Talking About Thinking Aloud: Perspectives from Interactive Think-Aloud Practitioners

uxpajournal.org/talking-about-thinking-aloud-perspectives-from-interactive-think-aloud-practitioners

May 26, 2023

Abstract

It is widely reported in the literature that intervening during usability testing sessions affects user behavior and compromises the validity of the test. However, this contrasts with the ongoing popularity of Interactive Think-Aloud (ITA) amongst practitioners. We report an in-depth qualitative study that explored this tension between theory and practice through nine interviews with ITA practitioners. Our findings add nuance to many established ideas about ITA but also reveal novel practices and attitudes. For example, ITA is sometimes used to slow down users as they navigate through a system, to manage external pressures such as recruitment difficulties, and to reframe a session as a kind of interview or participatory study. We also found that participants (ITA practitioners) experienced unexpected difficulties with ITA, including the risk that it results in overly reflective think-aloud and creates challenges in team working. Participants understood that ITA causes reactivity, and they reported taking steps to reduce it. However, overall, they did not see the traditional positivist objective of valid problem discovery as a realistic or high-priority goal for usability testing. They believed that ITA data can be useful and valid even if user behavior is not wholly realistic. Based on this, we argue against the narrow problem-counting approach often employed in the comparative usability evaluation studies that have sometimes seemed to discredit ITA. We also make the case for broadening how we think about the validity of usability testing data, and we argue that forms of ITA may be appropriate in some situations.

Keywords

usability testing, traditional think-aloud, interactive think-aloud (ITA), active intervention think-aloud, relaxed think-aloud

Introduction

Usability testing in which participants are asked to “think aloud” (TA) is a hugely popular usability evaluation method in industry (Nielson, 2019; McDonald et al., 2012). Verbalizations made by the participant while thinking aloud are useful in formative evaluations because they can help evaluators identify usability problems, which can then be addressed through redesign. Under the Traditional Think-Aloud (TTA) protocol, the moderator silently observes the test session, except for issuing occasional reminders to “keep talking” (Ericsson & Simon, 1980). However, since the conception of TTA, an alternative approach has emerged. In Interactive Think-Aloud (ITA), instead of remaining silent, the moderator makes interventions while the participant is using the system, such as asking questions of the participant. ITA interventions are an attempt to get more useful

TA data but have been shown to affect participant behavior, and the presence of this reactivity compromises the validity of the usability test (Hertzum et al., March 2009; Olmsted-Hawala et al., 2010; Alhadreti & Mayhew, 2017). Despite this issue, ITA has become the most popular usability testing approach among usability professionals (McDonald et al., 2012).

Studies comparing the usefulness of ITA and TTA data have delivered mixed results (McDonald et al., 2016) and have faced validity problems of their own (Hornbæk, 2010). Many studies have assessed ITA in experimental conditions, yet few have investigated the views of practitioners. We report a study exploring the perspectives of nine user experience (UX) practitioners in depth to provide a current view of how and why ITA is used. We investigated the practitioners' views about the usefulness of ITA data, their attitudes toward reactivity and validity, and other positive and negative factors which they consider when using ITA. By gathering rich qualitative data, we have tried to build a nuanced picture of why so many practitioners adopt ITA despite its well documented problems.

Origins of Think-Aloud

The original TTA protocol was developed from protocol analysis by Ericsson and Simon (1980) within the context of cognitive psychology before it was applied to usability testing by Clayton Lewis (1982). In TTA, the user is asked at the start of the session to "say out loud the things you normally say to yourself silently." In Ericsson and Simon's model, TA data (that is, what the participant says while thinking aloud) is divided into three levels. Level 1 and 2 verbalizations are based on information accessed directly from the participant's short-term memory as part of task completion. In contrast, Level 3 verbalizations involve additional cognitive processes not required for task completion. According to Ericsson and Simon, Level 1 and Level 2 verbalizations are the most valid forms of TA data. Level 3 verbalizations involve too much cognitive processing in addition to that required for task completion, so they should be avoided. Interventions from the moderator, such as probing questions, elicit undesirable Level 3 verbalizations, so Ericsson and Simon advise that moderators do not intervene except to issue a reminder to "keep talking." Despite the origins of TTA in the relatively narrow field of cognitive psychology, Ericsson and Simon's model is still widely referenced as the theoretical basis for the use of TA in usability testing.

Rise of Interactive Think-Aloud

Since the conception of TTA, a significant divergence between theory and practice in usability testing has emerged. Boren and Ramey (2000) were among the first to document this after they observed that practitioners "often intervene in theoretically inconsistent ways" (p. 265). Interventions observed by Boren and Ramey included, for example, requests for clarification and prompts for reflection. Boren and Ramey believed that, in a clear rejection of Ericsson and Simon's advice, practitioners delivered these interventions deliberately to elicit Level 3 verbalizations, which they seemed to find more useful than the Level 1 and 2 verbalizations that TTA produces.

Although Boren and Ramey's work is now quite dated, more recent studies show that Interactive Think-Aloud (ITA) (also sometimes called Relaxed, Active Intervention, or Talkative Think-Aloud) is still widely practiced. McDonald et al. (2012) conducted an international survey of 207 usability practitioners and found that about 70% practice ITA at least some of the time. Furthermore, when they intervene, ITA practitioners intervene a lot: Hertzum and Kristoffersen (2018) calculated an average rate of 1 word spoken by moderators for every 5-7 words spoken by users. The popularity of ITA is unsurprising given that ITA is encouraged by many practitioner websites (such as the website of Nielson Norman Group) and books (such as *A Practical Guide to Usability Testing* (Dumas & Redish, 1999)). For example, in an article published by Nielsen Norman Group, Pernice (2014) claims that "sitting completely mute… is not an advanced facilitation practice as it doesn't enable gathering the most possible information during a study;" instead, Pernice advocates for a form of ITA which involves "probing at the right times" (Facilitation Techniques for Handling Test Users' Questions During Usability Studies section).

Although the practitioner literature provides lots of advice on how to intervene, there is no standard protocol for ITA, and these articles sometimes contradict each other (McDonald et al., 2016). For example, Makri et al. (2011) recommend use of the word "why," whereas Dumas and Redish say that moderators should avoid this word as it may imply judgement. ITA practice itself also likely varies depending on individual, organizational, and cultural factors (Hertzum et al., July 2009). This lack of ITA standardization sits within the wider context of huge variability in many aspects of usability testing practice. The Comparative Usability Evaluation (CUE) studies detail many of these inconsistencies, which highlights disparities in how practitioners design tasks and report findings and in the problems they identify (Molich, 2018).

Separately, there is some inconsistency in the literature in how aspects of ITA practice are described, especially the categories of interjection made by the moderator (which we refer to as intervention types throughout this study). For example, Hertzum and Kristoffersen (2018) describe the moderator verbalization itself ("task instructions") whereas McDonald et al. (2012) sometimes describe situations ("user is stuck on a task") and sometimes describe the objective of the intervention ("to understand the impact of a problem"). As a result of these issues, the literature provides an overall account of ITA practice that is somewhat ill-defined.

Few studies exploring ITA have involved ITA practitioners themselves, but there are three notable exceptions that provide the best account of how and why ITA is used in practice. The survey conducted by McDonald et al. (2012) included a question that asked practitioners about the situations in which they would intervene. They asked about practical interventions necessitated by the "contingencies of usability testing" as Boren and Ramey (2000, p. 271) described them, and they also asked about probing interventions aimed at getting more useful TA data. Their findings detail the popularity of different intervention types within these broad categories. The second study that provides a good account of ITA practice involved the observation of 12 ITA sessions (Hertzum &

Kristoffersen, 2018). The third study of note is the 10th of the CUE studies, which involved getting usability professionals to comment on each other's usability testing moderation approaches (Molich et al., 2020). These studies' findings are discussed later in the context of our own, but it should be noted that some of this work is now quite old. The majority of UX professionals have less than 10 years of experience in the field (Krause & Rosala, 2020). Therefore, since McDonald et al. conducted their survey in 2012, a whole new generation of usability practitioners has entered the industry. It is possible that practices and attitudes have changed, and this was a key motivation for the study reported here.

Validity of Usability Testing Data

The theoretical underpinnings of usability testing have traditionally been firmly positivist, based on the idea that there is a set of usability problems out in the world that can be identified upon valid application of the method. Accordingly, usability testing is not simply aimed at identifying usability problems that *only* occur during usability studies. The study must have ecological validity, which is to say, "test performance predicts behaviors in real-world settings" (Barker et al., 2014), if the findings are to be useful. Therefore, researchers aim for participant behavior in the study to be as close as possible to user behavior in the real world. Any behavior that is "a reaction to being tested" is known as participant reactivity (Oates, 2012, p. 132). The presence of reactivity may affect the findings of the study. More specifically, problems identified in usability testing may not be real problems that would occur outside test conditions (false positives), and some real-world problems may not be observed in the study (false negatives). Under the Ericsson and Simon (1980) model, the moderator should stay mostly silent because interventions are likely to result in Level 3 verbalizations. These require cognition in addition to that needed for task completion, which constitutes reactivity and so compromises the validity of the study. The issue is not with the validity of the Level 3 verbalization itself, but with the effect of the additional cognition on future participant behavior, verbalizations, and task performance.

This reactivity is not just a theoretical risk; it can be measured in task performance metrics such as task completion rates and task times (Sauro, 2010). Several experimental studies have shown that ITA causes reactivity, affecting task performance compared to TTA and silent controls. For example, Hertzum et al. (2009) found that ITA altered participant behavior and resulted in a higher mental workload compared to TTA and a silent control. Olmsted-Hawala et al. (2010) ran a similar study and found that ITA led to higher task completion rates. Alhadreti and Mayhew (2017) compared TTA and ITA (and Speech Communication TA) and found that ITA resulted in a higher number of mouse clicks, a higher number of pages viewed, and longer task times. It is, therefore, reasonable to conclude that ITA causes reactivity, and that it does so to a greater extent than TTA.

Usefulness of ITA Data

Practitioners might adopt ITA despite its validity issues because they believe it improves the usefulness of the data for formative evaluation. Ericsson and Simon (1980) admit that TA data often lacks certain features of communicative speech, making it difficult to interpret. Therefore, ITA might help practitioners make sense of otherwise ambiguous verbalizations. Additionally, ITA may provide practitioners with a different and more useful kind of data, such as Level 3 verbalizations. Several studies have attempted to settle the issue by experimentally comparing the usefulness of ITA and TTA data, but these have had mixed results. Alhadreti and Mayhew (2017) found that ITA resulted in the identification of a similar number of usability problems as TTA. However, McDonald et al. (2016) ran a similar study and found that ITA resulted in the identification of *more* usability problems (although these were mostly low severity). The problem-counting approach used in these studies has been subject to heavy criticism for several reasons (Hornbæk, 2010). For example, it often gives equal weight to different kinds of problems, it ignores the usefulness of an identified problem for redesign, and it fails to account for anything a practitioner might learn about usability that cannot be captured neatly in the description of a problem.

Summary

Although the presence of ITA reactivity is well established, experimental studies have failed to show definitively whether ITA improves the usefulness of usability testing data. For several decades, usability researchers have been out of step with practice, advising against ITA even though the approach remains ever popular. One possible explanation for this divergence is that the academic community has failed to truly understand the perspectives of practitioners. ITA has been studied extensively in decontextualized experimental conditions, but few studies have involved ITA practitioners themselves. Those that have done so have often been observational and, as a result, practitioner attitudes to ITA are especially under-researched. In particular, we know relatively little about *why* practitioners use ITA, the extent to which they consider reactivity and validity, or what issues *other than reactivity* they might experience while using ITA. The study we report here is the first step toward addressing this gap in the literature.

Method

Study Design

We conducted nine in-depth semi-structured interviews with usability testing practitioners. The interviews were aimed at answering the following research questions:

- RQ1: Why do ITA practitioners use ITA?
- RQ2: What challenges do practitioners experience when using ITA?
- RQ3: How do practitioners view ITA's reactivity and validity problems?

We chose interviews as a method because they allow for the researcher to adapt the line of inquiry to facilitate deep exploration of attitudes and beliefs. Unlike observational studies, in which attitudes must usually be inferred, interviews provide the required

attitudinal data directly.

Participants

Participants all met the following criteria:

- Had at least 6 months of usability testing experience
- Had conducted usability testing in the last 6-months
- Practiced ITA at least some of the time
- Were over 18
- Were not considered vulnerable

A convenience sample was recruited by distributing a screener questionnaire to the first author's academic and industry contacts on online platforms, LinkedIn® and Slack®. Table 1 reports the characteristics (collected through the screener) of the participants who were recruited to the study.

Table 1. Participants and Their Characteristics

ID	Which option best describes your current role? [multiple choice]	What industry / sector do you work in? [text box]	How much experience conducting usability testing do you have? [multiple choice]	When was the last time you conducted usability testing? [multiple choice]	Job title	Highest level of education relevant to HCI (HCI, Psychology, UX Design, etc.)
P1	In-house	Fintech	3-5 years	In the last 6 months	Senior UX Researcher	Master's Degree
P2	In-house	Health sector	3-5 years	In the last 6 months	Senior UX Researcher	Master's Degree
P3	In-house	E-commerce	3-5 years	In the last 6 months	Senior UX Researcher	Industry Certification /Bootcamp / Short Course
P4	In-house	Health (NHS)	6-9 years	In the last 6 months	Senior User Researcher	Batchelors Degree
P5	Agency	Public Sector (but employed by a consultancy agency)	1-2 years	In the last 6 months	Senior User Researcher	Industry Certification /Bootcamp / Short Course
P6	Agency	Digital consultancy working on government projects	Between 6 months and 1 year	In the last 6 months	Senior UX Researcher	Industry Certification /Bootcamp / Short Course
P7	Agency	Digital Design (mainly gov/public sector)	1-2 years	In the last 6 months	Senior User Research Consultant	Batchelors Degree
P8	In-house	Energy / Tech	3-5 years	In the last 6 months	Senior User Researcher	Master's Degree
P9	In-house	Energy	3-5 years	Between 6 and 12 months ago	UX Researcher	Industry Certification /Bootcamp / Short Course

With an average of around 3.5 years of usability testing experience, participants were broadly representative of UX practitioners in industry at the time of this writing, over half of whom have less than 5 years of experience in the field (Krause & Rosala, 2020). Participants had attended a mixture of university-based and industry-based educational programs, which also makes them broadly reflective of the wider practitioner population (Krause & Rosala, 2020). All participants worked in the UK.

We ran our analysis concurrently with data collection, and the sample size was partly determined by the analysis. Rather than aiming for full data saturation, Braun and Clarke's (2021) objective of Information Power was used, that is, the data collection was concluded once the data seemed capable of answering the research questions.

Procedure

The study received delegated ethics approval from City, University of London, and participant consent was obtained using an online consent form. We prepared a discussion guide in advance. We phrased the introduction and questions carefully to limit the risk that participants would feel judged or assessed. Although the validity concerns associated with ITA are well understood in the academic community, we did not know whether practitioners would be familiar with these issues. We therefore structured the guide to begin with a broad exploration of how and why participants use ITA (RQ1). This was followed by a more reflective section exploring any challenges the participant associated with ITA (RQ2). For participants to talk freely about their use of ITA, it was key that the interviewer did not impose any views from the literature that the participant did not already hold themselves. In practice, this meant that if ITA's validity issues were not raised by the participant, then we delayed questions about validity until the end of the interview (RQ3).

The first author conducted an initial pilot with a single participant (P1), and then adapted the discussion guide slightly as a result. We deemed the data from this interview relevant enough to be included in the analysis. The first author then conducted the remaining interviews over a 4-week period (07/19/2021–08/20/2021), iterating the discussion guide several times to better target emerging themes and perceived data gaps.

The first author conducted the interviews remotely using video conferencing software. Both interviewer and participant had their cameras on, and we recorded audio. The first author transcribed the recordings and analyzed the transcripts using NVivo™ broadly following the classic Thematic Analysis approach described by Braun and Clarke (2012). The coding was primarily inductive, based on terminology and concepts present in the data. There were also elements of deductive coding, as concepts from the literature (such as participant reactivity) informed the analysis. The first and second authors discussed the codes as they evolved and then grouped the final set of codes into themes that directly answered the three research questions. We also identified intervention types in the coding and documented these in a tabular format. Our approach used for intervention

types was adapted from that taken by McDonald et al. (2012) and involved describing both the situation in which a practitioner would intervene and the purpose of the intervention.

Results

Nine themes emerged from the thematic analysis and are discussed in our results organized by research question.

RQ1: Why Do ITA Practitioners Intervene?

Theme 1—Getting More Useful Data: "You just won't get data unless you interrupt."

Unsurprisingly, the main reason participants in this study gave for using ITA was to get more useful TA data. We identified many specific intervention types aimed at getting more useful data (Table 2), although we did not attempt to capture an exhaustive list.

Table 2. Intervention Types: Data Usefulness

Situation in which practitioner would intervene	Purpose of intervention
User is not saying much.	To get the user to think aloud
User is providing mostly procedural verbalizations.	To get the user to provide the 'right kind' of TA (often Level 3 TA is desired)
User is using the system too quickly.	To slow the user down so the moderator can follow what is happening and potentially deliver additional interventions
User has not engaged with a feature / screen of interest.	To ensure the participant engages with the feature / screen so data can be collected about it
User is about to interact with a key or potentially problematic feature.	To elicit user expectations for comparison with the actual system state
User does or says something that indicates there is a problem.	To understand the problem and its cause more completely
User does or says something unclear or unexpected.	To understand the behavior or verbalization
User has encountered an important screen / feature.	To test the user's understanding of the screen / feature by asking them to explain it
User is acting without a clear purpose.	To understand what the user is trying to do
User strays from the task or appears unfocused.	To refocus the user on the task
User seems dissatisfied with functionality or content offered or makes a feature suggestion.	To understand a user need / requirement
User has encountered a page of interest.	To get the user's impression of the page
User asks a question about the system.	To redirect the user back to the task / to challenge the user to answer the question themselves

The perspectives of participants often added nuance to findings about intervention types reported by other studies. For example, like McDonald et al. (2012), we found that practitioners may intervene to direct users toward features of interest: "If the usability testing is focusing on… the [filtering] functionality, then I need to lead them" (P3). We found that although a particular feature of the system might be a research focus (or even *the* research focus), it is often a much broader journey that is actually tested. Therefore, there is a risk that users "skip over those bits" (P7) that are of interest, or simply "spend too much time on figuring out where" the relevant functionality is (P3). Accordingly,

practitioners intervene to ensure they get answers to their research questions about these features within the time available. This intervention may take the form of a direct instruction: “Here, you need to press here” (P3). Or it may be more subtle: “I’d be like, ‘Okay, anything else?’… I’d give them time to just find [the functionality]… And then I’d prompt them a wee bit more” (P5).

The elicitation of user expectations is another previously reported intervention type, according to Hertzum and Kristoffersen (2018). Interestingly, most participants in this study claimed this was among their most useful interventions. They said that they elicit user expectations for comparison with the actual system: “That helps [identify that] users have got these expectations, but we’re not meeting them because, actually, *this* is what happens instead” (P7). Mismatches between mental models and the system are used to inform alternative designs that better match user expectations. Participants believed it was essential to deliver this intervention before the user moved onto the next screen when they “haven’t been there yet and so they don’t know what is on the other side” (P8). Otherwise, there is a risk that users report what *actually happened* rather than their true prior expectation; the user’s retrospective report of their expectations can suffer from hindsight bias.

Some participants said they intervene for a reason that has not been reported in the existing literature—to slow the user down: “We have to stop people at key parts in the journey just to take a bit of a breath and try and get a little bit more of the finer detail before we move on with that” (P7). For some participants, this helps them understand behavior “that’s useful for you to be able to see what they’re doing” (P5). In other cases, stalling users on a screen of interest may provide the opportunity for other interventions about that screen. These interventions may focus on aspects of the system that the user did not use (and did not need to use) to complete the task: “You might learn more about other aspects that they don’t need to consider as they go through” (P5). Digitally competent users are particularly likely to navigate through the system too fast and require slowing down.

Our findings add particular nuance to the idea that ITA helps practitioners *understand* usability problems to “find out the ‘why’” (P9). Participants expressed that, although it may be possible to identify problems without ITA, often an intervention is required to understand what is causing the problem: “Sometimes people might click on something and they’re like ‘oooh’… But then they don’t necessarily say to you what they thought was gonna happen” (P5). An improved problem-understanding from ITA facilitates formative evaluation by directly informing redesign: “When it comes to analysis, it’s not just like, ‘somebody got stuck, cool, now, what do we do?’ We’ve got the data to explore what we do about that” (P4). Participants were wary about inferring the cause of a problem from observation alone, that is, without asking the participant: “I might have an idea, but it’s *just my idea*” (P1). Probing the user for a problem explanation was seen as an empirical way of investigating problem causation.

Some of the intervention types in Table 2 are ultimately aimed at identifying more usability problems. However, many of the interventions that practitioners said were most useful were more focused on helping them *understand* usability, such as by eliciting a problem explanation or a user's expectation for how the system works. ITA data is seen as useful primarily because it facilitates this deeper understanding of problems and illuminates avenues for redesign.

Theme 2—Managing the Contingencies of Usability Testing: "You are constrained by the thing that you're testing."

Participants confirmed that, as Boren and Ramey (2000) first reported, ITA is also used to manage many practical aspects of usability testing (Table 3).

Table 3. Intervention Types: Practical

Situation in which practitioner would intervene	Purpose of intervention
User is stuck or asks for help.	To offer task assistance or move the user on to the next task
User has misunderstood the task or asks about the task.	To provide task instructions (sometimes rephrased)
User does not engage with a feature of interest.	To provide an additional task-like request aimed at getting the user to engage with the feature
User encounters a technical problem (like a bug).	To circumvent the bug and allow the user to proceed with the task
User encounters part of the system that will work differently in the final version.	To explain how the system is supposed to function
User is about to take an action in a live system that could have wider negative consequences (like deleting data).	To stop the user from taking the action

In one example of a practical intervention type, participants in this study reported that they often intervene to manage technical problems. This is perhaps partly because they test designs at a wide range of different fidelity levels. Especially in the early stages of the design process, this can include very basic prototypes that are only "partially clickable" (P3). Participants use ITA to explain missing functionality, such as by telling the user that "FYI, the prototype at this point, we know it wouldn't do this in real life" (P4). Participants believed that, by explaining gaps in functionality, they give the user an experience that is more representative of how the final system will work. With very basic prototypes, some participants said they sometimes have to drive the interaction, getting users to "sort of direct me to interact with the prototype" (P7); when the user is not directly using the system, an ITA dialogue is used to allow the user to vicariously navigate through the system.

Participants said they would use ITA to move users on when they get stuck. This was unsurprising, as 77% of respondents to the survey conducted by McDonald et al. (2012) said the same thing. Our participants described delaying their intervention for a while to see if the user can overcome the problem themselves. Watching a user struggle with an issue can provide useful data, and practitioners want to give users enough time to see if they can overcome issues themselves. However, at some point, it becomes "clear that they're just not going to be able to progress" (P8), and it is better to intervene "to just continue with the rest of the session" (P5). This aligns with Molich et al.'s (2020) finding that the practice of moving the user on when the usability problem becomes clear is associated with more effective management of session time. Participants in the present study also identified an ethical dimension to this issue: "We're not going to be like an arse about it… we'll help them out… We don't want them to walk away like feeling shitty about the fact that they got stuck" (P4).

Although Ericson and Simon (1980) provide no advice on how to manage these logistical problems, Boren and Ramey (2000) take the sympathetic view that some practical interventions are effectively unavoidable. However, we found that ITA is sometimes used to manage session logistics in cases in which better preparation might have been preferable. For example, participants described having to intervene to explain usability testing tasks:

If it was clear they just didn't understand [the task], I'll kind of reaffirm it, and if I thought they maybe didn't understand or maybe they were just struggling, I would kind of ask them to play back to me what I'd asked them to do. And then if it wasn't lining up, I would apologize and reframe it. (P8).

This aligns with Hertzum and Kristoffersen's (2018) findings: In their observation of ITA sessions, the provision of task instructions was the most common intervention type (32%). Participants in the present study also reported issuing impromptu task-like requests as interventions while the participant is completing a broader task. For example, they might say "Can you find this?" or "Where would you look for this?" (P3) or "How would you get back to where you were before?" (P5). This kind of extensive task-related dialogue might be avoided by preparing clearer tasks that are designed to more completely test the parts of the system the evaluator is interested in. Alternately, perhaps we should see these moderator verbalizations not as ITA interventions but as mini tasks in their own right. A similar stepped approach to presenting usability testing tasks is recommended in some of the practitioner literature (Pernice, 2020).

In alignment with Molich et al. (2020), we found that practitioners are sometimes unfamiliar with the systems they are testing, such that they may be surprised by what the user is able to do in the system during the session. This increases the need to use ITA to bring the user back to the task or to manage unexpected technical issues.

Theme 3—ITA as Part of a Wider Research Process: "You don't really have the luxury of five new participants every two weeks."

Getting useful data and managing session logistics both sit within a wider research process. The organizational context around usability testing in industry has rarely been a focus in the ITA literature, but we found that practitioners face a range of external pressures that influence their decision to use ITA.

For example, failings in the recruitment of a representative sample of users may encourage use of ITA. Recruitment is “a struggle for everyone” (P3) and appropriate users are like “gold dust” (P2). So once a user has been recruited, our participants wanted to “get everything we can from the person” (P3), and ITA was seen as a way of achieving this efficiently: “So that’s more justified for the budget” (P5). For participants, recruitment problems were often a motivation specifically for intervening to focus users on features of interest. This is because they were not able to recruit enough users to have “the luxury” of running sessions in which users do not engage thoroughly with all the relevant areas of the system. Boren and Ramey (2000) note that instead of using ITA, an indeterminate number of users could be recruited until the research questions were answered, but this is obviously not a practical solution for practitioners facing recruitment difficulties. Additionally, ITA may also be used to extract some value from unrepresentative users: “If [they’re] not the right customer… you can always get some knowledge from this person” (P3). When the user does not match the participant specification, structured usability testing plans may be disregarded in favor of free-flowing ITA: “So I might go off on a completely different kind of tangent there because I don’t want to waste a session” (P7).

All participants worked in multidisciplinary teams and frequently mentioned their colleagues (designers and product managers) and wider stakeholders (clients) as playing a role in their application of ITA. Interventions might be used to specifically “explore what the team asks us to answer” (P7), for example, to probe about a feature that the team is unsure about. Occasionally, other members of the team may be brought into the ITA conversation, especially on highly domain-specific projects. For example, one participant described how, in a usability test of a system designed for software developers, she allowed the team’s engineer to “interrupt because he might know better when we need to probe on something” (P9). Some participants implied that they felt it was a part of their role as the UX researcher in the team to intervene in sessions: “If I didn’t want to interrupt, then why wouldn’t I ask people just to record their screens? Like what’s the function of the presence?” (P3). ITA is seen as an innate feature of *moderated* (rather than *unmoderated*) usability testing: “If you’ve got a moderator, the point is that you’re there to *moderate a session*” (P7). ITA session facilitation is not something that just anyone in the team can do; it demands “quite a lot of skill” (P2) that requires “honing” (P8) through experience. Although no participants said this directly, it is possible that practitioners use ITA partly to demonstrate their specialist moderation skills to their colleagues and to justify their presence in the team as a professional UX researcher.

Theme 4—The Moderator-User Dynamic: “It’s got that kind of human element… that kind of two-way communication.”

Separate from concerns about data usefulness and session logistics, participants had strong views on the role that ITA plays in their relationship with their users. Most felt that ITA improves the experience for users, and that it specifically reduces test anxiety and builds rapport. ITA helps users “feel more comfortable because… it feels a lot more like a conversation” (P8). In contrast, silent TTA makes the moderator seem like a “kind of creepy, weird, passive observer” (P7). Rapport building through ITA seems to be a widespread and varied practice: Molich et al. (2020) reported that moderators may intervene to encourage users when they self-blame, to laugh with the user, and to compliment the user. Some participants in the present study got personal satisfaction from making the session enjoyable for the user: “I really like getting that nice, positive feedback from users. They’ve enjoyed chatting… They found it interesting (P7)”. The experimental literature presents a mixed picture regarding whether users prefer ITA. In their comparison of different usability testing methods, Alhadreti and Mayhew (2017) found that users involved in ITA studies considered the evaluator more of a distraction than in TTA or Speech Communication TA. However, in a similar study, Zhao and McDonald (2010) found that 17 out of the 20 users preferred ITA over TTA because it felt more relaxed or natural. High levels of variation in ITA styles may explain this inconsistency.

Several participants reported that ITA allows them to shift the moderator-user dynamic from one of an observer and an observed subject to something more interactive, similar to a semi-structured interview. In these cases, the system can become an “interview kind of stimulus” (P8) and the session becomes a conversation about the system “rather than just trying to observe them use it as they would naturally” (P8). Interestingly, two participants (P3, P2) used the terms “usability testing session” and “interview” interchangeably. The back-and-forth “question and answer” (P7) dynamic of ITA certainly shares many qualities with interviews. The idea of the usability test as an interview can also be found in some of the usability guidebooks. For example, in *Observing the User Experience*, Goodman et al. (2012) describe usability tests as “structured interviews focused on specific features in an interface prototype.” Therefore, ITA could be characterized as a hybrid approach sitting somewhere between usability testing and semi-structured interviewing.

In another extension of the traditional moderator-user dynamic, some participants saw ITA as a participatory approach: “It’s got that kind of human element. It’s got that kind of two-way communication. It’s got that sense of exploring something together and trying to understand a little bit more together” (P7). Practitioners described introducing elements of “codesign” to the usability test to benefit from the user’s creativity to “get ideas for content and features” (P2). These participatory elements were also thought to improve the user experience: “Especially if it’s an hour, which can be incredibly taxing on the participant’s brain as well, it’s good to have as many different kinds of brain teasers or exercises as possible” (P9). This participatory reframing of the usability test would certainly not be acceptable under the Ericsson and Simon model. However, the approach resembles

other widely accepted (though not so widely used) participatory evaluation approaches like Cooperative Usability Testing (Frøkjær & Hornbæk, 2005), the Pluralistic Walkthrough (Bias, 1994), and the Cooperative Evaluation (Wright et al., 1991).

RQ2: What Challenges Do Practitioners Experience When Using ITA?

Theme 5—The Risk of Overly Reflective TA: "They kind of see it more as a design crit."

Although participants generally valued the reflective nature of ITA data, they also believed that ITA could backfire, leading users to provide undesirable design feedback: "I think probing a lot can have the cumulative effect of encouraging people to give you a critique or review rather than just *use the system*" (P8). ITA may encourage users to "see their role in the session as the expert who should be telling you what to change and how" (P5). Superficial comments about "the colors of the website and the wording that is used" (P1) were especially uninteresting to participants. Some users may even provide feedback on elements of the design that they personally find unproblematic. This can then leave the practitioner with a challenging analysis task of "distinguishing between people… using the system and… giving us design improvements, even though they understand things" (P8). If a participant starts "design critting away" (P8), another intervention may be required to indicate that the moderator is not interested in this kind of TA to bring the user back to talking about "how you use the website and kind of how it works for you" (P8).

This negative perception of user design feedback contrasts with McDonald et al.'s (2012) finding that practitioners are likely to intentionally gather design feedback by intervening to elicit a user's opinion (61%) or to seek a design recommendation (45%). In fact, there was a prevalent belief among participants in the present study that "it's not the participant's task to redesign something" (P9). The user is there to "tell us or to show us what their needs are" (P8), but they don't have the skills to make good redesign proposals: "They're less good at identifying the solution for that because they aren't typically designers" (P8). Perhaps we are seeing an evolution in practice as practitioners become more comfortable identifying problems through usability testing and designing solutions, rather than eliciting design proposals from users. This approach notably reflects the design-thinking doctrine that separates problem identification from solution ideation (Interaction Design Foundation, 2022).

Some participants even believed users' design suggestions were dangerous, as they might be overvalued by stakeholders: "If the participants tell [us] that they want this button to be pink, because it's just on their minds, people will take it very seriously and design a pink button" (P3). Where design suggestions *were* deemeduseful, it was because they helped participants understand an underlying problem: "It's not about [the user's suggestion] being a solution, but it's about understanding why they think the current as-is model doesn't work" (P7).

There is an obvious tension between these findings about participants' reluctance to capture users' design suggestions and the finding in Theme 4 that practitioners may use ITA to introduce participatory elements into usability testing sessions. This tension partly reflects wide variation in how and why ITA is used by different practitioners. However, as some participants expressed both views (P7 and P9), it also implies that individual moderators may value design feedback in certain contexts but not in others.

Theme 6—Challenges in Team Working: "Everybody's got a different style."

ITA can also lead to challenges in team working. Some participants said they got their team to take notes using structured note-taking templates during usability testing sessions. However, the unpredictability of ITA can make things difficult for session notetakers. When the ITA moderator "go[es] off script," note-takers may have to "scrabble around a little bit" (P7) to work out how to fit their notes into an existing template. Some note-takers who are less familiar with ITA might not record user responses to ITA interventions: "If it's not in that template, they think, 'Well, I'm not going to write it down because that's not what they want to know'" (P7). One participant said that, ideally, they try to "mirror" planned interventions in their note-taking template so that it's "really clearly delineated in the notes what we've asked people and what people have just organically talked about" (P8). Given that session notes (rather than recordings) are the primary data submitted for analysis by practitioners (McDonald et al., 2012), this delineation in the notes is extremely important.

In rare cases, ITA may give team members observing a usability test "the impression that they can just interrupt and ask questions" (P4). This can "completely invalidate the session or derail [it]" and lead to tensions within the team. Even when multiple experienced ITA practitioners collaborate, the fact that "everybody's got a different style" (P3) makes it difficult to standardize ITA sessions moderated by different practitioners. Differing types and frequencies of interventions are understood to affect user behavior and to lead to inconsistencies in the type of data collected. One participant reported that they plan interventions in a shared session guide to help mitigate this issue.

Theme 7—ITA Can Become an Interrogation: "They feel their behavior is being questioned."

Although most participants felt that ITA improves the user's experience of participating in usability testing, they also noted that bad ITA practices could make the user feel uncomfortable (see Theme 2 for a discussion of the experimental literature on this topic). In particular, too many questions, inappropriate questions, poorly timed questions, or an inappropriate tone may make users feel under "interrogation" (P5), or that they are being "examined" or "assessed" (P6). One participant said that poor ITA can make the user "feel like [they are] at school… feel that they are doing something wrong" (P3). Molich et al. (2020) also reported similar risks associated with poor ITA practice and found that even well-intentioned affirmations (such as "great" or "ok") may be seen to cause friction if they

are perceived as trivial or routine. All moderated research involves a power imbalance between researcher and participant, and it is possible that by increasing interaction, ITA exacerbates this imbalance compared to TTA.

However, participants were confident that this risk can be avoided by ensuring that one does not “interrupt in a wrong way” (P3). This problematic style of ITA was much more closely associated with less experienced researchers: “I think if you’re an experienced researcher, then you don’t make it seem like an interrogation. Whereas sometimes if you’re doing some with a junior member staff, it will sound like a survey” (P5). Delivering ITA without causing the user undue distress is seen as a skill that must be refined through experience.

RQ3: How Do Practitioners View ITA Reactivity and Validity?

Theme 8—Managing Reactivity: “There is probing and then there’s over-probing.”

In some earlier studies, researchers have assumed that ITA practitioners have little understanding of the reactivity caused by their interventions (Boren & Ramey, 2000). However, like McDonald et al. (2012), we found that practitioners often *are* aware of reactivity. In this study, they often described it as impacting the “realism” or “authenticity” of the usability test. There was an understanding that interventions have an impact on what “naturally would have happened if the user [had] been left to their own devices” (P7). As a result, some participants advised against using ITA in quantitative, summative, and later stage (beta) usability testing: “You have to use your judgment a bit and balance out what stage of design you’re at… and what risk you introduce [by] directing somebody” (P8).

However, participants generally saw ITA reactivity as something that should be managed rather than eliminated and that must be weighed against the perceived benefits of intervening. They achieve this by applying professional judgement “as a researcher, honing that skill of realizing when you’re going to be leading somebody… to balance out how much you need to understand and what risk you have by introducing that bias” (P8). This has echoes of Boren and Ramey’s (2000) advice that intervening to answer a particularly important question might be justified if one is prepared to do so “at the expense of less critical research questions later on” (p. 275). Participants were concerned that inexperienced moderators lack the judgement to use ITA appropriately and “interrupt in a wrong way” (P3). Again, good ITA moderation was seen as a practice that requires high levels of specialist skill.

Participants noted several specific approaches they use in an attempt to limit reactivity. For example, they might avoid leading language in their interventions, especially language that references the UI “like calls to action [or] content headers” (P2). P5 described a time when she wanted to draw a user’s attention to a button labelled Print and did this “not with the same words” as in the label but by asking “how would you keep a copy of that?” Similarly, Molich et al. (2020) found that practitioners advise against probing with closed or leading questions such as, “Did you find the European time format

confusing?" To an extent, all of this advice aligns with Boren and Ramey's (2000) recommendation to only intervene with minimal probes that do not introduce new concepts.

Some participants said that they try to deliver interventions in "natural breaks in the flow of the journey" (P9), especially when the user was "already in the mindset of exploring" (P4). They wanted to keep the user "focused, in the moment… in the context" and there is a danger that poorly timed interventions disrupt task focus (P4). Participants were particularly concerned that intervening "mid-thought process" might make the user "forget something that they were going to say" (P7). Some said they withhold interventions while a user is focused on the task, and then follow-up with their questions at the end of the session:

Sometimes I make a note of something and ask at the very end. So that's very, very common… something that they mention that is interesting and we want to follow up on, but it wouldn't have made sense to interrupt the flow (P9).

Although no participant mentioned Cooperative Usability Testing (Frøkjær & Hornbæk, 2005) by name, this intervention delaying approach sounds very similar. Some participants saw the methodological benefits of delaying interventions until the end but did not see this as a practical alternative to ITA:

So, in some ways, it would be really nice to have a really, really long usability… just do one run-through with everybody a bit more naturally, and then kind of go back through it with a fine-toothed comb… as much as we'd love to do that, there's not the time – not to collect the data, nor to process it, often (P8).

Theme 9—A Relaxed Attitude to Validity: "As long as you don't lead people… I don't think it's a big drama."

Practitioners held mixed views about the validity of ITA data. Interestingly, some argued that ITA *improves* validity. For example, one said that "using an interactive style with think-aloud, for me is helping maintaining a kind of realistic mindset… bringing a person back… to the task that they need to complete" (P1). As discussed previously, ITA is also thought to reduce test anxiety, thereby improving the realism of the study. Other participants believed that reactivity reduces data validity but viewed this issue with a degree of flippancy. The validity of data was described as "not a massive concern" (P7), not a "big drama" (P8), or not something to "worry so deeply" about (P3).

This attitude is partly explained by the fact that participants thought that they could compensate for validity issues through a variety of mitigating approaches. For example, participants described accounting for ITA reactivity in their reporting, making it clear that an "insight has come from direction rather than organically" (P8). Or they may account for it in their analysis. For example, if a user struggles to find a feature *even* when nudged toward it by the moderator, "then you're kind of like, 'okay, *that's problematic!*'" (P5). Here the practitioner believed they understood the reactivity caused by their intervention (such as making a feature easier to find), so they took that into account in their analysis (such

as concluding that the feature is *particularly* difficult to find). Similarly, one participant noted that when she documents task success/failure, she includes a “they needed prompting” (P8) option in addition to the traditional pass and fail options. The fact that a participant completed a task with assistance still tells the evaluator *something* about the usability of the system, even if it is not as clear-cut evidence as a regular unaided task completion. Other participants triangulate usability testing with other methods to compensate for ITA’s validity issues. Web analytics were mentioned as a particularly effective way of testing hypotheses developed from ITA data. Analytics can provide the behavioral “data from how [users] interact with the page in real life” rather than in the “artificial scenario” of an ITA session (P5). Other contextual methods were also mentioned, such as intercept surveys and follow-up interviews. If ITA cannot provide completely valid data about user behavior, practitioners have other, perhaps better, methods for collecting this behavioral data.

Another more profound motivation for this relaxed attitude to validity is a belief that *all* usability testing suffers from validity issues, whether or not ITA is used. In their defense of ITA, participants described several of these issues; for example, P9 cited the “Hawthorne effect [which is] always there… even if I shut up, [because] I’m still there” (P9). Participants also noted that users are more likely to persevere with a task in testing than they normally would: “In real life they might have tried once and gone ‘oh this is a piece of crap’ and given it up” (P5). The decontextualized nature of lab testing was also noted; for example, users may complete a task in a focused research session that, in real life, they would do when “they’ve got a spare moment in the evening in between feeding the kids and giving them a bath” (P8). Practitioners also held the classically interpretivist view that, even when test conditions are ideal, “there’s no one way of conducting objective true research” (P3). Together, these issues contribute to “the actual artificial nature of the usability test” (P9). This innate artificiality constitutes such a low baseline for the realism of the study that, in the eyes of participants, ITA has a negligible impact on overall data validity. The implication is that if user behavior in usability testing can never *really* represent user behavior out in the world, then moderators might as well use ITA to get the data they need.

In addition, participants challenged the traditional conception of qualitative usability testing as a primarily *behavioral* study. When asked about how the realism of the study affects the usefulness of the data, P8 said, “Because we’re looking for an ability to iterate the design when we’re doing qualitative usability studies, it [the realism of the test] doesn’t matter too much.” Valid behavioral data is knowingly sacrificed in favor of rich ITA data:

You can see what the person would do in a test environment, but you’re not so sure that’s what they’d do in a real environment if you weren’t there… So I think you can learn about their understanding of something, you can learn about their perception of something, but you can’t necessarily learn about what their natural behavior would be if they weren’t being prompted (P5).

For participants, this is an acceptable sacrifice because they can arrive at the insights they need about the user's "understanding" and "perception" of the system without the hard behavioral data traditionally associated with usability testing. Participants were open about their focus on a softer form of qualitative data. One claimed that ITA sessions were "not scientific… not able to give us 100% truth and concrete answers" (P9), and that they were instead meant "just to give us guidance… to inspire us" (P9). The collection of non-behavioral, self-reported data is understood to carry certain risks, but "as long as you don't lead people" (P8) and "you bear in mind that bias and what kind of data you're getting" (P9), this data is still deemed to be valid and useful. Indeed, within interpretivist schools, it is widely accepted that a skilled moderator can apply good practice to limit bias whilst using non-behavioral methods (such as interviews) and can obtain valid data as a result (Goh, 2020). As discussed in Theme 1, it is precisely this reflective ITA data that provides participants with the rich, nuanced understanding of usability problems that they require. Observing user behavior is certainly useful but, for practitioners, it is often of secondary importance to the collection of ITA verbalizations.

Conclusion

This study is the first to report in detail on the views of practitioners regarding their use of ITA. Overall, participants saw many diverse advantages in ITA, but unsurprisingly, getting more useful data was seen as the key benefit. In particular, participants used ITA to help them *understand* usability problems, rather than to identify a greater number of problems. This finding supports Hornbæk in his challenge to the dogma of usability research, which often assumes that better evaluation approaches simply identify more problems (2010). Studies comparing usability evaluation methods need to do more than this if they are to have relevance to practitioners. They must find ways of measuring problem-understanding and the impact of findings on the redesign process if they are to provide meaningful metrics for the usefulness of ITA data.

The perceived practical advantages of ITA were also a major factor for participants. Other methods do not offer the same flexibility to deal with limited numbers of users, unrepresentative users, basic prototypes, technical problems, and organizational pressures. This flexibility can even extend to the repurposing of the usability testing session from a primarily observational study to something more like a semi-structured interview or a participatory usability evaluation. Practitioners facing the challenges of doing research in industry require highly adaptable methods, and it is difficult to see how more rigid usability testing approaches can compete with ITA on this front.

Despite their preference for ITA, participants recognized that it comes with disadvantages. These included the complexity of coordinating with other ITA moderators, the risk of overly reflective ITA and the threat of bad ITA moderation to the participant experience. They were also aware of ITA reactivity and often took steps to intervene without causing too much of it. However, not all participants thought that reactivity was a problem for the validity of their data, and those who did were not as concerned as the literature suggests they should be. Participants often believed they could compensate for validity problems in

their analysis and reporting or through triangulation with other methods. However, they also rejected the idea that the observation of wholly realistic user behavior was needed to deliver valid, useful findings about a system's usability, and they did not think that usability testing could provide this kind of valid behavioral data anyway.

Some of the concerns that participants raised about the objectivity of usability research are reflected in the wider literature. In particular, the 10 CUE studies have documented a wide disparity in the findings of usability tests conducted by different practitioners (Molich, 2018). This has led some to question whether reliability should be a goal of formative usability research (Sauro, 2018). Additionally, some have presented a challenge to the classical positivist conception of usability testing by casting doubt on whether, given the contextual nature of usability, there is even a definite set of real usability problems that a well-designed study could uncover (Wilson, 2007).

The notion that usability research is in crisis (Alhadreti & Mayhew, 2017) is spurred on both by these reliability and theoretical problems and by the apparent irrelevance of the think-aloud literature to the majority of practitioners who continue to ignore its recommendations. However, between these two enduring issues there is an opportunity for reconciliation. Perhaps ITA practitioners are not concerned about reliability and external validity because they simply do not attempt to capture wholly realistic behavior or to identify usability problems from a definite real-world set. Instead, they focus on the broader and more achievable goal of building a deep qualitative understanding of how users experience the system, and they do this outside of the confines of purist observational research, instead adopting a hybrid method that combines semi-naturalistic observation, TTA, prompted elicitation, and semi-structured interview.

Although this conception of ITA may be a significant departure from traditional usability testing, this does not mean that ITA is a bad usability evaluation method. It is just a different kind of method, theoretically closer to those from interpretivist schools (interviews, contextual enquiries, etc.) than the Ericsson and Simon (1980) model and the classical positivist conception of usability testing. ITA is not a replacement for TTA, but there are some circumstances (methodological, organizational, and practical) in which ITA may be a better choice for practitioners than the traditional approach. These might include early-stage studies, tests of extremely limited prototypes, tests involving unrepresentative participants, tests of highly technical or domain-specific systems, mixed-methods studies that collect behavioral data in other ways, and studies focused on building a deeper or more contextual understanding of usability and usefulness.

Limitations

The qualitative approach employed in this study allowed us to explore the attitudes of the participants in detail. However, it did come with some limitations. Conclusions are not necessarily generalizable, and they are vulnerable to sample bias toward practitioners working in the UK. As with many interview-based studies, there is a risk that some findings are idiosyncratic, that some important attitudes were not captured, and that other researchers would have come to different conclusions from the same data.

Recommendations for Further Work

We have several recommendations for further work:

- The ITA literature is mostly focused on discouraging ITA use, so it provides little advice on how to apply ITA effectively. Rather than continually (and ineffectively) advising practitioners to moderate in silence, usability researchers should instead define an ITA protocol that provides advice on how reactivity can be limited through best practice intervention techniques and how more robust findings can be produced through a reactivity-aware analysis of ITA data. Researchers must look beyond the Ericson and Simon model and toward interpretivist theoretical frameworks, which are better aligned with the goals and priorities of ITA practitioners.
- The nuances of the improved understanding of usability problems that ITA provides are lost in the coarse problem-counting approaches employed by many comparative usability studies. These studies should instead explore the detail captured about usability problems and the extent to which this supports problem-understanding and facilitates redesign. Researchers could take a similar approach to McDonald et al. (2016) by assessing the usefulness of individual verbalizations (such as causal explanations) and counting their frequency under ITA and TTA. Other approaches might also be effective, such as taking a case study approach (Wixon, 2003) or focusing on the impact of identified problems on the redesign process (Hornbæk, 2010; John et al., 1996). Alongside data usefulness, studies should follow Alhadreti and Mayhew (2017) and measure other variables that are of interest to practitioners (such as the participant experience and session time).
- More work is required to assess the generalizability of our findings to other contexts. A quantitative survey, like that conducted by McDonald et al. (2012), might be valuable given that their survey is now quite dated, and that we report findings that were not covered by the options of its multiple-choice questions.
- The practice of triangulating using methods such as web analytics to compensate for the reduced validity of ITA studies could be a fruitful topic for further research. It would be beneficial to identify some case studies and to establish some principles for how these methods can be applied in a complementary way.
- Participants in this study thought they knew what reactivity was caused by their interventions and that they could draw valid conclusions about user behavior from ITA data by taking this into account. Future studies could experimentally investigate whether reactivity caused by certain intervention types is indeed predictable.

Tips for User Experience Practitioners

The following tips are based on our findings and provide pragmatic advice for practitioners who wish to use ITA while limiting its disadvantages:

- Plan and document the interventions you wish to use in the session. Avoid leading language or interventions that could be replaced with well-designed tasks. Consider how each intervention is likely to impact user behavior in the rest of the usability test and assess whether the value added by the intervention will be worth the reduced realism.
- If you are going to intervene about user expectations, do this before the user has completed the action you are asking about. Otherwise, there is a risk that they simply confirm what happened (hindsight bias).
- If a user gets stuck, do not intervene to help them straight away. Let them persist with the task for a brief period of time so that you can fully understand the usability problem. However, also be wary about causing undue distress and compromising the ethics of your research.
- Be aware that users' design suggestions could be over-valued by your team. If a user makes a design suggestion, use the suggestion to understand the underlying usability problem, but carefully consider whether to report the suggestion itself.
- When analyzing behavioral data from an ITA session, try to identify where reactivity might have occurred. Instead of taking the data at face value, think about how the user may have reacted to your interventions and take this into account when drawing conclusions. If possible, triangulate by using other methods such as web-analytics to get more valid data about user behavior in the real world.

Acknowledgements

We would like to thank the nine practitioners who took part in this study for their openness and enthusiasm when talking about their use of ITA.

References

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

Barker, A. A., Wilkes Musso, M., & Gouvier, W. D. (2014, October 28). Ecological validity. In *Encyclopedia Britannica*. Retrieved April 26, 2023, from https://www.britannica.com/science/ecological-validity

Bias, R. G. (1994). The pluralistic usability walkthrough: Coordinated empathies. In *Usability inspection methods* (pp. 63–76). John Wiley & Sons, Inc.

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261–278. https://doi.org/10.1109/47.867942

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper et al. (Eds.), *APA Handbook of Research Methods in Psychology* (Vol. 2, pp. 57–71). American Psychological Association.

Braun, V., & Clarke, V. (2021). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, *13*(2), 201–216. https://doi.org/10.1080/2159676X.2019.1704846

Dam, R. F. (n.d.). *The 5 stages in the design thinking process*. The Interaction Design Foundation. Retrieved July 2022 from https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process

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

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Frøkjær, E., & Hornbæk, K. (2005). Cooperative usability testing: Complementing usability tests with user-supported interpretation sessions. *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1383–1386. https://doi.org/10.1145/1056808.1056922

Goh, W. (2020, May 11). *5 types of biases in user interviews and how to mitigate them*. UX Collective. Retrieved December 2022 from https://uxdesign.cc/5-types-of-biases-in-user-interviews-and-how-to-reduce-them-6b863fb65af1

Goodman, E., Kuniavsky, M., & Moed, A. (2012). Chapter 11 – Usability tests. In E. Goodman, M. Kuniavsky, & A. Moed (Eds.), *Observing the user experience* (2nd ed., pp. 273–326). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-384869-7.00011-5

Hertzum, M., Hansen, K. D., & Andersen, H. H. 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., Hornbæk, K., Shi, Q., & Yammiyavar, P. (2009). Cultural cognition in usability evaluation. *Interacting with Computers*, *21*, 212–220. https://doi.org/10.1016/j.intcom.2009.05.003

Hertzum, M., & Kristoffersen, K. B. (2018). What do usability test moderators say? 'Mm hm', 'uh-huh', and beyond. *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, 364–375. https://doi.org/10.1145/3240167.3240181

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

Krause, R., & Rosala, M. (2020, April) *User experience careers: What a career in UX looks like today*. Nielsen Norman Group. Retrieved December 2022 from https://www.nngroup.com/reports/user-experience-careers/

Lewis, C. (1982). *Using the 'thinking-aloud' method in cognitive interface design*. IBM T.J. Watson Research Center.

Makri, S., Blandford, A., & Cox, A. (2011). This is what I'm doing and why: Methodological reflections on a naturalistic think-aloud study of interactive information behaviour. *Information Processing & Management*, *47*, 336–348. https://doi.org/10.1016/j.ipm.2010.08.001

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., Zhao, T., & Edwards, H. M. (2016). Look who's talking: Evaluating the utility of interventions during an interactive think-aloud. *Interacting with Computers*, *28*(3), 387–403. https://doi.org/10.1093/iwc/iwv014

Molich, R. (2018). Are usability evaluations reproducible? *Interactions*, *25*(6), 82–85. https://doi.org/10.1145/3278154

Molich, R., Wilson, C., Barnum, C., Cooley, D., Krug, S., LaRoche, C., Martin, B. A., Patrowicz, J., & Traynor, B. (2020, August 22). How professionals moderate usability tests. *Journal of User Experience, 15*(4), 184-209. https://uxpajournal.org/moderate-usability-tests/

Moran, K. (2019, December 1). *Usability testing 101*. Nielsen Norman Group. Retrieved December 2022 from https://www.nngroup.com/articles/usability-testing-101/

Oates, B. J. (2012). *Researching information systems and computing*. Sage Publications Ltd.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2381–2390. https://doi.org/10.1145/1753326.1753685

Pernice, K. (2014, January 26). *Talking with users in a usability test*. Nielsen Norman Group. Retrieved December 2022 from https://www.nngroup.com/articles/talking-to-users/

Pernice, K. (2020, February 9). *How to maximize insights in user testing: Stepped user tasks*. Nielsen Norman Group. Retrieved December 2022 from https://www.nngroup.com/articles/user-testing-stepped-tasks/

Sauro, J. (2010, November 23). *What metrics are collected in usability tests?* MeasuringU. Retrieved December 2022 from https://measuringu.com/usability-metrics/

Sauro, J. (2018, March 28). *How large is the evaluator effect in usability testing?* MeasuringU. Retrieved December 2022 from https://measuringu.com/evaluator-effect/

Wilson, C. E. (2007). The problem with usability problems: Context is critical. *Interactions*, *14*(5), 46. https://interactions.acm.org/archive/view/september-october-2007/the-problem-with-usability-problems1

Wixon, D. (2003). Evaluating usability methods: Why the current literature fails the practitioner. *Interactions*, *10*(4), 28-34. https://doi.org/10.1145/838830.838870

Wright, P. C., & Monk, A. F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies*, *35*(6), 891–912. https://doi.org/10.1016/S0020-7373(05)80167-1

Zhao, T., & McDonald, S. (2010). Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 581–590. https://doi.org/10.1145/1868914.1868979