Journal of Usability Studies
Vol. 15, Issue 2, February 2020, pp. 85–102

Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey

Mingming Fan
Assistant Professor
School of Information
Rochester Institute of Technology
Rochester, NY, 14623
USA
mxfics@rit.edu

Serina Shi
Master's graduate
Faculty of Information
University of Toronto
140 St. George ST.
Toronto, ON, M5S3G6
Canada
serina.shi@mail.utoronto.ca

Khai N. Truong
Professor
Department of Computer Science
40 St. George ST, Rm 7268
University of Toronto
Toronto, ON, M5S2E4
Canada
khai@cs.toronto.edu

Abstract

Think-aloud protocols are one of the classic methods often taught in universities for training UX designers and researchers. Although previous research reported how these protocols were used in industry, the findings were typically based on the practices of a small number of professionals in specific geographic regions or on studies conducted years ago. As UX practices continuously evolve to address new challenges emerging in industry, it is important to understand the challenges faced by current UX practitioners around the world when using think-aloud protocols. Such an understanding is beneficial for UX professionals to reflect on and learn from the UX community's practices. It is also invaluable for academic researchers and educators to understand the challenges faced by professionals when carrying out the protocols in a wide range of practical contexts and to better explore methods to address these challenges. We conducted an international survey study with UX professionals in various sized companies around the world. We found that think-aloud protocols are widely and almost equally used in controlled lab studies and remote usability testing; concurrent protocols are more popular than retrospective protocols. Most UX practitioners probe participants during test sessions, explicitly request them to verbalize particular types of content, and do not administer practice sessions. The findings also offer insights on practices and challenges in analyzing think-aloud sessions. In sum, UX practitioners often deal with the tension between validity and efficiency in their analysis and demand better fast-paced and reliable analysis methods than merely reviewing observation notes or session recordings.

Keywords

Think-aloud protocols, usability test, user experience, industry practices and challenges, international survey

Copyright © 2019–2020, User Experience Professionals Association and the authors. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. URL: http://uxpajournal.org/.

Introduction

Think-aloud protocols, in which participants verbalize their thoughts when performing tasks, are used in usability testing to elicit insights into their thought processes that are hard to obtain from mere observation. Think-aloud protocols are often taught in UX courses to train professionals (Dumas & Redish, 1999; Nielsen, 1993; Preece, Rogers, & Sharp, 2015; Rubin & Chisnell, 2008) and are considered as the "gold standard" for usability evaluation (Hornbæk, 2010). Boren and Ramey probably were the first to note the discrepancies between the theory introduced by Ericsson and Simon (1984) and the practice of using think-aloud protocols in the UX field (Boren & Ramey, 2000). The discrepancies, however, were identified by their field observations and by reviewing usability guidebooks and literature. Therefore, there was a lack of empirical reports on how the protocols are used in industry.

Previous research has examined the practices of using think-aloud protocols in local geographic regions. For example, Nørgaard and Hornbæk studied a small number of UX practitioners' practices in Danish enterprises and offered insights on how they conducted and analyzed think-aloud sessions (2006). Similarly, Shi reported practices of and particular challenges in using think-aloud protocols (2008). In contrast, McDonald, Edwards, and Zhao conducted an international survey study to understand how think-aloud protocols were used in a broader scale and distributed the survey to UX professional and academic listservs (2012). However, as the survey was conducted in 2011 and new UX testing software and tools have emerged over this period, the extent to which think-aloud protocols are currently being used in industry is unclear. Moreover, recent research has also urged the community to learn more about the current UX practices in industry (MacDonald & Atwood, 2013).

To better understand how think-aloud protocols are currently used in industry, we designed and conducted a survey study with UX practitioners who had different levels of experience and worked in different industries around the world. In this paper, we present and discuss the key findings and implications of the survey study to inform UX practitioners and researchers about the practices and challenges surrounding the use of think-aloud protocols in industry.

Methods

The goal of this study was to understand how think-aloud protocols are being used by UX professionals in different fields around the world. We chose survey over other methods (e.g., interview, focus groups) because it allowed us to gather data from a broad range of UX practitioners located in different geographic regions who work in different industrial fields.

Respondents

We contacted the organizers of local chapters of the User Experience Professional Association (UXPA), the largest organization of UX professionals around the world, to promote the survey study. We received support from the organizers of the UXPA's local chapters in Asia, Europe, and North America, who helped us distribute the survey link to their listservs. We also promoted the survey link in UX professionals-related LinkedIn groups and other social media platforms. Thus, the members of these UXPA local chapters and the LinkedIn groups were our potential samples. We conducted the survey study for about three months—July-September in 2018. The inclusion criterion was that respondents must work in industry as a UX practitioner.

Survey Design

The survey was conducted as an online questionnaire using Google Form. The survey contained a list of multiple-choice (required) and short-answer (optional) questions to understand whether and how UX professionals are currently using think-aloud protocols in addition to their basic profile information (i.e., the organization and/or the usability testing team that they work in and their current positions). No personally identifiable information was collected.

We were inspired by the previous survey study conducted in 2010 by McDonald et al. (2012) but at the same time made important changes. The previous survey was distributed to UX practitioners working in both academia and the industry, which made it hard to isolate the use and practical impact of think-aloud protocols in industry. Instead, our survey study was focused on the practices around the use of think-aloud protocols in industry and thus was only distributed to UX practitioners in industry. We also collected the respondents' years of

experience as a UX professional, which allowed us to understand the effect of the years of experience on their usage patterns. Furthermore, as new tools and procedures for conducting usability test sessions have entered the market since 2010, such as the Agile-UX design (Jurca, Hellmann, & Maurer, 2014), we wanted to understand how the use of think-aloud protocols has evolved in light of the introduction of new practices.

Data Analysis

Answers to multiple-choice questions are quantitative data and were analyzed to identify the statistical trends in using think-aloud protocols. Answers to short-answer questions are qualitative data. Two researchers first independently analyzed the qualitative data using open coding and then discussed to resolve any conflicts. They then used affinity diagramming to identify common themes that emerged from the data.

Results

We received valid responses from 197 UX practitioners in industry around the world. Next, we reported the aggregated information about the respondents' profile information and their practices of conducting and analyzing think-aloud usability tests.

Respondents' Profile

Work role: We asked respondents about their current job titles and allowed them to report more than one title if applicable. The majority of the respondents reported their current job title as UX researcher (54%) or UX designer (36%). Others identified their job title as UX team lead (11%), UX manager (8%), or design strategist (6%).

Location: In terms of the geographic locations, the majority of the respondents worked in North America 63.5% (n = 125), followed by Asia 19.3% (n = 38) and Europe 14.7% (n = 29). Other respondents worked in Australia 1.5% (n = 3), Africa 0.5% (n = 1), and South America 0.5% (n = 1).

Companies or organizations: The respondents worked in various sized companies or organizations (see results in Table 1). Furthermore, 81 respondents also reported the actual companies that they worked in. These companies covered a wide range of industrial fields, including ads and marketing, banking, gaming, health care, IT and software, professional services, supply chain, telecommunication, and UX consulting.

Table 1. Number of Employees in the Companies/Organizations that Respondents Worked In

Self-employed	< 100	100–999	1,000–9,999	> = 10,000
6.1% (n = 12)	15.2% (n = 30)	21.3% (n = 42)	21.3% (n = 42)	36.1% (n = 71)

UX team size: We asked respondents about the size of the UX team that they worked in and found that they worked in different sized UX teams: 1 (n = 21), 2–5 (n = 55), 6–10 (n = 42), 11–15 (n = 22), 16–20 (n = 16), 21–30 (n = 16), 31–50 (n = 5), and >50 (n = 20).

Experience: We asked respondents about the number of years that they had worked in HCI/UX/usability testing fields (see results in Table 2). The distribution of the years of experience in industry covered all ranges among the respondents.

Table 2. Number of Years that Respondents Had Spent in HCI/UX/Usability Testing Fields

< 1 year	1–2 years	3–5 years	6–9 years	> = 10 years
12.7% (n = 25)	20.3% (n = 40)	22.8% (n = 45)	16.8% (n = 33)	27.4% (n = 54)

Methods for detecting usability problems: We asked respondents about their three most frequently used methods for detecting usability problems (see results in Figure 1). The most frequently used methods for detecting usability problems among the respondents were as follows: usability testing (n = 168, 86%), interview (n = 118, 60%), heuristic evaluation

(n = 81, 41%), field studies/observation (n = 66, 34%), A/B testing (n = 53, 27%), cognitive walkthrough (n = 46, 23%), card sorting (n = 26, 13%), and focus groups (n = 25, 13%).

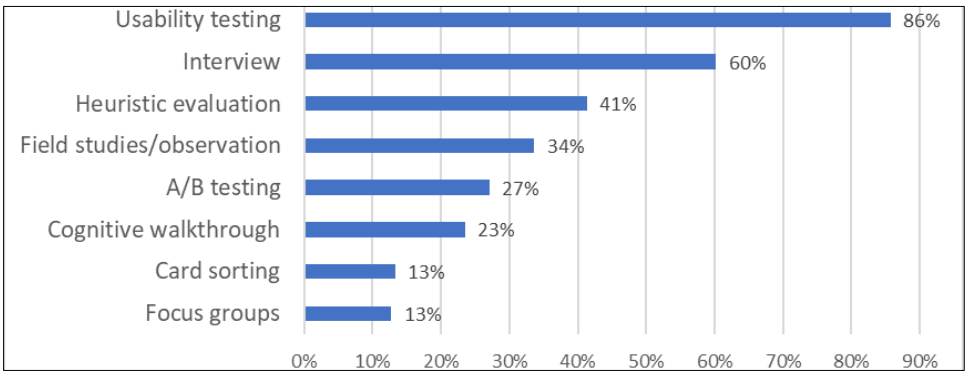


Figure 1. The frequently used methods for detecting usability problems among our respondents.

General Use of Think-Aloud Protocols

Where respondents learned think-aloud protocols: Among the 197 respondents, 91% (n = 179) reported that they had learned think-aloud protocols, and the remaining 9% (n = 18) reported that they were unfamiliar with think-aloud protocols. For the 179 respondents who had learned think-aloud protocols, 49% of them (n = 87) reported that they had learned the protocols in university/college, 36% (n = 65) at work, and 15% (n = 27) from UX online/offline bootcamps.

General use and non-use of think-aloud protocols: When conducting usability tests, 86% of all respondents (n = 169) reported that they used think-aloud protocols. In other words, 95% of the respondents who had learned think-aloud protocols (169 out of 179) used them. We carried out the following analysis based on the responses of these 169 respondents who used think-aloud protocols because the remaining survey questions were about how UX practitioners used think-aloud protocols.

We also asked those respondents who had learned think-aloud protocols but did not use them (n = 10) about their reasons for not using the protocols as an optional short-answer question and received seven responses. The reasons were as follows: conducting think-aloud sessions is not part of their role (n = 2), their study subjects may not verbalize their thoughts easily (e.g., children) or unbiasedly (e.g., internal users; n = 2), conducting think-aloud sessions takes too much time (n = 1), think-aloud protocols may distract their users (n = 1), and there are alternative methods (n = 1).

The frequency of using concurrent and retrospective think-aloud protocols: *Concurrent think-aloud protocols*, in which users verbalize their thoughts while working on tasks, and *retrospective think-aloud protocols*, in which users verbalize their thoughts only after they have completed the tasks (usually via watching their session recordings) are the two types of protocols. We asked respondents about their frequency of using concurrent and retrospective think-aloud protocols (see results in Figure 2). Specifically, 61% of them (n = 103) used the concurrent think-aloud protocols in almost every usability tests, and 91% of them (n = 154) used the concurrent think-aloud protocols in at least half of their usability tests. In contrast, only 21% of them (n = 36) used the retrospective think-aloud protocols in almost every usability tests, and the majority of them (61%, n = 104) almost never or only occasionally (i.e., roughly a quarter of the tests) used the retrospective think-aloud protocols.

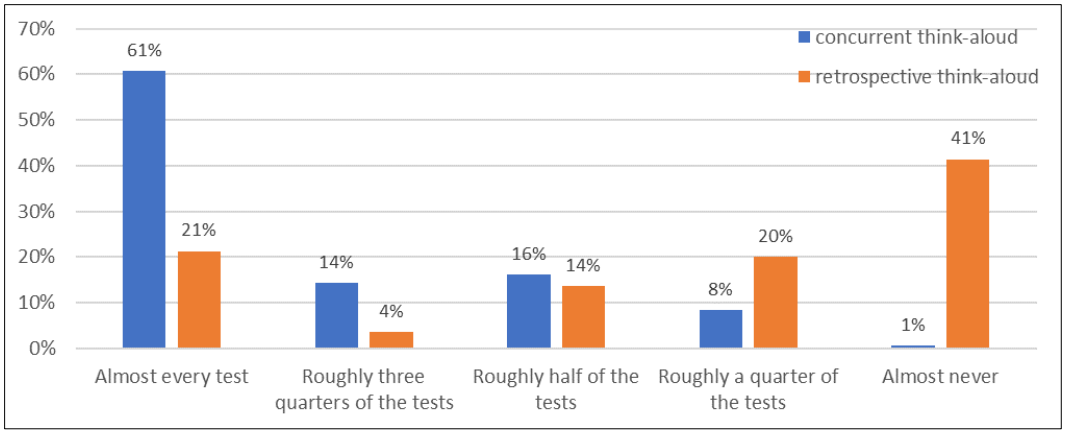


Figure 2. The frequency of using concurrent-think-aloud protocols and retrospective think-aloud protocols among the respondents.

Motivation: We asked respondents about their motivation for using think-aloud protocols and found that 51% of the respondents (n = 86) used the think-aloud protocols to both inform the design (e.g., problem discovery) and to measure the performance (e.g., success rate); 48% of them (n = 81) only used the protocols to inform the design and only 1% of them (n = 2) only used the protocols to measure the performance.

Testing environments: We asked respondents about the test environments in which they used think-aloud protocols (see results in Figure 3). Specifically, 75% of the respondents (n = 127) used the protocols in controlled lab studies, 72% of them (n = 121) used the protocols in remote usability testing, and 48% of them (n = 81) used the protocols in field studies. The total does not sum up to 100% because respondents can use the think-aloud protocols in more than one test environment.

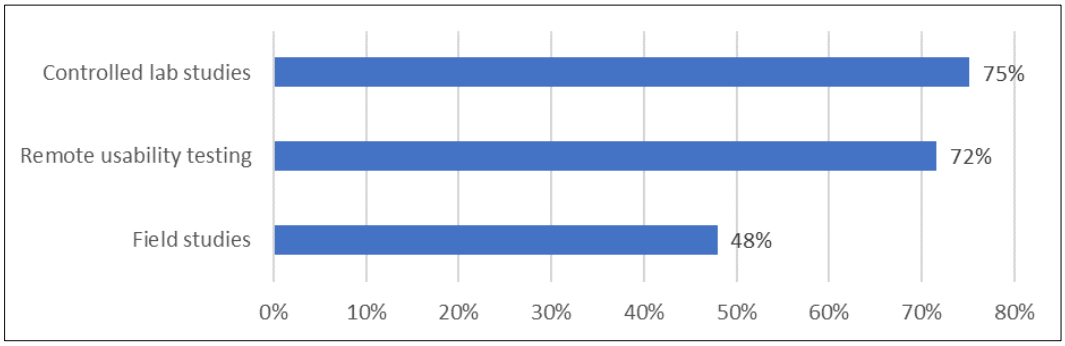


Figure 3. The testing environments in which UX practitioners use think-aloud protocols.

Conducting Think-Aloud Sessions

Types of tasks for think-aloud sessions: We asked respondents about the types of tasks that they ask their participants to work on during think-aloud sessions (see results in Figure 4). Specifically, 27% of them (n = 46) only ask their participants to work on tasks without instruction steps to follow (e.g., navigating a website), while 12% of them (n = 20) only ask their participants to work on tasks with instruction steps to follow (e.g., setting up a TV with its manual). In contrast, the majority of the respondents (61%, n = 103) used both two types of tasks during think-aloud sessions.

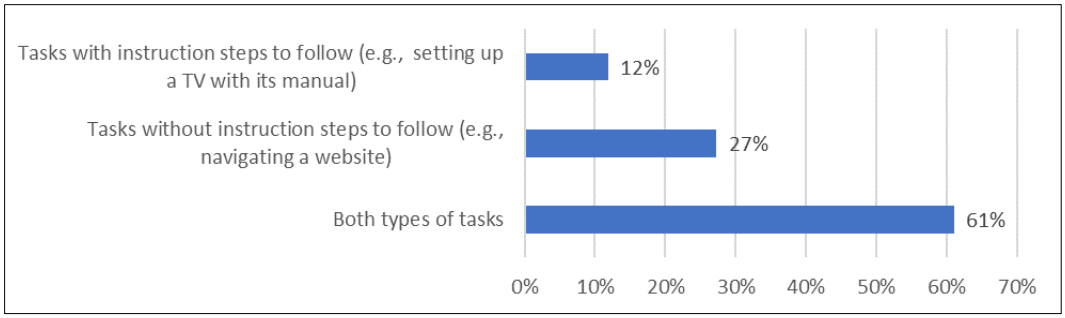


Figure 4. The types of tasks that UX practitioners ask their participants to work on during think-aloud sessions.

Practice sessions: Ericsson and Simon have suggested that practitioners should ask their participants to practice thinking aloud before conducting the actual think-aloud sessions (1984). We asked the respondents about the frequency of conducting a practice session before starting the actual think-aloud test sessions (see results in Figure 5). Specifically, the majority of the respondents (61%, n = 103) almost never do it, 7% (n = 12) only do it roughly a quarter of the time, 6% (n = 10) do it roughly half of the time, 2% (n = 4) do it roughly three-quarters of the time, and 24% (n = 40) do it almost all the time. The result shows that the majority of the UX practitioners seldom ask their participants to practice think-aloud before conducting the actual think-aloud sessions.

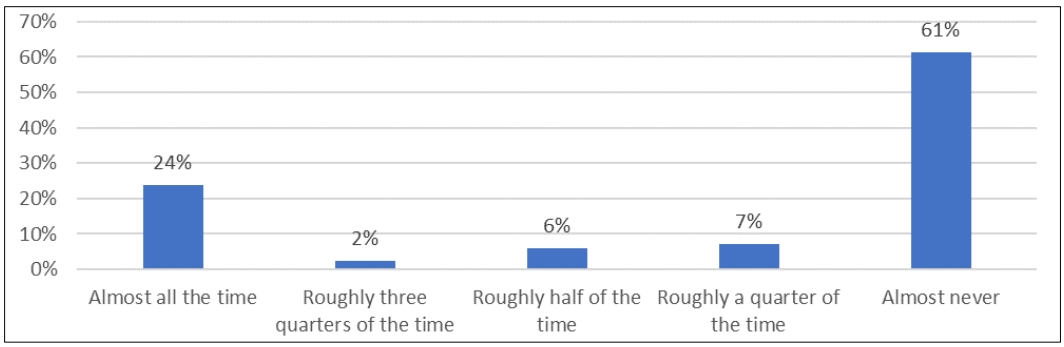


Figure 5. The frequency of conducting practice sessions before actual think-aloud sessions.

Instructions for requesting verbalizations: When using the classic think-aloud protocol (Ericsson & Simon, 1984), moderators are required to only ask their participants to say out loud everything that *naturally* comes into the mind. We asked respondents what else they explicitly ask their participants to verbalize during think-aloud sessions in addition to the thoughts that naturally comes into the mind (see results in Figure 6). Specifically, only 7% of the survey respondents (n = 12) reported that they do not ask their participants to verbalize anything beyond what naturally comes into their mind. In contrast, 80% (n = 136) mentioned that they also explicitly ask their participants to verbalize their feelings, 70% (n = 119) explicitly ask their participants to verbalize their feedback, 55% (n = 93) explicitly ask their participants to verbalize their actions on the interface, and 33% (n = 55) explicitly ask their participants to verbalize their design recommendations.

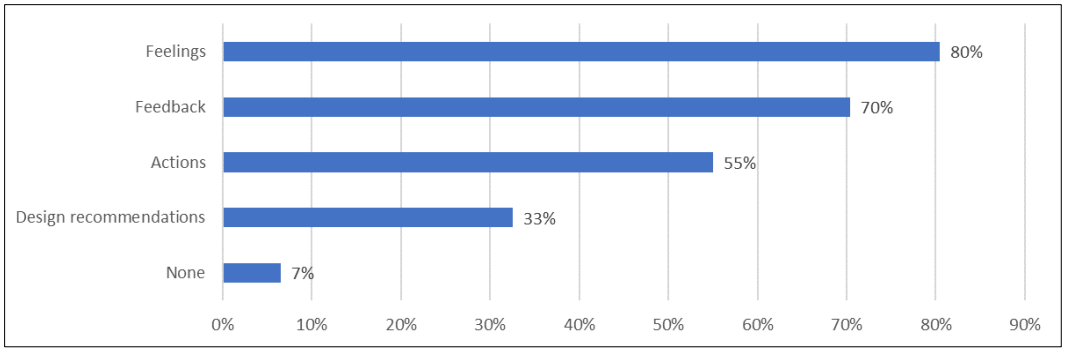


Figure 6. The content that respondents ask their participants to verbalize in addition to the thoughts that come naturally into the mind.

To better understand what types of content that respondents often request their participants to verbalize together, we counted the number of occurrences of different combinations of content that they ask their participants to verbalize in addition to the thoughts that come naturally into the mind (see results in Figure 7).

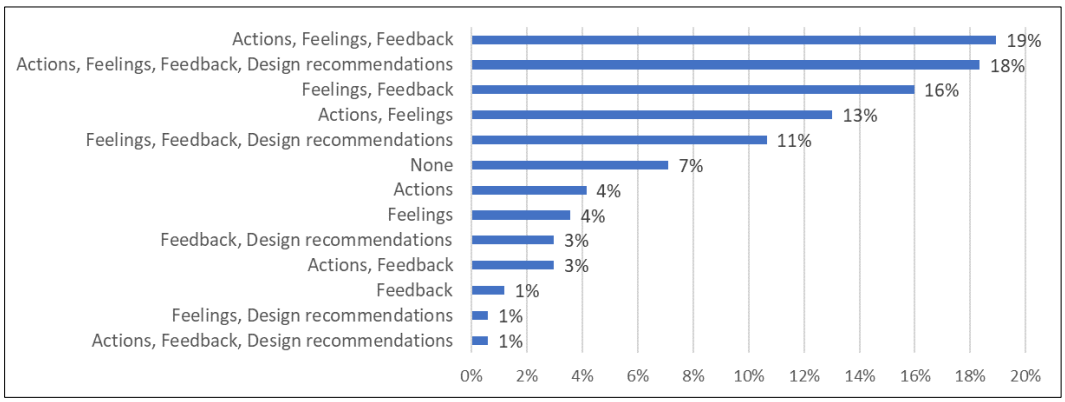


Figure 7. The percentages of different combinations of the content that respondents ask their participants to verbalize.

Prompting participants: When using the classic think-aloud protocol (Ericsson & Simon, 1984), moderators are required to keep the interaction with their participants to a minimal level and only remind them to keep talking if they fall into silence. We asked respondents whether they prompt their participants during think-aloud sessions and found that only 22% of the respondents (n = 37) keep the interaction minimal and do not prompt their participants with questions. In contrast, 78% of the respondents (n = 132) prompt their participants.

In addition, 91% of the respondents (n = 154) also reported how the frequency of prompting their participants had changed compared to when they just started their UX career (see results in Figure 8). Among these respondents, 44% (n = 67) felt that the frequency with which they prompt their participants remained roughly the same, 41% (n = 64) felt that the frequency for prompting their participants had only slightly changed, and 15% (n = 23) felt that the frequency had changed significantly.

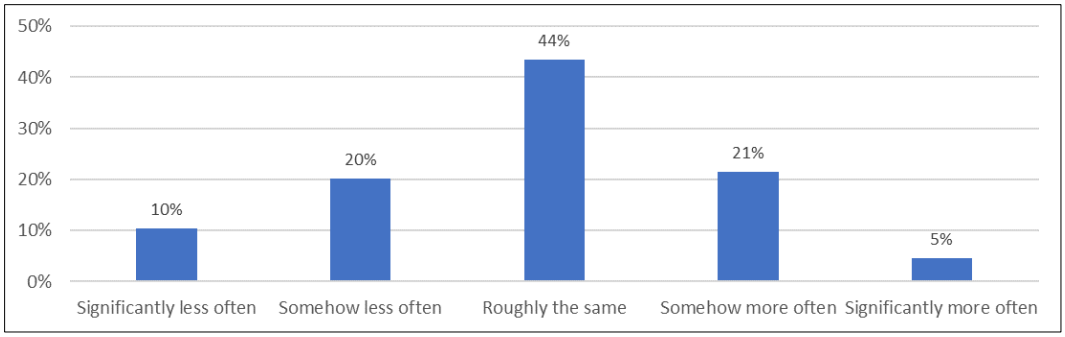


Figure 8. How the frequency with which respondents prompted their participants during think-aloud sessions had changed compared to when they just started their UX career.

Correlation analysis: We examined whether there was any correlation between respondents' profile info and their practices of using think-aloud protocols. Specifically, we performed the Spearman's rank-order correlation test when both variables were ordinal data and the Chi-square test when there was categorical data (see results in Table 3). In sum, the tests did not find any significant correlation for most pairs except between the size of respondents' companies and whether respondents request their participants to verbalize content beyond what comes into the mind, $\chi^2(4, N = 169)$=14.403, p = 0.006.

Table 3. Correlation Analysis Between Responders' Profile Information and Their Practices of Conducting Think-Aloud Sessions

Respondents' profile info	**Frequency of conducting practice sessions** (ordinal data)	**Whether asking users to verbalize content beyond what comes into the mind** (categorical data)	**Whether prompting users during the study session** (categorical data)
The size of their companies (ordinal data)	$r_s(167)$ = -0.0294, p = 0.7043	$\chi^2(4, N = 169)$ = 14.403, p = 0.006*	$\chi^2(4, N = 169)$ = 1.3939, p = 0.8453
The UX experience (ordinal data)	$r_s(167)$ = -0.0166, p = 0.8308	$\chi^2(4, N = 169)$ = 2.6906, p = 0.6109	$\chi^2(4, N = 169)$ = 2.7057, p = 0.6082

* indicates significance

Analyzing Think-Aloud Sessions

Activities performed for analyzing sessions: We asked respondents about specific activities they did when analyzing think-aloud sessions. The activities were the following: review observation notes of the usability test, review the test session recording, review post-task interview data, review post-task questionnaire data, or transcribe and review the transcript of the session. These options were based on a prior survey (McDonald et al., 2012) and were updated via a pilot study (see results in Figure 9). Specifically, 89% of the respondents (n = 151) review observation notes, 77% of them (n = 130) review the session recordings (e.g., audio/video recordings), 70% of them (n = 118) review post-task interview data, 60% of them (n = 102) review the questionnaire/survey data, and 56% of them transcribe and review the transcripts (i.e., what participants said).

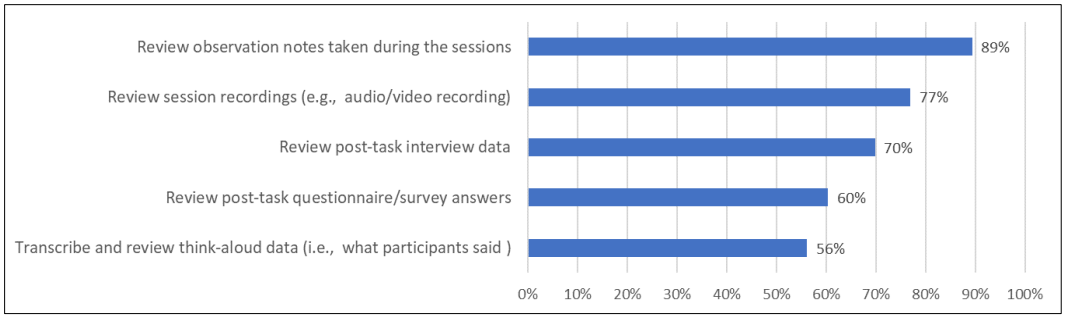


Figure 9. The activities that UX practitioners perform when analyzing think-aloud sessions.

Information for locating usability problems: We asked respondents about the types of information they thought would help locate usability problems (see results in Figure 10). Specifically, when reviewing think-aloud sessions to identify usability problems, 94% of them (n = 159) thought what participants were doing (e.g., user actions on the interface) is helpful, 86% of them (n = 145) thought what participants said during the sessions is helpful, and 76% of them (n = 128) also thought how participants said it (e.g., pauses, tone) is helpful.

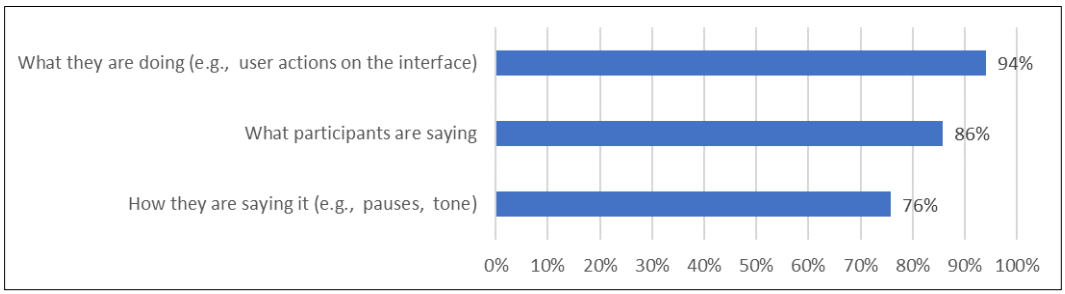


Figure 10. The types of information that are helpful for UX practitioners to locate usability problems.

Information sought out from users' verbalizations: We asked respondents about the information that they looked for when analyzing their participants' verbalizations (i.e., utterances; see results in Figure 11). Specifically, 94% of them (n = 153) looked for expressions of feelings (e.g., excitement, frustration), 89% (n = 145) looked for their participants' comments (e.g., feedback), 74% (n = 119) looked for their participants' action descriptions, 70% (n = 116) looked for their participants' explanations, and 30% (n = 49) looked for their participants' design recommendations.

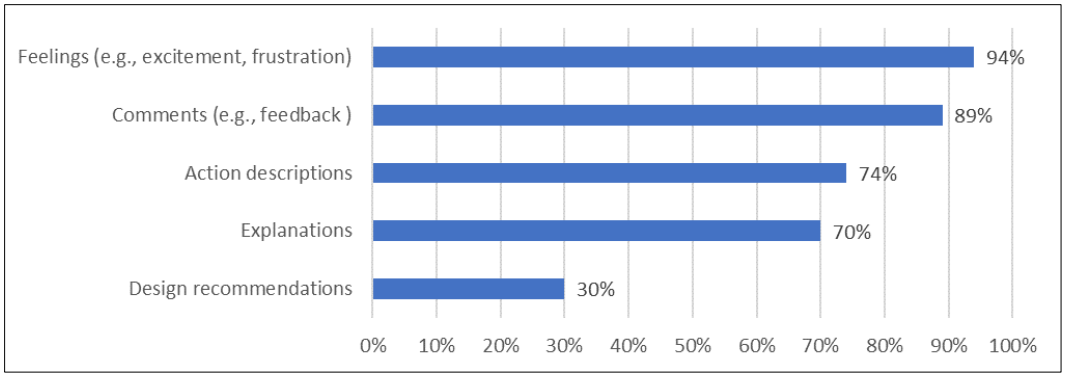


Figure 11. The types of information that UX practitioners seek in users' verbalizations.

Delivering analysis results: We asked respondents what activities they performed when delivering analysis results. The following were the three activities: write an informal usability test report, write a formal usability test report, and have a data analysis discussion meeting. We did not provide definitions for these activities to make them open to interpretation. They could choose multiple options if applicable (see results in Figure 12). Specifically, when analyzing a think-aloud session, 69% of them (n = 116) wrote an informal usability test report, 58% (n = 98) wrote a formal usability test report, and 57% of them (n = 97) had a data analysis discussion meeting.

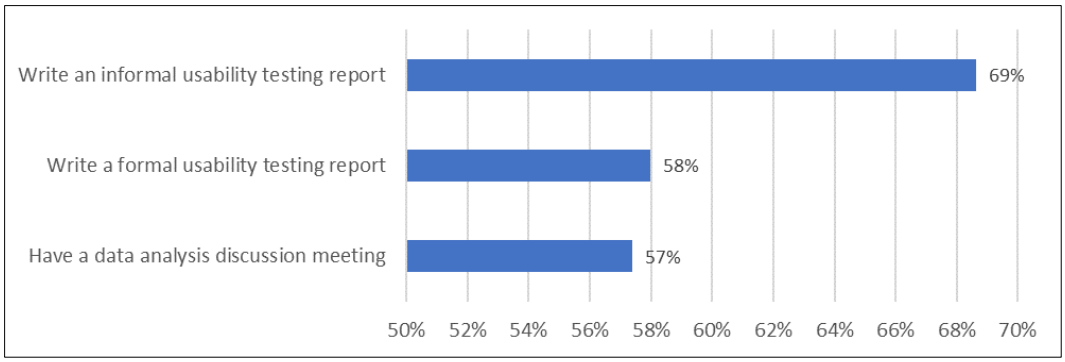


Figure 12. The ways in which UX practitioners deliver their analysis results.

Participation in the three types of data analysis: We asked the respondents who write formal and informal usability reports about how they did this. We gave them the following six options: Only myself, UX designers/researchers, UX team lead, Lead of non-UX teams (e.g., engineering, marketing), Other non-UX team members (e.g., engineers), and C-level executives (e.g., CEO). In addition, we also asked respondents who would attend data analysis discussion meetings with the same set of options except "Only myself." They could choose multiple options if applicable (see results in Figure 13). More than half of the respondents (56%, n = 95) wrote informal usability testing reports alone and nearly half of the respondents (42%, n = 71) also wrote formal usability testing reports alone. In addition, UX team members were the primary authors of informal/formal reports with occasional help from outside of the UX team. In contrast, non-UX team members were more involved in data analysis discussion meetings.

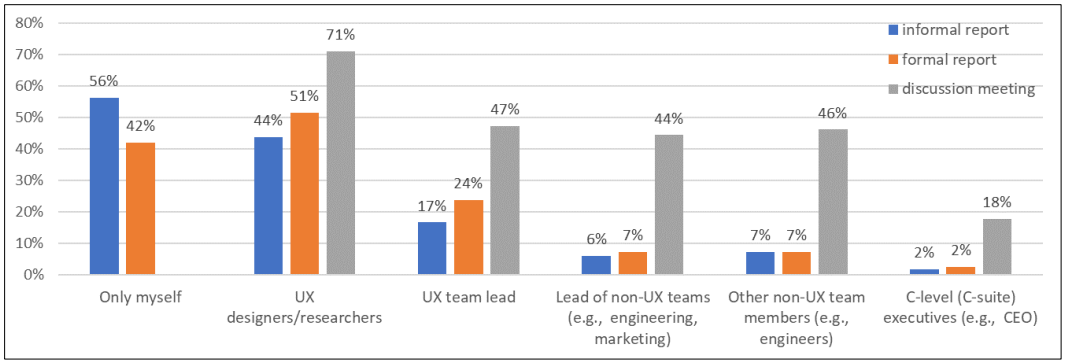


Figure 13. Participation in three types of data analysis activities: writing an informal usability test report, writing formal usability test report, and having a data analysis discussion meeting.

Challenges of Using Think-Aloud Protocols

We asked respondents what their biggest inefficiencies or difficulties had been in conducting and analyzing think-aloud sessions as an optional short-answer question. We present the key findings from the responses in the following paragraphs.

Challenges for conducting sessions: Our qualitative analysis reveals three main challenges that respondents encountered when conducting think-aloud sessions. First, getting their participants to think aloud is a challenge. Participants' personality and their ability to verbalize thoughts and the complexity and duration of the tasks are factors that influence the amount of content that they verbalize. For example, some people tend to be able to verbalize more readily than others, which can create an unbalanced representation of potential users. For some products, the target population may not be able to verbalize properly, for example, children. Participants may also feel less comfortable verbalizing their thoughts when the task is complex. Furthermore, it may also be fatiguing for users to verbalize their thoughts if the task takes too long to complete.

Another challenge facing respondents is to create a comfortable and neutral environment that encourages participants to honestly verbalize their thought processes. This is challenging because participants might want to say nice things or may be reluctant to offer criticism during the test sessions, which could preclude UX practitioners from identifying usability bugs.

Finally, being patient and knowing when to interrupt participants is challenging. It is valuable to observe and understand how participants deal with the tasks themselves and recover from errors. Interrupting the process with prompts too early could change their way of interacting with the test interface. Moreover, because part of the goal of usability evaluations is to gather data on what is difficult/impossible for users, it is often necessary to observe users struggle a bit during the evaluation to understand their "pain points." That being said, it is also bad to let participants get stuck for too long as they can be too frustrated, which could, in turn, affect the rest of the test session and consequently the amount of feedback that can be gained from the test session.

Challenges for analyzing sessions: While previous research reported general practices in analyzing usability evaluation (Følstad, Law, & Hornbæk, 2012), our survey study found specific challenges that respondents faced when analyzing think-aloud test sessions. This survey study showed that respondents reviewed think-aloud session notes (89%) more often than the session recordings (77%; see Figure 9). Respondents felt that reviewing think-aloud session video recordings was arduous because recorded think-aloud sessions often contain so much data that transcribing and coding them takes a significant amount of time. Consequently, instead of transcribing sessions and reviewing transcripts, respondents often rely on "their memory of participants' sentiments and actions" or the notes.

Despite the convenience of observation notes, respondents realized that it is "easy to make judgments that might be off if they don't refer back to actual transcripts or recordings" and thus considered reviewing think-aloud session recordings a necessary part of their analysis process.

First, it is necessary to match the observation notes with the corresponding segments in the session recordings to understand the context of the notes. Second, it is necessary to review the session recordings to capture points that might have been missed by observation notes because notetakers can only write down the points that seem to be important from their perspective, and any individual perspective can be incomplete or biased. Indeed, previous research also suggested that while some of the usability problems may be captured by notes, much of the insight is often lost and needs to be reconstructed by conducting video data analysis later (Kjeldskov, Skov, & Stage, 2004).

This survey study further identifies two challenges associated with reviewing think-aloud sessions. One challenge is to compare users' verbalization data with other streams of data to triangulate the issues that users encountered. One such comparison is to pair the user's actions on the interface with what they are saying (i.e., utterances) during the session. In scenarios where multiple streams of data are acquired, respondents had to correlate the verbalizations with other sensor data. Recent research has shown that considering verbalizations with other sensor data, such eye-tracking data (Elbabour, Alhadreti, & Mayhew, 2017; Elling, Lentz, & de Jong, 2012), EEG data (Grimes, Tan, Hudson, Shenoy, & Rao, 2008), or functional near-infrared spectroscopy (fNIRS; Lukanov, Maior, & Wilson, 2016), can potentially increase the reliability and validity of the findings. Another challenge is to match the observation notes with the context in which the notes are taken. It is not always possible to notate the exact timestamps when notes are taken. Consequently, matching notes (e.g., observations about users' facial expressions) with the audio stream often require evaluators to watch the entire recording. Another example of this challenge comes from the emerging VR and AR applications. To make sense of users' verbalizations when they interact with a VR or AR application, evaluators need to correlate the verbalizations with the visual content that participants observed during the sessions.

Reviewing think-aloud sessions is time-consuming. Our respondents reported that they often had limited time to complete the analysis and faced the tension between achieving high reliability and validity in their analysis and completing their analysis efficiently. To cope with the tension, respondents reported using strategies such as developing better note-taking skills or having a team of UX professionals observe a think-aloud test session. Respondents also proposed to discuss the session afterward with their peers.

In addition to reviewing sessions, respondents also pointed out that it can be valuable to keep track of the examples of different types of usability problems that they had observed over time and develop a taxonomy to describe the patterns in the data that commonly occur when users encountered usability problems. Such patterns, examples, and the taxonomy could act as templates that potentially help them quickly identify common issues that users encounter and the solutions that they had accumulated in a new test context.

Discussion

Our study respondents worked in different geographic locations, in different industrial fields, and different sized UX teams. They also played different roles and possessed different levels of experience as UX professionals. Thus, the survey responses have uncovered a wide range of UX practitioners' practices surrounding the conduct and analysis of think-aloud sessions. Next, we discuss the implications of the survey responses.

General Use of Think-Aloud Protocols

This survey study found that 86% of all respondents (169 out of 197) used think-aloud protocols when conducting usability tests, which was viewed by respondents as the most popular method to detect usability problems. Among the 91% of all respondents (n = 179) who had learned think-aloud protocols, 95% (169 out of 179) actually used the protocols in their usability tests. This result shows that think-aloud protocols are still widely used in industry. This result is consistent with that of the survey study conducted in 2010 (McDonald et al., 2012), which showed that 90% of the usability practitioners often use think-aloud protocols.

Our study shows that concurrent think-aloud protocols are much more popular than the retrospective think-aloud protocols among UX practitioners. Of the respondents, 91% used the concurrent think-aloud protocols in at least half of their usability tests (see results in Figure 2).

In contrast, only 39% of the respondents used the retrospective think-aloud protocols in at least half of their usability tests.

Our study also shows that think-aloud protocols are widely and almost equally used in both controlled lab studies (75%) and remote usability testing (72%). Compared to the most recent survey study conducted by McDonald et al. (2012), our survey study identified that remote usability testing is increasingly popular and think-aloud protocols are widely used in the remote usability testing as well as in controlled lab studies. Remote usability testing allows UX practitioners to recruit geographically distributed and diverse users to participate in usability testing in their native work environments. Previous research has shown that remote synchronous usability testing, in which there is a test facilitator, is virtually equivalent to conventional lab-based controlled user studies in terms of the number of identified usability problems and the task completion time (Andreasen, Nielsen, Schrøder, & Stage, 2007). Furthermore, previous research also showed that although participants experienced higher workload in remote synchronized usability testing (i.e., a web-based two-dimensional screen-sharing approach and a three-dimensional virtual world) than conventional lab-based user studies as measured by post-task NASA TLX questionnaire, they generally enjoyed the remote synchronous usability testing (Madathil & Greenstein, 2011). Similarly, despite remote asynchronous usability testing, in which there is no test moderator and may reveal fewer problems than conventional lab-based user studies, it requires significantly less time and thus is cost-effective (Bruun, Gull, Hofmeister, & Stage, 2009).

Conducting Think-Aloud Sessions

To ensure the validity of participants' verbalizations, Ericsson and Simon (1984) provided three guidelines for conducting classic think-aloud sessions: keep the interaction minimal (i.e., only remind users to think aloud if they fall into silence for a period of time), use neutral instructions (i.e., instructions that do not ask for specific types of content), and have practice session(s). A meta-analysis of 94 think-aloud studies showed that an artificial change in performance can happen if these guidelines are breached (Fox, Ericsson, & Best, 2011). However, previous research has documented that the gap between the theory and the practice of using think-aloud protocols existed (Boren & Ramey, 2000), and our survey study provides evidence that such a gap between the theory and the practice still exists. Specifically, we found that respondents did not always adhere to the three guidelines. We analyze potential reasons for violating each guideline in the following paragraphs.

Our study shows that only 16% of the respondents reminded their participants to keep talking when they fell into silence for a substantial period without actively probing them with questions while they were thinking aloud. Previous research has attributed the reason for not adhering this guideline to the differences between the original goal of think-aloud protocols and the goal of using them in usability testing (Boren & Ramey, 2000). The original goal is to study the unaltered human thought processes. Numerous studies have shown that probing or intervention (i.e., interaction with participants) could potentially alter the participant's thought processes, which could make the reported verbalizations not be an authentic representation of their thoughts (Alhadreti & Mayhew, 2017; Ericsson & Simon, 1984; Fox et al., 2011). Thus, UX practitioners should keep their intervention or probing minimal if possible. However, we also acknowledge that the goal of using think-aloud protocols in usability testing is mainly to identify usability bugs or to evaluate potential users' performance instead of just acquiring unaltered thought processes. Because of this difference, previous research suggests that UX practitioners may deviate from the guidelines and interact with their participants in two situations (Nielsen, 1993). One is when participants are frustratingly stuck. In this situation, interacting with them to help them recover from the error would allow the test to continue again, which would, in turn, allow UX practitioners to identify further usability issues. Another situation is when participants are struggling with a familiar problem, whose impact has been identified and well understood with previous test participants. In this situation, it is less meaningful to sit and observe participants struggle with the problem again. Furthermore, as previous research suggested that audio interruptions (e.g., a beeping sound) during think-aloud sessions may affect participants more than visual interruptions (e.g., an on-screen notification; Hertzum & Holmegaard, 2013), future research should examine the possibility of probing participants through the visual modality, such as showing an onscreen notification with a question, to acquire richer data while minimizing the risk of altering their thought processes.

Despite Ericsson and Simon's guidelines that recommend practitioners use neutral instructions (i.e., only ask participants to report the content that naturally comes into their mind), our study reveals that only 7% of the respondents adhered to this guideline. Most of the respondents explicitly asked their participants to verbalize other types of content, such as feelings, comments, actions, and even design recommendations. This is concerning because research explicitly asking participants to verbalize a particular type of content can change their task-solving behavior (McDonald & Petrie, 2013), which may mask potential usability problems.

Our study also shows that UX practitioners also do not always follow the third guideline. For example, most of the respondents (61%) rarely asked their participants to practice thinking aloud before conducting the actual sessions. Unfortunately, previous research showed that without practicing thinking aloud, participants often have difficulty verbalizing their thought processes (Charters, 2003). Consequently, instead of treating the practice session as a burden, UX practitioners should treat it as an opportunity to help their participants become familiar with thinking aloud, which would help the participants verbalize their thoughts more naturally and frequently. This would potentially reduce the need for probing participants or asking them to verbalize their thoughts and feelings, which could, in turn, enhance the adherence to the other two guidelines. In sum, allowing participants to practice thinking aloud would ultimately help UX practitioners acquire more rich data to understand their user experiences.

Analyzing Think-Aloud Sessions

When analyzing think-aloud sessions, UX practitioners reviewed observation notes more often than the session recordings and the transcriptions. One potential reason was that transcribing and reviewing the session recordings is arduous and time-consuming. Previous research pointed out that UX practitioners often face time pressure for their analysis (Chilana, Wobbrock, & Ko, 2010). Indeed, the qualitative feedback from our survey respondents echoed this finding. Although the survey respondents largely knew that their judgments might be inaccurate if they did not refer to the actual session recordings, they often had to make trade-offs between achieving high reliability and validity and being efficient in their analysis. Currently, there are no known methods to deal with this tension effectively. The methods that survey respondents used include developing better note-taking skills and referring to the notes during analysis or having multiple UX practitioners observe a test session and then discuss to recap the session afterward. However, it remains unknown whether these methods are effective or if there are other more effective methods available. Indeed, recent research also suggested gaining a richer understanding of the tradeoffs that evaluators make and the impact of their decisions (MacDonald & Atwood, 2013). Therefore, future research should investigate methods and processes that can better balance the reliability, validity, and efficiency of the analysis of think-aloud sessions.

Our study also reveals a need to identify common patterns from users' data that point to the moments when they experience problems in think-aloud sessions. Research has shown that users' verbalizations can be classified into different categories (Cooke, 2010; Hertzum, Borlund, & Kristoffersen, 2015). Recently, Fan et al. found subtle patterns that tend to occur when users encounter problems in concurrent think-aloud sessions (Fan, Lin, Chung, & Truong, 2019). Specifically, when users encounter problems, their verbalizations tend to be in the observation category (e.g., comments and remarks) and include negative sentiments, questions, more verbal fillers, abnormal pitches, and speech rates (Fan et al., 2019). In addition to subtle verbalization and speech patterns, do users' eye movements also exhibit certain patterns when they experience problems in think-aloud sessions? Similarly, do users' facial expressions and physiological signals (e.g., heartbeat, skin conductance) tend to change in a predictable way when they encounter usability problems or enjoy the interaction? Future research should explore whether such patterns exist. If these patterns do exist, they could be leveraged to design systems that automatically highlight portions of a think-aloud test session in which the user more likely experienced a problem, which in turn could help UX practitioners better allocate their attention during analysis.

Conclusion

We conducted an international survey study to understand the practices and challenges of using think-aloud protocols in industry. Based on the responses from 197 UX practitioners who worked in different industrial fields and different geographic locations, we have identified the practices and challenges surrounding the conduct and analysis of think-aloud sessions. The findings of the survey study could potentially inform UX practitioners about how their peers perceive and use think-aloud protocols. Our survey study also reveals opportunities in developing better methods and tools to make conducting and analyzing think-aloud sessions more effective, for example, by identifying patterns in users' data (e.g., verbalizations, actions, and physiological measures) that commonly occur when they encounter problems and by developing a taxonomy of the patterns that would allow UX practitioners to improve the efficiency and communication of usability analyses in a field- and scale-invariant manner (i.e., independent of industrial fields or the amount of test sessions).

Tips for Usability Practitioners

Our survey study discovered that many UX practitioners' current practices deviate from Ericsson and Simon's three guidelines (1984). Considering the number of empirical studies that examined the effect of these deviations, we offer the following recommendations for UX practitioners to consider:

- Conduct a practice session before actual study sessions. This would help participants practice and get used to verbalizing their thoughts more frequently and reduce the need for prompting or intervention during the study sessions (Charters, 2003).
- Use neutral instructions to ask participants to report whatever comes into their mind naturally and avoid instructing them to report a particular type of content (McDonald, McGarry, & Willis, 2013; McDonald & Petrie, 2013).
- Keep interaction with participants minimal (Alhadreti & Mayhew, 2017; Fox et al., 2011; Hertzum, Hansen, & Andersen, 2009) .
- Consider using think-aloud protocols in both controlled lab studies and remote (synchronous and asynchronous) usability testing (Andreasen et al., 2007; Bruun et al., 2009).

We have further derived the following tips and recommendations based on the current practices and challenges facing UX practitioners when using think-aloud protocols:

- Pay attention to participants' actions, their verbalizations, and how they verbalize (e.g., speech rate, pitch) when analyzing think-aloud sessions.
- Acknowledge that tension exists between achieving high validity and reliability and maintaining high efficiency in analyzing large amounts of think-aloud sessions. Create more open dialogues to discuss fast-paced and reliable analysis methods to cope with the tension.
- Design methods to understand, capture, and categorize patterns emerged from participants' data (e.g., verbalizations, actions, eye-tracking, and physiological measures) that commonly occur when users encounter usability problems. These patterns could then be organized into a taxonomy that would allow UX practitioners to improve the efficiency and communication of usability analyses in a field- and scale-invariant manner.

These tips are based on the practices of a large percentage of UX practitioners. In practice, UX practitioners conduct and analyze think-aloud sessions in different contexts (e.g., different user groups and different types of products) with different constraints. Therefore, we suggest UX practitioners evaluate the underlying motivations and justifications underlying the use of these tips to make informed decisions.

Acknowledgements

This work was complete when the first author was a PhD student at the University of Toronto. We would like to thank UXPA organizers Chris Bailey, Adams Banks, Bendte Fagge, Yvonne Liu, Nicole Maynard, Hannes Robier, Nabil Thalmann, Rik Williams, and Jackeys Wong for helping us distribute the survey to their associated UXPA local chapters in Asia, Europe, and North America. We would also like to thank all the anonymous UX professionals who participated in the survey study. Finally, we would like to thank our anonymous reviewers and Editor-in-Chief Dr. James Lewis for their constructive reviews and feedback and also thank Sarah Harris for copyediting the manuscript. We have made a PDF version of the survey available here: http://mingmingfan.com/doc/ThinkAloudSurvey-FAN-Mingming.pdf.

References

Alhadreti, O., & Mayhew, P. (2017). To intervene or not to intervene: An investigation of three think-aloud protocols in usability testing. *Journal of Usability Studies*, *12*(3), 111–132.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What happened to remote usability testing?: An empirical study of three methods. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1405–1414). ACM. https://doi.org/10.1145/1240624.1240838

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09* (pp. 1619–1628). ACM Press. https://doi.org/10.1145/1518701.1518948

Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, *12*(2), 68–82. https://doi.org/10.26522/brocked.v12i2.38

Chilana, P. K., Wobbrock, J. O., & Ko, A. J. (2010). Understanding usability practices in complex domains. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2337–2346). ACM Press. https://doi.org/10.1145/1753326.1753678

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication*, *53*(3), 202–215. https://doi.org/10.1109/TPC.2010.2052859

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Intellect Books.

Elbabour, F., Alhadreti, O., & Mayhew, P. (2017). Eye tracking in retrospective think-aloud usability testing: Is there added value? *Journal of Usability Studies*, *12*(3), 95–110.

Elling, S., Lentz, L., & de Jong, M. (2012). Combining concurrent think-aloud protocols and eye-tracking observations : An analysis of verbalizations. *Ieee Transactions on Professional Communication*, *55*(3), 206–220. https://doi.org/10.1109/TPC.2012.2206190

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. MIT Press.

Fan, M., Lin, J., Chung, C., & Truong, K. N. (2019). Concurrent think-aloud verbalizations and usability problems. *ACM Transactions on Computer-Human Interaction.*, *26*(5), 28:1--28:35. https://doi.org/10.1145/3325281

Følstad, A., Law, E., & Hornbæk, K. (2012). Analysis in practical usability evaluation: A survey study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2127–2136). ACM Press. https://doi.org/10.1145/2207676.2208365

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316.

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., & Rao, R. P. N. (2008). Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 835). ACM Press. https://doi.org/10.1145/1357054.1357187

Hertzum, M., Borlund, P., & Kristoffersen, K. B. (2015). What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction*, *31*(9), 557–570. https://doi.org/10.1080/10447318.2015.1065691

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165–181. https://doi.org/10.1080/01449290701773842

Hertzum, M., & Holmegaard, K. D. (2013). Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction*, *29*(5), 351–364. https://doi.org/10.1080/10447318.2012.711705

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour and Information Technology*, *29*(1), 97–111. https://doi.org/10.1080/01449290801939400

Jurca, G., Hellmann, T. D., & Maurer, F. (2014). Integrating agile and user-centered design: A systematic mapping and review of evaluation and validation studies of agile-UX. In *Proceedings - 2014 Agile Conference, AGILE 2014* (pp. 24–32). IEEE. https://doi.org/10.1109/AGILE.2014.17

Kjeldskov, J., Skov, M. B., & Stage, J. (2004). Instant data analysis: Conducting usability evaluations in a day. In *Proceedings of the third Nordic conference on Human-computer interaction - NordiCHI '04* (pp. 233–240). ACM Press. https://doi.org/10.1145/1028014.1028050

Lukanov, K., Maior, H. A., & Wilson, M. L. (2016). Using fNIRS in usability testing. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (pp. 4011–4016). https://doi.org/10.1145/2858036.2858236

MacDonald, C. M., & Atwood, M. E. (2013). Changing perspectives on evaluation in HCI: Past, present, and future. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (pp. 1969–1978). ACM Press. https://doi.org/10.1145/2468356.2468714

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, *55*(1), 2–19. https://doi.org/10.1109/TPC.2011.2182569

McDonald, S., McGarry, K., & Willis, L. M. (2013). Thinking-aloud about web navigation: The relationship between think-aloud instructions, task difficulty and performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 2037–2041. https://doi.org/10.1177/1541931213571455

McDonald, S., & Petrie, H. (2013). The effect of global instructions on think-aloud testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (pp. 2941–2944). ACM Press. https://doi.org/10.1145/2470654.2481407

Madathil, K. C., & Greenstein, J. S. (2011). Synchronous remote usability testing: A new approach facilitated by virtual worlds. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (pp. 2225–2234). ACM Press. https://doi.org/10.1145/1978942.1979267

Nielsen, J. (1993). *Usability engineering*. Academic Press, Inc.

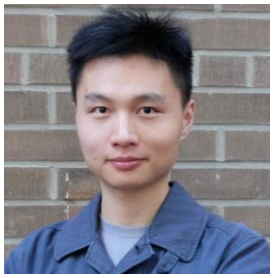
Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the 6th ACM conference on Designing Interactive systems - DIS '06* (p. 209). ACM Press. https://doi.org/10.1145/1142405.1142439

Preece, J., Rogers, Y., & Sharp, H. (2015). *Interaction design : Beyond human-computer interaction*. Wiley.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design and conduct effective tests*. John Wiley & Sons.

Shi, Q. (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* (p. 344). ACM. https://doi.org/10.1145/1463160.1463198

About the Authors



Mingming Fan
Dr. Fan is an Assistant Professor at Rochester Institute of Technology. He received his PhD from the University of Toronto. His dissertation focuses on leveraging subtle verbalization and speech patterns to assist UX evaluators with analyzing think-aloud sessions. His research interests include AI-assisted UX design and analysis methods, assistive technology, and human-sensing technologies and applications.

Khai N. Truong
Dr. Truong is a Professor in the Department of Computer Science at the University of Toronto. He received his BS and PhD from Georgia Institute of Technology. His research interests are in human-computer interaction and ubiquitous computing. Much of his current work focuses on the design and evaluation of health and assistive technologies.

Serina Shi
Ms. Shi is a graduate of the Faculty of Information at the University of Toronto who specialized in UX design and information management. She owns a small dumpling shop in downtown Toronto. Her current interests are social project design and understanding human behavior from her business perspective.