \cup P_j}$ (P_i and P_j are the sets of problems identified by tow evaluator i and j) [18] for each think-aloud session that was evaluated by two evaluators. We then computed the average any-two agreement for all test products (the first column in Table 17) and that for each test product separately (the second to the last columns in Table 17).

The values of the average any-two agreement measure in Table 17 show that our evaluators had reasonably high agreement. The patterns between think-aloud verbalizations and the usability problems that we identified in this research were based on the analysis of the joint problems identified by participants ($P_i \cup P_j$). The relatively high agreement between our evaluators suggests that the identified patterns would be largely applicable to each individual evaluator although they might disagree if a verbalization segment indicates a problem sometimes.

The average any-two agreement in our study was similar to the average any-two agreement reported in other studies (e.g., 0.71 in [50]) and was higher compared to other studies in the literature. For example, the average any-two agreement was 31% for moderated sessions, wherein a moderator presented and probed the user, and was 30% for unmoderated sessions, wherein no moderator was present [19]. Many factors could contribute to the differences. One factor was the amount of time that was allocated to evaluators to analyze the sessions. Evaluators in our study were allocated 1.5 times the length of a think-aloud session to spend on their analysis. In contrast, the evaluators in Hertzum et al.'s study [19] spent on average 22 hours to analyze the sessions, which were on average 33 minutes. Therefore, in our study, evaluators may be more likely to focus on the more significant issues, leading to a higher agreement between evaluators. In fact, the evaluators in Hertzum et al.'s study [19] also had a much higher any-two agreement for critical problems (53% for moderated sessions, and 69% for unmoderated sessions), which was more closer to our measures. However, there are other factors that may have resulted in the difference between the any-two agreement measures. For example, our study used four test products, which consisted of two physical devices and two digital websites, while their test product was one digital website. Further, in our study, two evaluators examined each think-aloud session while their evaluation had nine or ten. The background and experience of think-aloud participants and evaluators may have also contributed to the differences.

6 LIMITATIONS

We used different sets of test products, different pools of think-aloud participants, and different sets of usability evaluators for the two studies to evaluate the validity and generalizability of the

findings. The number of test products, however, is still relatively small compared to the ever-growing number of products and websites that are available. It would be valuable to replicate the research with different products and websites to further examine the findings. Although we included UX professionals who were working in industry as evaluators in both studies ($N = 2$ for Study 1, $N = 5$ for Study 2), many of the evaluators were graduate students majoring in UX. Thus, the overall experience of our evaluators in evaluating data from usability tests is relatively less compared to usability evaluators who have worked in industry for years. It would be valuable to examine whether and how the years of experience in conducting think-aloud tests might affect the findings of this research.

The facilitators in the two studies informed the usability evaluators about the products that the think-aloud users used before they started to evaluate the recordings. Specifically, the facilitator showed the test product, described its main functions, and the tasks that the think-aloud users worked on. Although this introduction provided information about the products that they would evaluate in think-aloud sessions, we did not provide a chance for our evaluators to use the test products. We designed the studies in such a way so that our evaluators could identify usability problems that the think-aloud participants experienced in the recordings without being primed by their own experience of using the products. In practice, usability evaluators may have access to the test products and previous research suggests that double experts with knowledge in both usability evaluation and the specific domain might yield better insights [30]. Thus, it would be interesting to further explore whether having usability evaluators use the test products prior to evaluating think-aloud sessions would have any effect on the finding of this research.

In our two studies, we requested usability evaluators to identify problems that users were experiencing. Therefore, the findings of this research reveal the verbalization patterns that are likely associated with usability problems in general. We did not, however, request evaluators to rate the severity of the problems primarily because the amount of workload was already considered to be high for the allocated study time. As certain verbalization patterns might not only cue usability problems but also might appear more often when users are experiencing more severe problems. Thus, in future, it would be worth exploring whether there are correlations between verbalization patterns and a usability problem's severity.

We followed Ericsson and Simon's guidelines when conducting think-aloud sessions and did not probe or intervene during the sessions except to remind users to keep talking if they fall into silence for a long time [11]. In practical settings, however, usability evaluators do not always conform to these guidelines (e.g., [32, 42]) and may instead employ alternative protocols (e.g., relaxed think-aloud [17], speech-communication [39]), although these alternatives have received mixed results regarding their impact on task performance and the user's ability to make verbalizations [17, 47]. When using these alternative protocols, practices may vary in terms of the instructions, intervention, and prompts as no universal guidelines exist for conducting these thinking aloud protocols. Only recently have some researchers started to study the verbalizations in relaxed think-aloud sessions [16], but how these verbalizations relate to usability problems remains largely unknown. Thus, it is interesting to examine whether and how intervention during the think-aloud sessions affects the findings of this research.

All protocols discussed so far are variations of concurrent think-aloud protocols. Another type of protocol is the retrospective think-aloud protocol. When using retrospective think-aloud protocol, participants verbalize their thought process after they complete a task. Although the reported verbalizations rely on participants' memory and may suffer from post-task rationalization [22], this protocol does have one advantage, in that verbalizations do not have a direct interference with participants' thought processes during tasks. It is worth exploring how verbalization patterns in the

retrospective think-aloud protocol suggest problems. For example, would the *Observation category* still be most associated with problems? Would *Explanation* still be the least popular category?

Another direction is to look at the combination of concurrent and retrospective think-aloud protocols, referred to as the hybrid protocol [1]. Recent research shows that when using a hybrid protocol, the interpretations given after completing the concurrent think-aloud task helped to identify more problems [12] and provide insights into reasons for difficulties that participants encountered during concurrent think-aloud [29]. Thus, we conjecture that there would be more verbalizations labeled as the *Explanation* category in retrospective think-aloud than in concurrent think-aloud, leading to difference in the categories and category pairs that are more likely associated with problems. A controlled experiment that compares the verbalizations patterns in concurrent think-aloud and the combined think-aloud method is needed to ascertain this.

Lastly, the think-aloud sessions in our two studies were conducted with young adults. Recent research has suggested that age might have an influence on think-aloud usability testing [34, 44]. One interesting research question is to study how the findings of this research may differ for other age groups, such as older adults.

7 FUTURE WORK

7.1 Automatic Detection of Verbalization Categories and Usability Problems

Although we developed a tool to help us reduce some of the workload, the task of manually segmenting and categorizing the think-aloud verbalizations was still time-consuming and laborious. Future work should study how to automate or semi-automate the verbalization categorization process. This process may require building machine learning models to distinguish the semantic differences among the four categories. Another approach could be to learn from human demonstrations, although this would involve determining how to leverage small amounts of human data effectively, since collecting a large dataset would be difficult given how time-consuming it is to conduct and process large amounts of think-aloud sessions. Furthermore, since it is now known that certain verbalization and speech patterns tend to occur when users experience problems, it is possible to leverage the patterns to design systems that automatically detect when in a recorded think-aloud session users experience problems. Such information could then be used to draw usability evaluator's attention to parts of the session that are more likely to reveal problems.

7.2 Visualization of Verbalization Patterns to Suggest Problems to Usability Evaluators

We have discovered and validated patterns in verbalizations that are related to usability problems. Revealing these patterns to usability evaluators could potentially improve their productivity. However, how to present these patterns to evaluators so that they can identify usability problems easier or more efficiently might be challenging. As our studies have demonstrated, visualizing verbalization categories and speech features at the same time overwhelmed the evaluators, leaving them unsure of which ones to focus on. Future research should investigate how to effectively present cues to usability evaluators so that they could benefit from the verbalizations patterns that were identified through this research and at the same time do not feel overwhelmed.

7.3 Identification and Evaluation of Visual Cues for Locating Usability Problems

Usability evaluators in our studies pointed out several important cues in the video stream of think-aloud recordings that could be useful for locating usability problems, such as facial expressions and body gestures. However, these cues, as they commented, may vary across people. It is important to study whether these visual cues (e.g., facial expression or body gestures) are reliable source of information for identifying problem in the future.

When analyzing verbalizations, there are certain moments when video streams could potentially enrich usability evaluators' understanding of the context. These moments include when think-aloud users fall into silence or use certain words, such as demonstratives or adverbs of place. It is worth investigating where and how in the video stream should evaluators' attention be drawn to.

7.4 Analysis of Multiple Think-Aloud Sessions Conducted by Different Users

Analyzing multiple users' think-aloud sessions enables usability evaluators to detect common problems encountered by different users. Problems encountered by more users should be potentially given higher priority. Thus, an important next step is to extend the tool presented to allow evaluators to simultaneously visualize, analyze, and compare think-aloud sessions to identify common problems encountered by users.

8 CONCLUSION

Through the two studies, we systematically studied the relationship between verbalizations, speech features, and usability problems in concurrent think-aloud sessions. Our findings show that certain patterns of verbalization and speech features act as telltale signs of usability problems. Segments labeled as the *Observation* category were most likely associated with usability problems. Segments labeled as the *Procedure* category that also contain description of repeated actions were likely associated with usability problems. Segments labeled as the *Reading* category that last for a long period of time were also likely associated with usability problems. On the contrary, segments labeled as the *Explanation* category were relatively rare and did not have a clear relationship with usability problems. Our findings further show that evaluators often identified problems using combinations of verbalization categories since category combinations were helpful in providing contextual information as to why users were encountering problems. Furthermore, pairs of verbalization categories that contained the *Observation* category were generally more likely associated with problems than those without the *Observation* category.

Our analysis shows that the *F*-measure of using the *Observation* category to locate usability problems was around 0.5. To increase the chance of locating a problem, *sentiment* and *speech features* should be considered in conjunction with the category information. For example, when experiencing problems, users tended to use *negations, verbal fillers*, words indicating *uncertainty, repetitions*, or *questions*. Therefore, the sentiment of these verbalizations was often *negative*. Furthermore, users tended to verbalize their thought units in *high or low pitch* or with *low speech rate* but rarely changed the loudness of their voices when experiencing problems.

Our research demonstrates that these findings are largely generalizable to three factors as follows: the types of test products (i.e., physical devices vs. digital systems), the modality used to record the think-aloud sessions that evaluators were provided with (i.e., audio vs. video recording), and access to a visualization of the verbalizations. The implication is that the same set of verbalization patterns can be used to identify problems that users were experiencing when thinking aloud regardless of whether a physical device or a digital system was used. Usability evaluators can rely on verbalization and speech features alone to identify problems by and large, although certain cues in video streams have additive values to their analysis, such as facial expression and body language. However, whether these visual cues are consistent across users for locating problems remains to be examined. Moreover, the video stream of a think-aloud session can be informative when the think-aloud user remains silent or frequently uses demonstratives (e.g., this, that) or adverbs of place (e.g., here, there), which makes it difficult to infer what the user is referring to from the audio stream alone. As a result, in such situations, it would be preferable to draw evaluators' attention to the video stream. Our research also reveals that visualizations of verbalizations as provided in our studies did not affect the number of problems identified or the verbalization

patterns that were associated with problems. Although each individual verbalization feature was informative, presenting them all at once in addition to the audio or video recording of a think-aloud session, as evidenced in our studies, was overwhelming. Future work should explore better ways to visualize verbalization patterns to facilitate usability problems identification.

ACKNOWLEDGMENTS

We thank all our reviewers for their insightful and constructive reviews, which have helped us revise the manuscript. This research was supported by NSERC.

REFERENCES

[1] Obead Alhadreti and Pam Mayhew. 2018. Rethinking thinking aloud. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–12. DOI: https://doi.org/10.1145/3173574.3173618

[2] Obead Alhadreti, Pam Mayhew, and Senior Lecturer. 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 http://www.upassoc.org.

[3] Pamela M. Auble, Jeffery J. Franks, Salvatore A. Soraci, Salvatore A. Soraci, and Salvatore A. Soraci. 1979. Effort toward comprehension: Elaboration or "aha"? *Mem. Cognit.* 7, 6 (1979), 426–434. DOI: https://doi.org/10.3758/BF03198259

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

[5] Elizabeth Charters. 2003. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Educ. J.* 12, 2 (2003), 68–82. DOI: https://doi.org/10.26522/brocked.v12i2.38

[6] Michelene T. H. Chi, Nicholas De Leeuw, Mei Hung Chiu, and Christian Lavancher. 1994. Eliciting self-explanations improves understanding. *Cogn. Sci.* 18, 3 (1994), 439–477. DOI: https://doi.org/10.1016/0364-0213(94)90016-7

[7] Jason M. Chin and Jonathan W. Schooler. 2008. Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing. *Eur. J. Cogn. Psychol.* 20, 3 (2008), 396–413. DOI: https://doi.org/10.1080/09541440701728623

[8] H. H. Clark and J. E. Fox Tree. 2002. Using "uh" and "um" in spontaneous speaking. *Cognition* 84, 1 (2002), 73–111. DOI: http://dx.doi.org/10.1016/S0010–0277(02)0.

[9] 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

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

[11] K. Anders Ericsson and Herbert A. Simon. 1984. *Protocol Analisys: Verbal Reports as Data.* MIT Press.

[12] Asbjørn Følstad. 2007. Work-domain experts as evaluators: Usability inspection of domain-specific work-support systems. *Int. J. Hum. Comput. Interact.* 22, 3 (2007), 217–245. DOI: https://doi.org/10.1080/10447310709336963

[13] Mark C. Fox, K. Anders Ericsson, and Ryan Best. 2011. Do procedures for verbal reporting of thinking have to be reactive? *Psychol. Bull.* 137, 2 (2011), 316.

[14] Amy M. Gill and Blair Nonnecke. 2012. Think aloud. In *Proceedings of the 30th ACM International Conference on Design of Communication (SIGDOC'12)*. 31–36. DOI: https://doi.org/10.1145/2379057.2379065

[15] F. Goldman-Eisler. 1986. *Cycle Linguistics: Experiments in Spontaneous Speech.* Double Day, New York, NJ.

[16] 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

[17] 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

[18] Morten Hertzum and Niels Ebbe Jacobsen. 2009. International journal of human-computer interaction the evaluator effect: A chilling fact about usability evaluation methods the evaluator effect: A chilling fact about usability evaluation methods. *Int. J. Hum. Comput. Interact.* 13, 4 (2009), 421–443. DOI: https://doi.org/10.1207/S15327590IJHC1304_05org/10.1207/S15327590IJHC1304_05

[19] Morten Hertzum, Rolf Molich, and Niels Ebbe Jacobsen. 2014. What you get is what you see: Revisiting the evaluator effect in usability tests. *Behav. Inf. Technol.* 33, 2 (2014), 143–161. DOI: https://doi.org/10.1080/0144929X.2013.783114

[20] Masahiro Hori, Yasunori Kihara, and Takashi Kato. 2011. Investigation of indirect oral operation method for think aloud usability testing. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, 38–46. DOI: https://doi.org/10.1007/978-3-642-21753-1_5

[21] Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. 216–225.

[22] M. D. T. de Jong, P. J. Schellens, and M. J. Van den Haak. 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. Retrieved from https://academic.oup.com/iwc/article-abstract/16/6/1153/769631.

[23] Claire-Marie Karat, Robert Campbell, and Tarra Fiegel. 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'92)*. 397–404. DOI: https://doi.org/10.1145/142750.142873

[24] Emiel Krahmer and Nicole Ummelen. 2004. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Trans. Prof. Commun.* 47, 2 (2004), 105–117. DOI: https://doi.org/10.1109/TPC.2004.828205

[25] Petri Laukka, Clas Linnman, Fredrik Åhs, Anna Pissiota, Örjan Frans, Vanda Faria, Åsa Michelgård, Lieuwe Appel, Mats Fredrikson, and Tomas Furmark. 2008. In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *J. Nonverbal Behav.* 32, 4 (2008), 195–214. DOI: https://doi.org/10.1007/s10919-008-0055-9

[26] Darryn Lavery, Gilbert Cockton, and Malcolm P. Atkinson. 1997. Comparison of evaluation methods using structured usability problem reports. *Behav. Inf. Technol.* 16, 4–5 (1997), 246–266. DOI: https://doi.org/10.1080/014492997119824

[27] S. Mcdonald, T. Zhao, and H. M. Edwards. 2016. Look who's talking: Evaluating the utility of interventions during an interactive think-aloud. *Interact. Comput.* 28, 3 (2016), 387–403. DOI: https://doi.org/10.1093/iwc/iwv014

[28] Sharon McDonald and Helen Petrie. 2013. The effect of global instructions on think-aloud testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. 2941–2944. DOI: https://doi.org/10.1145/2470654.2481407

[29] 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

[30] Jakob Nielsen. 1993. *Usability Engineering*. Elsevier. DOI: https://doi.org/10.1145/1508044.1508050

[31] Jakob Nielson. 2014. Demonstrate thinking aloud by showing users a video. In *Evidence-Based User Experience Research, Training, and Consulting*. Nielson Norman Group. Retrieved October 2, 2018 from https://www.nngroup.com/articles/thinking-aloud-demo-video/.

[32] Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? In *Proceedings of the 6th ACM conference on Designing Interactive Systems (DIS'06)*. 209–218. DOI: https://doi.org/10.1145/1142405.1142439

[33] K. R. Ohnemus and D. 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/10.1177/107118137902300152

[34] Erica Olmsted-Hawala and Jennifer Romano Bergstrom. 2012. Think-aloud protocols: Does age make a difference? In *Proceedings of the STC Technical Communication Summit*.

[35] Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. Think-aloud protocols analyzing three different think-aloud protocols with counts of verbalized frustrations in a usability study of an information-rich web site think-aloud protocols alterna.pdf. In *Proceedings of the 2010 IEEE International Professional Communication Conference (IPCC'10)*. 60–66.

[36] Alex Pentland. 2009. *Honest Signals: How They Shape Our World*. MIT Press. DOI: https://doi.org/10.1145/2072298.2072374

[37] Lloyd R. Peterson. 1969. Concurrent verbal activity. *Psychol. Rev.* 76, 4 (1969), 376–386. DOI: https://doi.org/10.1037/h0027443

[38] Matthew F. Pike, Horia A. Maior, Martin Porcheron, Sarah C. Sharples, and Max L. Wilson. 2014. Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14)*. 3807–3816. DOI: https://doi.org/10.1145/2556288.2556974

[39] J. Ramey and T. Boren. 2000. Thinking aloud: Reconciling theory and practice. *IEEE Trans. Prof. Commun.* 43, 3 (2000), 261.

[40] Detlef Rhenius and Gerhard Deffner. 1990. Evaluation of concurrent thinking aloud using eye-tracking data. *Proc. Hum. Factors Soc. Annu. Meet.* 34, 17 (1990), 1265–1269. DOI: https://doi.org/10.1177/154193129003401719

[41] Sharon McDonald, Helen M. Edwards, and Zhao Tingting. 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

[42] Qingxin Shi. 2008. A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction Building Bridges (NordiCHI'08)*. 344–352. DOI: https://doi.org/10.1145/1463160.1463198

[43] Glen Shires and Hans Wennborg. 2012. *Web Speech API Specification*. Speech API Community Group, W3C. DOI: https://doi.org/10.1021/jf001495e

[44] Andreas Sonderegger, Sven Schmutz, and Juergen Sauer. 2016. The influence of age in usability testing. *Appl. Ergon.* 52 (2016), 291–300. DOI: https://doi.org/10.1016/j.apergo.2015.06.012

[45] Siegfried L. Sporer and Barbara Schwandt. 2006. Paraverbal indicators of deception: A meta-analytic synthesis. *J. Appl. Cogn. Psychol.* 20, 4 (2006), 421–446. DOI: https://doi.org/10.1002/acp.1190
[46] Howard Tamler. 1998. How (much) to intervene in a usability testing session. *Common Gr.* 8, 3 (1998), 11–15.
[47] S. Tirkkonen-Condit. 2006. Think-aloud protocols. In *Encyclopedia of Language and Linguistics*. Elsevier, 678–686. DOI: https://doi.org/10.1016/B0-08-044854-2/00479-X
[48] R. B. Wright and S. A. Converse. 1992. Method bias and concurrent verbal protocol in software usability testing. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 36, 16 (1992), 1220–1224. DOI: https://doi.org/10.1177/154193129203601608
[49] Zhao Tingting and McDonald Sharon. 2010. Keep talking. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries (NordiCHI'10)*. 581–590. DOI: https://doi.org/10.1145/1868914.1868979
[50] Zhao Tingting, McDonald Sharon, and Helen M. Edwards. 2014. The impact of two different think-aloud instructions in a usability test: A case of just following orders? *Behav. Inf. Technol.* 33, 2 (2014), 162–182. DOI: https://doi.org/10.1080/0144929X.2012.708786
[51] Praat: Doing Phonetics by Computer. Retrieved from http://www.fon.hum.uva.nl/praat/.

Received May 2018; revised March 2019; accepted April 2019