Exploring Think-Alouds in Usability Testing: An International Survey

Research Article

—SHARON MCDONALD, HELEN M. EDWARDS, AND TINGTING ZHAO, MEMBER, IEEE

***Abstract*—Research problem:** *The study explored think-aloud methods usage within usability testing by examining the following questions: How, and why is the think-aloud method used? What is the gap between theory and practice? Where does this gap occur?* **Literature review:** *The review informed the survey design. Usability research based on field studies and empirical tests indicates that variations in think-aloud procedures may reduce test reliability. The guidance offered on think-aloud procedures within a number of handbooks on usability testing is also mixed. This indicates potential variability in practice, but how much and for what reasons is unknown.* **Methodology:** *An exploratory, qualitative survey was conducted using a web-based questionnaire (during November–December 2010). Usability evaluators were sought via emails (sent to personal contacts, usability companies, conference attendees, and special interest groups) to be cascaded to the international community. As a result we received 207 full responses. Descriptive statistics and thematic coding were used to analyze the data sets.* **Results:** *Respondents found the concurrent technique particularly suited usability testing as it was fast, easy for users to relate to, and requires limited resources. Divergent practice was reported in terms of think-aloud instructions, practice, interventions, and the use of demonstrations. A range of interventions was used to better understand participant actions and verbalizations, however, respondents were aware of potential threats to test reliability, and took steps to reduce this impact.* **Implications:** *The reliability considerations underpinning the classic think-aloud approach are pragmatically balanced against the need to capture useful data in the time available. A limitation of the study is the focus on the concurrent method; other methods were explored but the differences in application were not considered. Future work is needed to explore the impact of divergent use of think-aloud instructions, practice tasks, and the use of demonstrations on test reliability.*

Index Terms*—International survey, think-aloud methods, usability testing.*

THE USE of think-aloud methods as a tool to support usability testing has become the focus of much recent debate [1]–[4]. While think-aloud methods are no doubt popular among usability specialists [5], [6], we know little about the extent to which the individual methods are being used in practice, or indeed how they are being applied. The results of several field studies suggest that the use of the concurrent think-aloud approach within usability testing is undergoing change with practitioners moving away from the classic think-aloud approach espoused by Ericsson and Simon [7] toward a more relaxed and interactive style [8]–[10]. However, while the insights these studies provide are invaluable, we do not know if their results are representative. In this paper, we present the results of a survey which explored the use of think-aloud methods within usability testing; focusing, in particular, on how the concurrent think-aloud method is being used and the potential gap between the theory and practice of thinking aloud. Following a discussion of the literature on the use of think-alouds in usability testing we present the design and results of our survey. In the concluding section of this paper, we relate our findings to the literature and discuss the limitations of our work and areas for future research.

Manuscript received May 26, 2011; revised November 26, 2011; accepted December 26, 2011. Date of current version February 13, 2012.

The authors are with Faculty of Applied Sciences, University of Sunderland, Sunderland, SR6 0DD, U.K. (email: Sharon.mcdonald@sunderland.ac.uk; Helen.edwards@sunderland.ac.uk; Tingting.zhao@sunderland.ac.uk).

IEEE 10.1109/TPC.2011.2182569

LITERATURE REVIEW

In this section, we review the literature in order to identify the themes that our survey will explore. In particular, we focus on the procedural application of think-aloud methods and the relationship between procedure and test reliability. Specifically, we discuss: think-aloud methods and professional communication, types of think-aloud methods, theory versus practice in the use of the concurrent think-aloud, and the relaxed or interactive think-aloud. We conclude this section by presenting the aims of our study.

Think-Aloud Methods and Professional Communication Writers in the field of technical communication have argued for some time now that for usability work to be meaningful, it must be informed by the context in which products are used [11]. This position recognizes that a variety of factors shape people's use and experience with

digital products and argues therefore that usability must be contextualized in terms of its contribution to the design of interactive systems and their subsequent evaluation.

The most effective way to contextualize usability evaluation is through the use of field-based or ethnographic studies [12]. Where this is not possible or practical, usability professionals may strive to recreate the context of use within the usability lab. However, such a context necessarily includes factors that are not present in the user's real-world context. One key factor is the addition of another social agent who is there to facilitate and observe: the evaluator.

Technical communicators were among the first to highlight that while usability testing focuses on observing the interactions between the user and system, the interactions between the usability professional and the user are instrumental in the process [8]. As Still and Albers state *usability as science alone is not tenable. Users use products in context. Culture is always present...* [13, p. 189]. It is perhaps ironic therefore that one of the key methods used to elicit verbal data from users, the concurrent think-aloud technique and the framework that guides its use, is dependent upon eliminating interactions between users and evaluators as far as is possible as these interactions may serve to influence what users say and do at the interface [7].

However, usability tests do not take place within a social vacuum, and communication between participant and evaluator is an inevitable part of the process and a key area where the skills and methods of technical communication can make a significant impact. This paper explores the current use of think-aloud methods in usability testing, focusing, in particular, on how evaluators communicate with users and the procedures they follow when eliciting verbal data. An understanding of how these methods are used at the moment is likely to help technical communicators identify those areas of evaluator-participant communication in which they may continue to add value in terms of the future use of think-aloud methods and their subsequent development.

Types of Think-Aloud Methods Think-aloud methods have long been used in psychology in order to study task-based cognitive processes [7]. However, they have been much criticized; the primary objections are that thinking aloud is an unnatural process, and that the very act of thinking aloud may actively change the cognitive demands of a task. In response to these and other criticisms, Ericsson and Simon [7] undertook their seminal work on think-aloud methods and made specific recommendations to underpin their use.

Ericsson and Simon [7] identified two basic types of think-aloud: the concurrent think-aloud in which participants verbalize their thoughts during task execution and the retrospective think-aloud in which participants do so after task completion. While Ericsson and Simon encourage the use of these techniques together, within usability testing they have emerged as separate approaches, both receiving attention from researchers interested in usability evaluation. (See, for example, [14] and [15].) Other approaches to gathering verbal data have also entered the fray; these include constructive interaction [16], sometimes referred to as co-discovery [17], where users work together in pairs. Constructive interaction may take a number of forms: users may work together on an equal footing [18] or alternative formats may be used such as teach-back scenarios, in which the user teaches another person about the focus system. (See, for example, [19].) Pair-based approaches elicit data that are conversational in nature, therefore they address one of the key criticisms of the think-aloud technique: its artificial nature [18]. While all of these methods have been the focus of empirical investigations into their utility, the concurrent method is currently at the center of a debate that has resurrected the question of its validity and reliability as a tool to study cognitive processes. The debate hinges on observed divergence in the theory and practice of eliciting concurrent verbal protocols.

Theory versus Practice in the Use of the Concurrent Think-Aloud The Classic approach to gathering concurrent think-aloud data is based upon Ericsson and Simon's framework [7]. Ericsson and Simon undertook a meticulous study of the concurrent think-aloud method and developed a framework for gathering think-aloud data that incorporated a number of measures to safeguard the integrity of the resulting data. These measures include the use of neutral instruction scripts to avoid biasing the resulting verbalizations, the use of warm-up (practice) think-aloud sessions prior to data collection, and the use of a neutral "Keep Talking" reminder to maintain the think-aloud should the participant fall silent. (The interaction between evaluator and participant is kept to an absolute minimum.)

From the literature, it appears, however, that, in practice, evaluators may not follow Ericsson and Simon's advice. More than a decade ago, Boren and Ramey [8] first documented an emerging dichotomy between the theory and practice. They observed usability professionals working in two companies, and found practitioners frequently deviating from Ericsson and Simon's methodological recommendations by failing to give participants instructions in the prescribed manner, or the opportunity to practice thinking aloud. Perhaps the most contentious change to the classic technique observed by Boren and Ramey was the use of evaluator probes rather than simple "keep talking" reminders. Similar findings have been observed in other field investigations [9], [10]. Nørgaard and Hornbæk [9] conducted a study of 14 usability tests based in seven different organizations. They found that evaluators frequently engaged in interventions and questions that went beyond the user's actual experience with the system to hypothetical situations. They also found that detailed analyses of the resulting verbal data were rarely undertaken; a finding echoed by [20]. Shi [10] observed six usability tests in five different companies in Beijing, China. The findings suggest that users did not actively think-aloud and required encouragement from evaluators, which usually took the form of probing questions rather than reminder prompts. The evaluators decided what the important issues were prior to testing, and would question users about these if they were not mentioned, and would call users' attention to things they had not noticed.

The variation in the practice of think-aloud methods may be influenced by the range of advice on how to apply methods that is offered in the many texts on usability testing. For example, some recommend using a general instruction akin to Ericsson and Simon's guidelines [21], while others suggest that evaluators instruct participants to direct their verbalizations to aspects of the user experience, such as their likes and dislikes, and their affective or emotional response during interaction [17], [22]. Some texts recommend a think-aloud demonstration [5], [21], [23], whereas others do not [17]. Similarly, some recommend users have think-aloud practice [17], [21], [23]; whereas others do not mention this [5]. There are also differences in the nature of the practice and demonstration activities suggested. However, the one recommendation that all of these texts make is to use evaluator probes or interventions: this is in direct conflict with Ericsson and Simon's classic approach [7].

Relaxed or Interactive Think-Aloud The use of a more interactive style of gathering think-aloud data will be influenced by a number of factors. It is likely that, in practice, evaluators modify their style in response to individual differences in test users, or the cultural characteristics of the users they are working with [24]. Moreover, evaluators may feel it necessary to intervene in order to better understand a verbalization. As Ericsson and Simon note [7], concurrent verbal protocols frequently lack the characteristics of communicative speech that aids understanding on the part of the listener. Therefore, evaluator interventions may serve to increase understanding and, in doing so, reduce analysis time. Alternatively, evaluators may feel that the classic approach simply does not provide the type of data they require [25], [26]. For example, Makri, Blandford, and Cox [26] comment that users may not give reasons for their actions during think-aloud, only a description of the actions themselves. Recent analyses of the contents of concurrent protocols would lend support to this argument. Cooke [2] found that the verbalizations from concurrent think-alouds were mainly procedural in nature, with participants often reading from the screen. Similarly, Zhao and McDonald [3] categorized the utterances produced during a concurrent think-aloud study and found that the majority related to procedural information, with only a small number having direct relevance to usability analysis. Their study also included a comparison with a relaxed think-aloud: which they found produced a small increase in those categories related directly to usability analysis. However, evaluator probes may be a significant threat to the validity and reliability of the resulting data. Hertzum, Hansen, and Andersen [1] found that a relaxed think-aloud, which included evaluator interventions, increased test time and led to changes in behavior at the interface, with participants engaging in increased link traversal and scrolling behaviors. They may also serve to artificially improve the performance of interface tasks [4]. For example, Olmsted-Hawala et al. [4] found that the use of a relaxed think-aloud led to artificial improvements in task completion. It would appear, therefore, that the risks of intervening might well outweigh the benefits.

Aims of Our Study The picture painted thus far in the research literature suggests that use of the think-aloud approach is changing; with potentially negative consequences for test reliability and validity. However, this picture is based on the fruits of a small number of detailed field and empirical

studies. While these studies have produced very useful and insightful data, we do not know how widespread the interactive style is, nor do we know the reasons for this shift. There have been a number of large-scale surveys investigating user-centred design practices and the use of different methods within usability [6], [27], [28]. However none has focused exclusively on think-aloud methods. The aim of the work presented here is to build on the valuable data from earlier field studies. Since this is an exploratory study, we present no hypotheses but examine the following themes: the extent to which think-aloud methods are being used in usability research and practice, how these methods are being applied, and, in particular, the extent of the theory-practice gap in the use of the concurrent technique.

METHODOLOGY

In this section, we discuss the design, piloting, and implementation of our web-based survey which gathered qualitative and quantitative data from usability professionals (both in commercial practice and research/academia). We discuss the choice of method, identification of questions, piloting, participant recruitment, and data analysis.

Choice of Method Our knowledge of how think-aloud methods are being used in practice has been informed by a number of valuable field studies [8]–[10]. These studies have provided a detailed account of how a small number of practitioners are using the concurrent think-aloud approach. We wanted to understand how this method and other think-aloud approaches were being used by a larger sample; a web-based survey, therefore, was the obvious choice of method to follow.

Identification of Questions The survey was conducted as a web-based questionnaire and deployed using the Survey Methods software. Question areas were identified through analysis of the literature on think-alouds in usability testing. We wanted to understand what methods were being used, their perceived effectiveness, and the issues affecting use. In particular, we were interested in exploring the process of using concurrent think-alouds and the extent to which respondents were followed Ericsson and Simon's guidelines. The question areas identified were: method use, the think-aloud process, the nature of evaluator-participant interactions, and analysis activities.

We also gathered demographic (but anonymous) data from the participants so that subsequent analyses could give insight into differences in usage and viewpoints between different roles. A combination of closed questions, Likert scales, and open questions were used to enable us to determine not only how think-alouds are used but to also better understand the decisions that people take in choosing to adapt certain approaches.

Piloting The questionnaire was piloted by the authors on an iterative basis to refine the questions, eliminate ambiguities, and confirm that the pathways through the survey were clear. This gave confidence that the themes of interest were addressed and that it could be completed within 10 minutes. The survey was then piloted with 35 respondents: 10 external personal contacts within the usability field and 25 attendees at Nordi CHI'10. This pilot group gave insight into think-aloud usage and provided feedback on the clarity of the questions asked and the structure used. Some comments revealed misinterpretations of questions, others identified issues that individuals believed were important but had been neglected. The feedback was used to reassess the questionnaire and identify those areas that needed refinement: for instance, this resulted in a reduction in the number of questions asked and the number of open-ended questions. We found that participants chose not to complete these open-ended questions as they were "off-putting."

The final questionnaire consisted of 22 questions across 11 pages. A progress bar was included to provide participants with an indication of their status within the survey. "Next" and "Previous" buttons allowed respondents to navigate through the questionnaire. A welcome page outlined the purpose and scope of the survey and provided contact details for the research team. The survey was set up so that it could only be taken once at an individual machine and gave participants the option to return to the survey if they did not complete it in one session.

The structure of the final questionnaire had the following sections:

- "about you" gathering descriptive information about respondents
- "think aloud method usage" which asked about types of methods people used, why they used them, and the types of study in which they used them, we also collected information about frequency of usage using Likert scales.

- We then drilled down into the use of the concurrent method looking at instructions, demonstration and practice activities (both frequency and type), communication between evaluator and participants (whether or not evaluators would use interventions, the types of interventions where they did, and the reasons respondents had for and against using interventions).
- Finally, we asked about the frequency with which they engaged in different analysis techniques, how useful they found think-aloud data in comparison to other techniques for understanding usability problems. (These data were collected using Likert scales.)

Participant Recruitment The finalized web-based survey was live between November and December 2010. We invited individuals who had a proven interest in usability studies to complete the questionnaire. Individuals were identified from personal contacts and conference proceedings. A number of special interest groups were also used and list owners contacted for permission to circulate the questionnaire electronically to their groups. The key groups who gave permission for this were:

- SIGCHI (the mailing list for CHI Announcements)
- BCS HCI specialist group (Interactions)
- Usability Professionals Association

We also examined usability company information to elicit additional potential respondents thus maximizing the pool. Other individuals who may work in the area, but are not associated with usability companies or a professional body may not have been contacted, however, we do not think this is a significant limitation of the study. In January 2011, a summary of the preliminary (descriptive) results was published for the respondents on the survey website, whereas the detailed quantitative and qualitative analyses and resulting findings are reported here. The study did not require ethical approval.

Data Analysis Qualitative data from open questions were analyzed using a bottom-up card sorting process. Respondent comments were segmented and arranged on cards. These were sorted into themes by the first author, which were crossed checked by the second and third author. Statistics, where used, are nonparametric since the reported data are ordinal in nature.

RESULTS

The analysis that follows is based upon the data from 207 full completions. Some questions were not answered by all participants (those using skip logic); where this is the case, the reduced sample size is indicated so that percentages can be more clearly understood. The results are presented in the following order: respondents' profile, think-aloud method use, the concurrent think-aloud process, and analyzing test results.

Respondents' Profile

Location: More than half of our sample 54% ($n = 112$) was based in North America, 33% ($n = 68$) in Europe, 10% ($n = 20$) in Asia, and 3% ($n = 6$) in Australia. Only 1 respondent was based in South America.

Work Role: 58% ($n = 120$) of our sample described themselves as practitioners. Other large groups were drawn from research 24% ($n = 49$), academia 11% ($n = 23$), and 5% ($n = 10$) of respondents described themselves as User Experience Designers. The remaining work roles (2% ($n = 5$) of the sample) included roles such as Project Manager and Chief Executive. Combining data across groups 65% ($n = 135$) of respondents identified their roles as related to the practice of usability testing (in terms of design, development, or evaluation) and 35% ($n = 72$) were related to academic or research-based activity. Two respondents indicated that they were involved in practice and research. For these, we took the first role identified as their primary activity.

Primary Subject Area: Respondents reported a diverse range of background disciplines from Architecture to Education. Fig. 1 depicts the primary educational backgrounds reported by at least 5% of our sample. The "other" category (20%) contained disciplines that individually constituted 4% or less of the sample; these included Engineering, Education, Humanities and Science.

Experience: Our sample mainly comprised individuals who had worked in usability testing for a number of years. 37% ($n = 77$) had more than 10 years experience; 19% ($n = 39$) had between 6–9 years experience, 26% ($n = 53$) had between 3–5 years experience, 14% ($n = 28$) had between 1–2 years experience. Only 5% ($n = 10$) had been working within usability testing for less than a year.

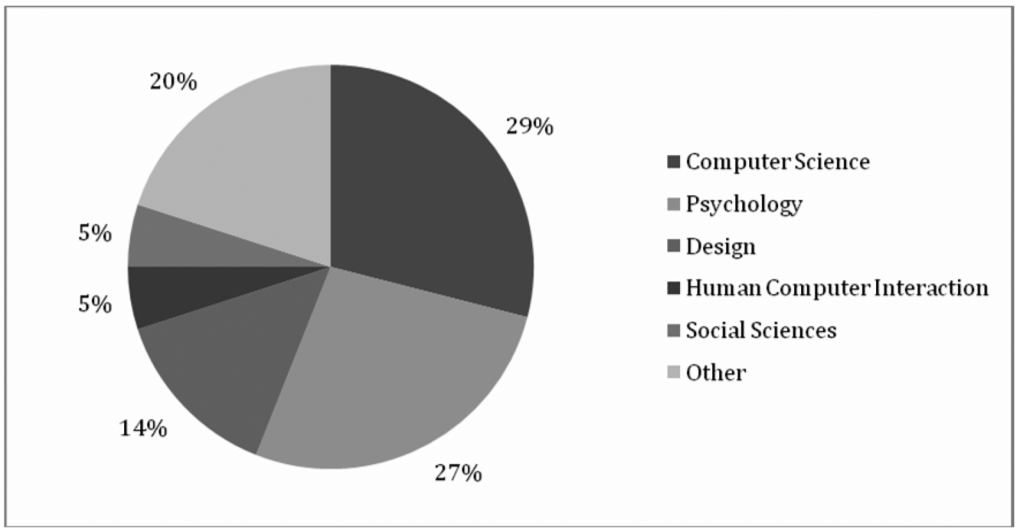

Fig. 1. Respondents' primary educational background (n = 207).

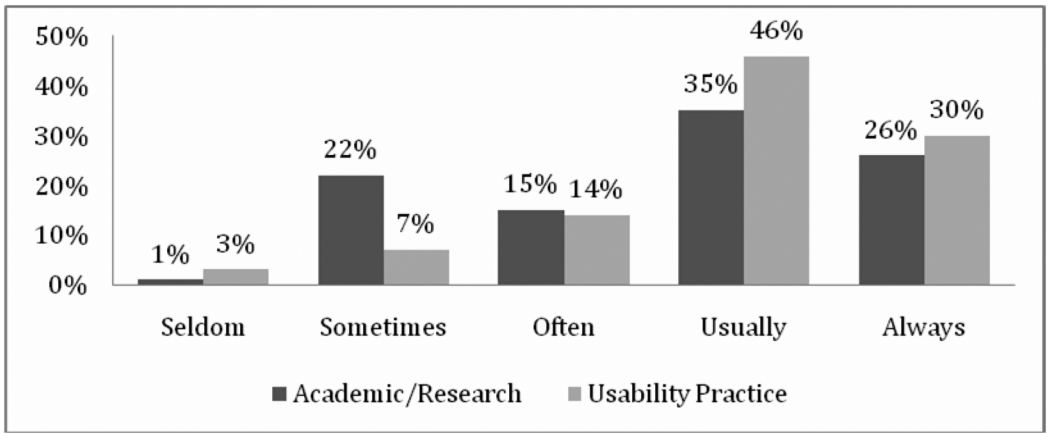

Fig. 2. Frequency of think-aloud method use (academic/research n = 72; usability practice n = 135).

Think-Aloud Method Use

Frequency of Use: We asked participants to tell us about the frequency with which they used think-aloud methods and the type of tests in which they were used. Overall, 29% of respondents indicated that they always used think-alouds in their tests, 42% reported that they usually did, 14% used them often, 13% sometimes used them with only 2% seldom using think-alouds. Fig. 2 breaks this result down by primary work role.

Test Type: We asked respondents about the type of studies in which they would use think-alouds: 86% reported that they would use them in formative usability studies: evaluations that are typically conducted at an early stage in the design process, where the results of the evaluation can be used to improve the system. 67% reported that they would also use think-alouds in summative tests: tests conducted at the end of the development process that typically use performance-based measures to determine or assess the overall quality of the product. Some participants used think-aloud methods in formative and summative tests, hence the reported values do not total 100%. Think-aloud methods were used to a lesser extent in field studies (49%) and research studies (51%). Fig. 3 shows the breakdown of study types by primary job role. It can be seen that respondents in the academic/research community used think-alouds more in summative and research tests than those involved in usability practice. A number of strong opinions were expressed against the use of think-alouds for summative usability tests. One participant commented "I would never use think-alouds in summative or benchmark tests"

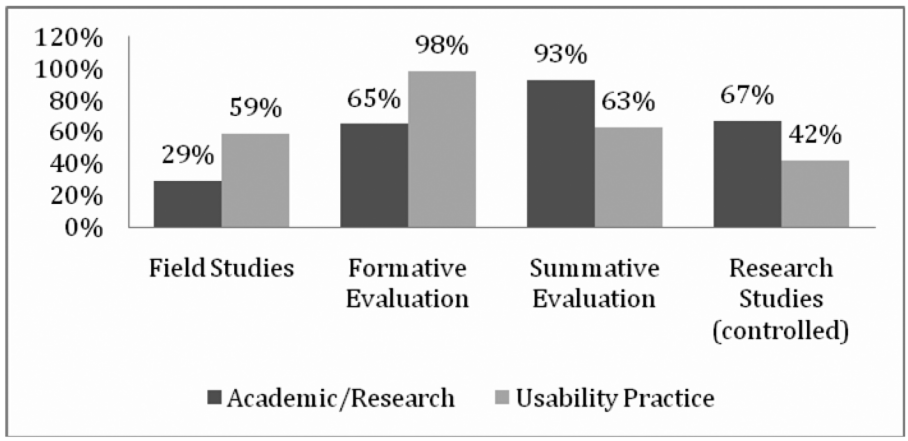

Fig. 3. Types of tests in which participants use think-aloud studies (academic/research $n = 72$, usability practice $n = 135$).

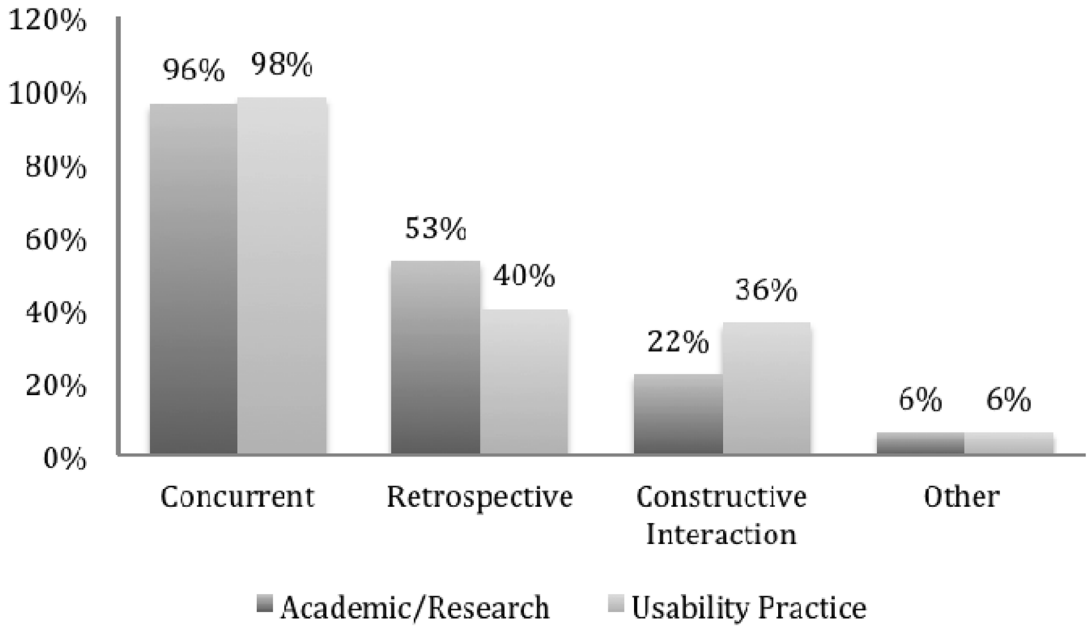

Fig. 4. Range of methods used by participants (academic/research $n = 72$; usability practice $n = 135$).

another indicated "of course I would not use think-alouds for timed tasks."

Experience With Methods: To determine the breadth of experience we asked respondents to tell us which think-aloud methods they had used. Overall, 97% ($n = 201$) of respondents had used the concurrent method, 44% ($n = 92$) retrospective think-aloud, 31% ($n = 64$) constructive interaction, and 6% ($n = 12$) identified additional methods. Fig. 4 shows the results for each method by the participants' associated work role.

The "other" methods included Cooperative Usability Testing (CUT) [29] and interviews based on issues logged during the test. One respondent indicated that they used "interrupt protocols" but gave no explanation of what this entailed.

Overall, 43% had used only one method, 39% had experienced two methods, and 18% had experienced three or more methods. Fig. 5 shows method experience by primary work role.

Method Used Most: Overall, the concurrent method was used most frequently by 89% ($n = 185$) of respondents. The next most frequently used method was retrospective (5%, $n = 10$), followed by constructive interaction (4%, $n = 8$) and CUT (2%, $n = 4$). When separated by primary work role,

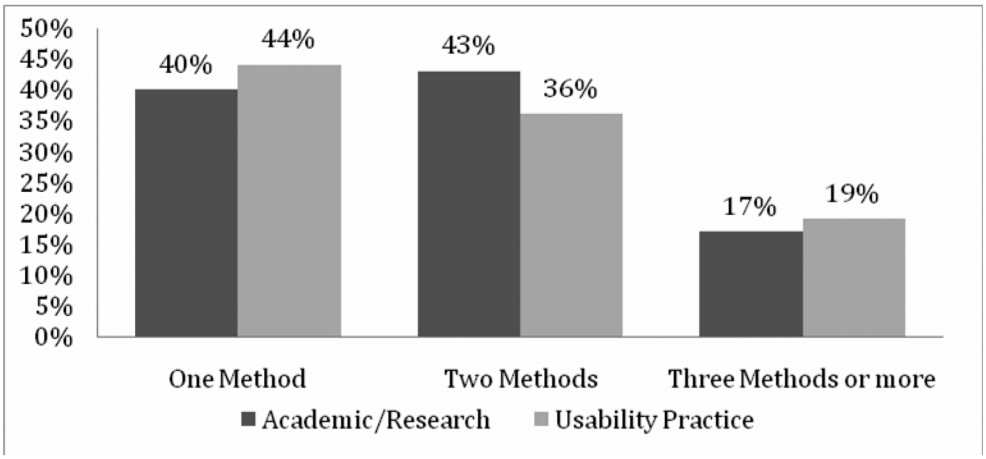

Fig. 5. Experience of using different methods (academic/research $n = 72$; usability practice $n = 135$).

the results still follow this pattern and are almost equivalent.

We asked participants why they adopted their most frequently used approach. Since the retrospective and constructive interaction approaches were used most by only a small number of respondents, only a few qualitative comments were provided. Users of the retrospective think-aloud adopted the method because it does not interfere with task performance. Comparing retrospective and concurrent think-aloud, one respondent commented: *concurrent causes task splitting and raises anxiety*. Others highlighted that the method was particularly useful following eye-tracking studies. Users of constructive interaction highlighted ecological validity as a key factor influencing method choice. One user commented that it was *more natural and spontaneous*, and *resulted in less data distortion as participants appear less aware of being observed*, another person said *the situation is more natural. I really don't like the inquisitive nature of a single person think-aloud.*

A large number of qualitative comments were posted detailing respondents' reasons for using the concurrent think-aloud method most frequently. The main comments are related to efficiency, added-value and understanding behavior and avoiding bias in data interpretation; these are discussed below.

Efficiency: The speed of the concurrent method appeared important with a number of respondents highlighting time as a key issue. For example, one respondent commented, *It takes less time per session and therefore less time overall.* Another said, *It suits fast iterative testing.* A number of people compared the concurrent technique to other methods, "*time needed is shorter than retrospective*; *if I had to use retrospective then it would take too long*; *I use it more than Constructive Scenarios because it's easier to schedule one participant than two, another reasons I use it more than retrospective (more than reliability) is that it takes less time per session (and therefore less time overall).*

Ease of instructing users to think-aloud also appeared important. One respondent noted that concurrent think-aloud is the *easiest method to explain and to remind users to employ.* Others indicated that test participants see the relevance and benefit of the technique "*people relate to it*" and are able to work with the technique effectively, one respondent commented *it works for the participant making it possible for them to engage with both the material and me.*

Understanding Interaction and Avoiding Bias: Our respondents repeatedly highlighted the insights concurrent thinking aloud provides in terms of understanding task-related cognitive processes that might otherwise be difficult to interpret: *reveals thought processes I can hear the users logic*; *users self reports are essential to understanding what a user is doing. Since user behavior goals are often ambiguous or opaque.* Respondents also emphasized the importance of reactive or live data that had not been subject to posttask (inaccurate) rationalizations *captures initial reactive thought processes rather than summarized processes*; *gets the immediate response of users and that is what counts.*

Added Value: A number of respondents found that the concurrent think-aloud tests were tests more interesting for clients who might be observing the

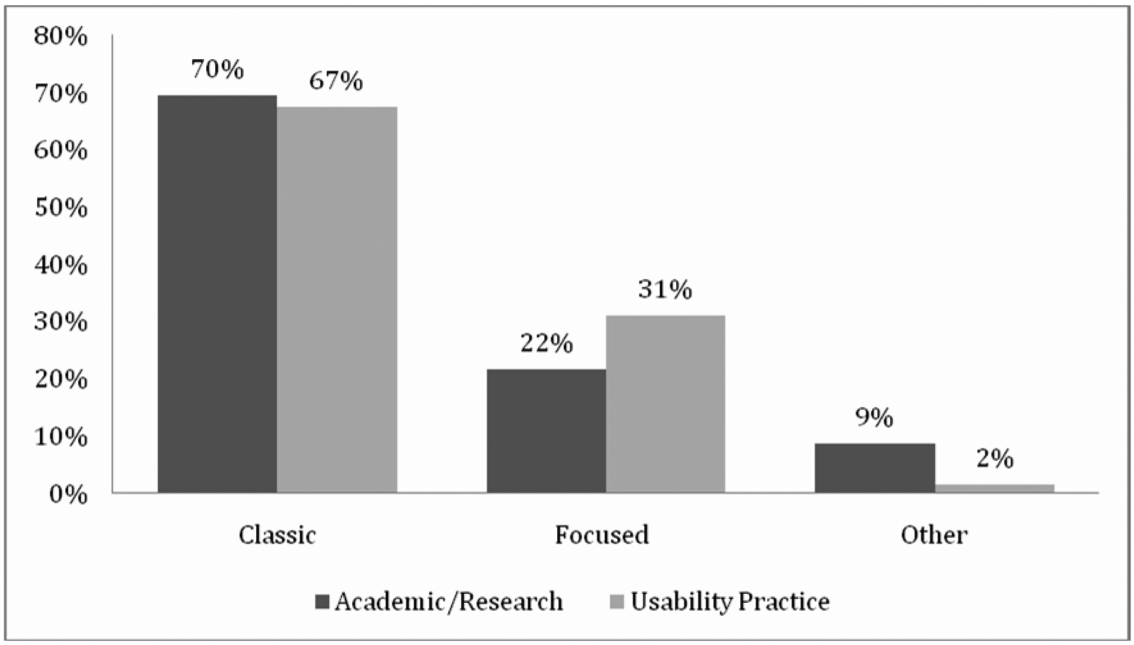

Fig. 6. Use of instructions during concurrent think-aloud (academic/research $n = 69$; usability practice $n = 132$).

test *gives viewing clients something to watch rather than silence*; *stakeholders can watch the interaction with the product by watching sessions as they play out*. Another respondent hinted toward the persuasiveness of the technique, *this method allows for great comments that can be used for highlight videos added to reports for our clients*.

There was no relationship between participants' length of experience within the field of usability testing and their use of or preference for different types of methods. The concurrent approach was the primary method used by the majority of respondents regardless of length of experience.

Concurrent Think-Aloud Process To determine the extent to which respondents adhere to Ericsson and Simon's guidelines [7], we asked about think-aloud instructions, practice, and evaluator-participant interactions. In addition, we explored whether think-aloud demonstrations were being used as advocated by some texts [17], [21], [23]. The percentages that follow in this section are based on a sample of 201 rather than 207 as six people had not used the concurrent method.

Instructions: Ericsson and Simon [7] note that think-aloud content may be influenced by the wording of the instructions. Instructions that require the participant to produce specific information may distort the think-aloud because they direct the participant's thought processes to their own procedures. Therefore, in their framework, Ericsson and Simon advise the use of a general instruction which simply asks participants to say outloud their thoughts as they complete tasks. By contrast, some usability tests recommend the use of focused instructions that direct the user to specific areas of the user interface or user experience [17], [22].

Overall, 68% of our sample indicated that they used a general think-aloud instruction in line with Ericsson and Simon [7]. An additional 28% indicated their instructions were focused toward the user experience, and 4% described alternative approaches. Looking at the impact of primary work role, we found that the use of these instructions was similar between the groups with only a small increase in the use of focused instructions among respondents involved in usability practice. (See Fig. 6.) The other category included the use of instructions relating to users working in pairs or retrospectively, for example one person commented *I ask participants to perform a task, then depending [on] behavior ask them to examine how and or why they did in that particular way*.

Some respondents provided us with examples of the type of instruction they use; these often focused on the need to respond to differences in users, for example *I tell them to focus on the task and then talk later if needed. I do this particularly if you get a "talker" that cannot focus on the interface and simply talks*. Others highlighted the danger that a

TABLE I
FREQUENCY OF USING A THINK-ALOUD PRACTICE SESSION

	Academic/Research	Usability Practice
Always	21.7%	12.9%
Sometimes	17.4%	15.9%
Rarely	17.4%	14.4%
Never	43.5%	56.8%

Note: Academic/Research $n = 69$; Usability Practice $n = 132$

focused instruction might influence the reliability of the verbal data. For example *I don't tell them what I'm interested in as they'll make statements just to keep me happy*. A common theme was to use a general instruction and then to follow up experience issues after the think-aloud session, for example, *Usually, the instruction at the start of the set of tasks is to talk through what they are thinking. After they complete the task, there are follow up questions asking them to specify what surprised them or didn't make sense, and how it could be improved*.

Think-Aloud Practice: Only 16% of our sample used a think-aloud practice session in all of their tests; 16.4% did so occasionally, 15.4% used one rarely, but more than half (52.2%) of our sample indicated that users never practiced. Breaking down these data by primary work role, it can be seen from Table I that respondents in Academic/Research used a practice think-aloud session more often than those involved in usability practice.

Some respondents said they had never needed to use a practice session since their participants could relate easily to the method. Others indicated that they took a flexible approach to practice, only using it when it proved necessary, *we don't plan for it but will include when appropriate*. It appeared that rather than have a dedicated or named practice session a number of respondents allowed for practice through the use of an ice-breaker task for which the results are not critical. For example, one respondent commented that they used *a related interface, which might appear to be the test, but for which the results can be tossed if they are too variable due to participant 'training' issues. Not always communicated as practice*.

There were also differences in the type of task used. Of those who used practice ($n = 96$), only 9% reported that they used activities in accordance with Ericsson and Simon's recommendations (such as math problems, simple mental arithmetic) and the majority of those who did were in the academic/research community. An additional 27% used procedural tasks akin to those recommended in usability textbooks [17], [21], such as disassembling a ball-point pen. The remaining 64% had users practice with either a related interface or the interface under test.

Think-Aloud Demonstrations: Overall, only 22% of our respondents always used a think-aloud demonstration and 26% sometimes did. Thirty-five percent never provided participants with a think-aloud demonstration and 17% did so only rarely. The proportions of responses in each of these categories were equivalent when the sample was subdivided into a primary work role. Of those who did use demonstrations, 71% ($n = 93/131$) used an unrelated product while 15% ($n = 20$) based it on the product that was the focus of the test. Fourteen percent ($n = 18$) used an alternative type of demonstrations, such as video clips, from a previous think-aloud.

Some respondents used demonstrations to illustrate the type of verbalizations participants might produce, or indeed the type of utterances that might be useful to them. For example, one commented *I use an unrelated product, but I am very critical of it rather than being neutral*, another said *I give examples of items they can react to E.G. a button on the homepage, the wording of the menus, etc.* Another said *I provide general examples of the type of comments I'm expecting (only with participants that seem confused or unsure of the instructions), which is extremely rare*.

Other respondents were concerned about not influencing the users; for example, *I never demonstrated with the product because I would want them to start fresh with the product the first task of the set*.

Evaluator-Participant Interactions: We asked respondents whether they typically avoided interventions, intervened actively, or if they modified their approach based on participant characteristics. Overall, 53% modified their approach, responding flexibly to test participants and situations, and 33% always tried to avoid intervening during think-aloud. Only 14% said that they always adopted an interactive approach. Fig. 7 depicts the pattern of response by primary work role.

More than half of those in the usability practice group modified their approach to suit the characteristics of the user. Limiting the use of interventions was more typical for the academic/research group. However, the number of practitioners who always intervened was low. The largest number (proportionally) who always avoided

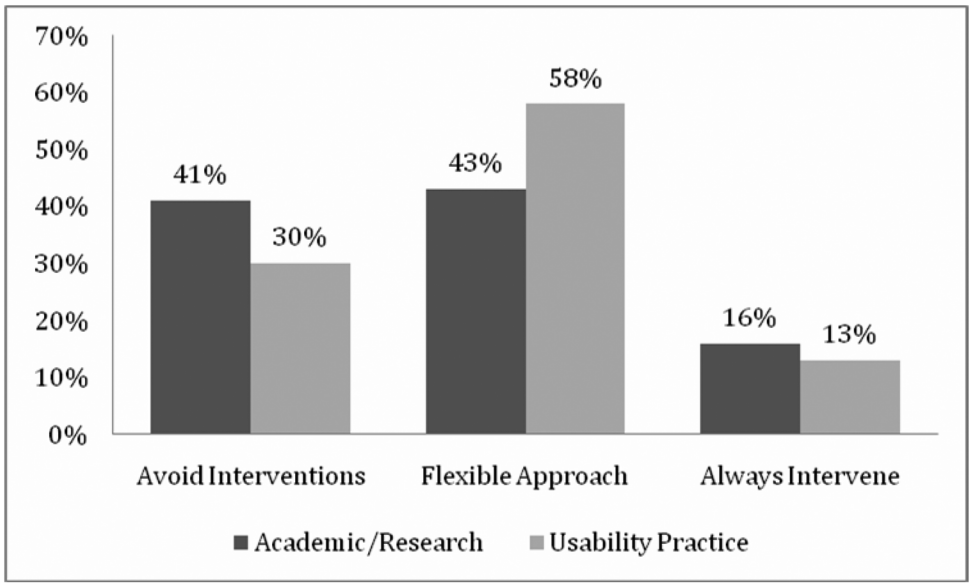

Fig. 7. Approach to gathering think-aloud data (academic/research $n = 69$; usability practice $n = 132$).

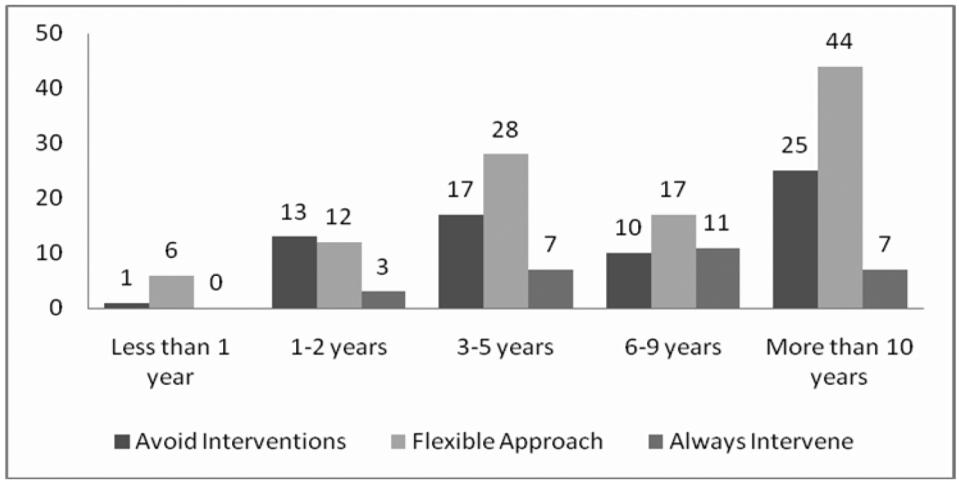

Fig. 8. Approach by experience (academic/research $n = 69$; usability practice $n = 132$).

interventions were in the most experienced group. (See Fig. 8.)

The details provided by respondents about their individual approaches and the types of interventions they typically made were analyzed using a bottom-up card sort.

Maintaining the Think-Aloud: A number of respondents only used interventions that were neutral in tone to maintain the think-aloud; for example, *What are you thinking?*. However some respondents believed that they needed to give positive feedback to maintain the think-aloud, for example; *If a participant is not thinking aloud, I will remind the participant to do so—in neutral terms, 'What are you thinking now?' I will also praise participants for thinking aloud—again in neutral terms. For example, at the end of a task, I might say, 'Thanks for thinking aloud through that task. Your comments are very useful'.*

Responding to Differences in Users: Many people reasoned that individual differences in the quantity and quality of people's verbal data could necessitate using a more interactive style. For example, *Some people have a harder time dividing their attention and need reminders. Others give a blow by blow and do not need 'coaching'*. Another respondent commented *Sometimes people are shy and need to be coaxed, others produce useful information without much prompting*. Others who prefer not to intervene often ended up doing so in response to

quieter participants. One respondent commented, *It is always the best case that they spontaneously think-aloud, but this is rarely the situation. With some participants it is necessary to modify the approach and use a more probing type of technique to encourage them to speak.*

While these comments focused primarily on encouraging more reserved users to verbalize, some respondents said that they would intervene if the participant was not providing the right kind of data, either because he or she was talking about an unrelated event or because the verbal data produced was not perceived to be useful. For example, one respondent commented, *If a user is not providing useful data I will probe*, another commented, *I will provide specific feedback if the person is not verbalizing the correct type of data or are taking blame for misunderstandings.*

Avoiding Bias: Respondents were aware of the possibility that interventions might affect test reliability. Two respondents (who avoided interventions where possible) said *Too easy to fall into coaching or questioning mode so keep interventions to a minimum and focused on getting them to think-aloud*; *I want to see the users natural response and not 'guide' them so I keep my involvement to minimum.* Others saw a need to intervene to gain a better understanding of the user's experience, but took steps to reduce the impact of this by questioning users after use: *I prefer to save the questions from the moderator until the task is completed, and then allow time for clarifications and interpretation before the next task*, or by adhering to a set of neutral interventions *As little as possible, and as much as possible sticking to a set number of phrases, to avoid biasing anything at all, even tangentially or subconsciously*, another respondent commented, *When users say something I don't understand, I will try to prompt for clarification as unobtrusively as possible. For example, I may repeat a word they said as a question, or say 'tell me more about that.' I avoid hypotheticals or tangents and stick to what they have already mentioned for follow up.*

Use of Interventions Interventions can broadly be categorized into two types: (1) those which might be necessitated by the context of the test, what Boren and Ramey call the contingencies of usability testing, and (2) those which seek to gain a deeper understanding of participant utterances or behaviors or reactions to the interface. We explored these two types of interventions further with our respondents.

Contingencies of Usability Testing: We asked all of the respondents, regardless of their stated approach (even those who had indicated that they would avoid interventions), if there were any situations, such as those identified by Boren and Ramey as the contingencies of usability testing, in which they would consider interacting with participants during a concurrent think-aloud session. Across the 201 respondents who had used the concurrent method, most said they would intervene in those situations that arise out of user confusion about a task or the inability to pursue a task: 77% where the user was stuck on a task, 61% where a user mistakenly thought a task was complete, and 70% where the user asked for help. Fewer respondents intervened in situations in which users behaved in a way that was counter to expectations, such as sidestepping features of interest (49%) or completing a task in an unexpected way (47%), 7% asserted that they never intervened. Using the primary work role as a comparator (See Fig. 9.), we can see that respondents with a focus on the practice of usability testing are more prone to intervening than those with an academic focus.

The "other" category here consisted of additional intervention scenarios, the most commonly identified being quiet users, or users who appear anxious or nervous. For example, one respondent commented, *Asking into what the participant say makes the participant feel more comfortable*; another commented, *to ensure that the participant does not feel that any difficulties they experience are their fault.*

Intervention Types: We were interested in the types of intervention respondents reported making in order to gain a better understanding of users' utterances and behaviors. This question was asked only of those individuals who stated that they would intervene during a think-aloud session or were at least flexible in terms of their approach to interventions. It was not asked of those respondents who indicated that they would not intervene. Table II shows the percentage of participants for each specific type of intervention: respondents were free to select as many options as appropriate. Our data suggest that respondents used interventions to clarify their understanding of participant utterances and to gain insight into the users' strategies through explanations of actions they make.

Other interventions involved responding to participants who were having difficulties with particular tasks. For example: *When participant*

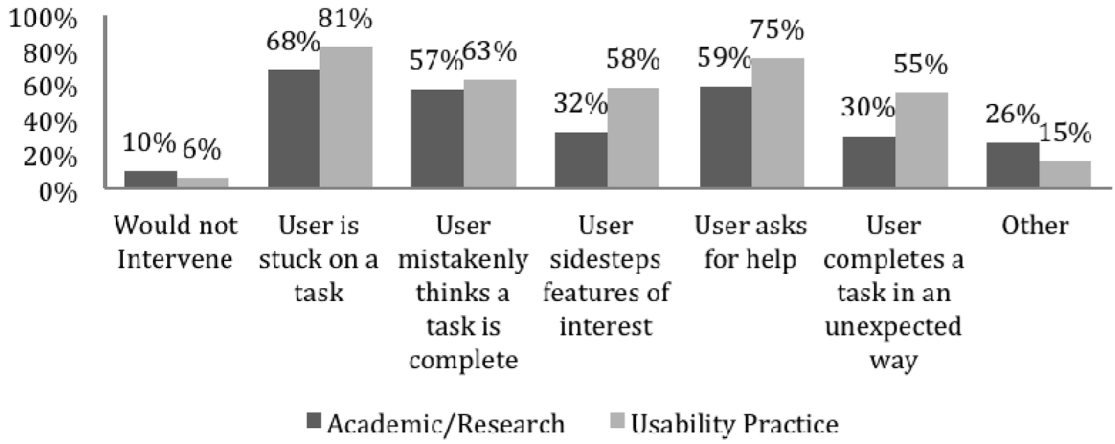

Fig. 9. Situations in which evaluators would intervene (academic/research $n = 69$; usability practice $n = 132$).

TABLE II
ADDITIONAL SITUATIONS IN WHICH INTERVENTIONS WOULD BE MADE

	Overall	Academic/ Research	Usability Practice
Clarify understanding of a comment	95%	88%	98%
An explanation of a user action	87%	83%	88%
To understand an Interjection	85%	78%	88%
To understand an experienced problem	82%	76%	85%
To understand the cause of a problem	76%	78%	75%
To understand the impact of a problem	69%	68%	70%
Seek the user's opinion	61%	49%	67%
Seek a recommendation from the user	45%	39%	47%
Other	7%	10%	5%

Note: Overall $n = 134$; Academic/Research $n = 41$; Usability Practice $n = 93$

is clearly frustrated with the task; *to support a user whose self-esteem is threatened by having difficulties with a product*. Some respondents indicated that interventions might be suggested by the client *if the client has directed us to ask specific questions of the participants*. Other respondents used probing questions during the think-aloud question to either gain a better understanding of a comment or a person's reaction to the interface, for example, *I prompt them to find out what they think of something as they make a comment, or ask them general questions about a page*.

Reasons Not to Intervene Overall, 33% of respondents told us that they avoided making any interventions during think-aloud sessions (67 out of 201 participants who had used the concurrent method). Of those who avoided interventions, 84% indicated that this was to avoid evaluator-introduced bias. Looking at the reasons further, Table III shows the key issues and the extent to which the respondents identified these. (Participants could select as many items as they thought applicable.)

TABLE III
REASONS NOT TO INTERVENE DURING THINK-ALOUD

	Overall	Academic/Research	Usability Practice
Influence behavior	90%	86%	92%
Influence verbalizations	75%	75%	74%
Distract the user	64%	57%	69%
Other	10%	11%	10%

Respondents' qualitative comments raised concerns that evaluator interventions may reduce the reliability of test data. However, there were also concerns about how interventions might influence the participant's feelings about their performance.

TABLE IV
ANALYSIS ACTIVITIES

Activity	Mean Rating	Standard Deviation
Review test notes	4.47	0.82
Transcription	3.21	1.37
Review test videos	3.13	1.30
Real-time coding	3.0	1.40

For example, one respondent commented, *I don't want to intimidate the user and imply there is a right answer*; another said, *I don't want the user to feel as if I need to take over and they're not thinking aloud correctly. Many users I've worked with on usability studies feel as if it is a test of how well they can do their jobs and it is very important to put them at ease and not make them feel like anything they're doing is wrong or being monitored for correctness.*

Analyzing Test Results We asked all participants about the frequency with which they analyzed data using the following activities: transcription of think-aloud data, review test notes, review test videos, and real-time problem coding using specialist software. Respondents rated activities on a 5-point scale ranging from 1 (never) to 5 (always). Table IV shows the mean rating and standard deviation for each of these activities.

A Friedman Anova test revealed that there was a significant difference in the ratings for each activity ($c^2(3) = 142.53$, $p < 0.05$). Pairwise comparisons were made using Wilcoxon tests with a Bonferroni correction and a significance threshold of 0.008. Reviewing test notes was the most frequently used data-analysis method (notes versus transcription: $Z = -8.624$; notes versus real-time coding $Z = -9.39$; notes versus reviewing videos: $Z = -8.89$). There were no other significant differences. The mean ratings for the use of each activity followed the same pattern regardless of identified work role.

Perceived Contribution of Test Measures We asked participants to rate the extent to which they believed different measures contributed to the identification of usability problems and understanding their causes during analysis. The measures were: think-alouds, posttask interviews, performance metrics (e.g., time on task), and behavioral data (e.g., from video data) and eye tracking. A 5-point scale ranging from 1 (not at all useful) to 5 (very useful) was used. Since respondents might not have experienced all measures, we included an option of Don't Know. As a consequence, there were some missing values on some of the measures. An analysis of the missing values revealed that more than 20% of cases were missing for the eye tracker; therefore, we excluded this measure from our analysis. Table V shows the mean ratings provided by respondents.

TABLE V
PERCEIVED CONTRIBUTION OF METHODS TO PROBLEM IDENTIFICATION AND UNDERSTANDING PROBLEM CAUSATION

	Identification		Causation	
	Mean	*SD*	Mean	*SD*
Think aloud	4.57	0.80	4.62	0.71
Interview	4.18	0.91	4.17	0.87
Performance	2.89	1.33	3.63	1.19
Behavior	3.55	1.22	3.95	1.03

Problem Identification: A Friedman's Anova revealed that participants rated the methods differently in terms of their contribution to problem identification ($c^2(3) = 227.07$, $p < 0.05$). Pairwise comparisons using Wilcoxon tests with a Bonferroni correction revealed that think-alouds were rated higher than interviews ($Z = -4.777$); interviews were rated more highly than behavioral measures ($Z = -5.42$) and behavioral measures were rated more highly than performance data (−6.745). When the results are separated by work role, this pattern is maintained.

Problem Causation: A Friedman's Anova revealed that participants rated the methods differently in terms of their contribution to understanding the cause of usability problems ($c^2(3) = 111.62$, $p < 0.05$). Wilcoxon tests revealed that think-alouds were rated higher than interviews ($Z = -5.966$); there was no significant difference between interviews behavior measures ($Z = -2.087$) but behavior was rated more highly than performance (−3.720). When the results are separated by work role, this pattern is maintained.

CONCLUSIONS, LIMITATIONS, AND SUGGESTIONS FOR FUTURE RESEARCH

In this section, we relate our findings to the literature on think-alouds in usability testing, identify the limitations of our work, and discuss areas for further research.

Conclusions The survey results give an insight into current think-aloud practice in usability studies and the underlying reasons determining the approaches adopted. Some of the issues raised in the literature review (which have been primarily based on in-depth studies of specific cases) are supported by the survey findings; however, there are others where the responses challenge previous

reports. This discussion focuses on three main areas: method use, in particular, the dominance of the concurrent technique; the divergence in theory and practice of using think-aloud methods; and the nature of evaluator interventions.

Method Use: The literature suggests that think-aloud methods are the technique of choice within usability testing [5], [6], and the results of our survey would support this with 98% of respondents using think-aloud methods at least sometimes, and 71% using them at least on a usual basis. In comparison to other available techniques (interviews, behavioral measures, performance data), the respondents consistently rated think-alouds as better for problem extraction and identification of problem causation. This finding lends support to those of Gulliksen et al. [6], who found think-alouds to be rated in the top five methods used by usability professionals in Sweden. Given the results of Gulliksen et al.'s survey, we were not surprised that think-alouds were rated so highly; we were, however, surprised at behavioral data during interaction being rated as less useful than interviews. We have no firm explanation for this finding, other than to comment that it may be related to our other finding that respondents reviewed test videos less often than referring to their test notes.

The concurrent method was the dominant approach used by our respondents with 97% having used it and 89% stating that it was their most frequently used approach to gathering think-aloud data. Respondents highlighted the benefits the concurrent approach offered in understanding user actions, helping to avoid errors of interpretation. However, the most often cited reason for choosing this method over the alternatives was its apparent fit to the context of testing. Respondents consistently reported that the method was fast, efficient, and easy for users to relate to, whereas retrospective think-aloud makes greater demands on time, and constructive interaction makes greater demands on the process of test user recruitment and test scheduling. Time seems to be of particular importance in terms of method selection and in terms of how the resulting data are analyzed. For example, there is a greater reliance upon test notes for analysis rather than transcription or reviewing test videos; as one participant commented:*it is sometimes impossible to find the time to process and review videos...you have to learn to adapt your methodologies to fit client needs, budgets and time constraints*. The pressures of time and working contexts have been highlighted by [30] who commented that practitioners frequently work with only limited resources.

Theory versus Practice: The Concurrent Method: The rigor of the classic approach rests on three basic tenets: a general instruction to think-aloud, the use of think-aloud practice, and limited interaction between participant and evaluator [7]. A number of studies [8]–[10] suggest that practitioners are moving from the classic approach toward a more relaxed and interactive style by either omitting or modifying these tenets.

Instructions: Boren and Ramey [8] observed that evaluators failed to give participants instructions in the manner prescribed by Ericsson and Simon [7]. In contrast, our findings suggest that 68% of respondents follow this recommendation with 28% using a more focused instruction akin to those recommended by some usability texts [17], [22], [23]. This may be because it is thought that they offer a means of securing the type of data analysts require, rather than the procedural descriptions that have been uncovered in research studies for example [2], [3]. However, such priming may also serve to distort behavior and some of our participants warned of the potential pitfalls in trying to lead the user. Moreover, there is, as yet, no empirical evidence to suggest that such instructions would yield more useful data or identify what impact they might have on performance.

Practice: Boren and Ramey noted that think-aloud practice was omitted in their observed sessions. Our findings echo these with 52% never using practice and 15% doing so only rarely. Typically practice was used when the evaluator judged that a specific user needed it. Most respondents seemed comfortable in dispensing with these preliminary activities, with some commenting that they no longer needed them since participants are comfortable with the concept of thinking aloud. It may well be that evaluators are, in part, driven by the need to reduce time spent on costly test sessions and, thus, have no qualms in eliminating what are seen as nonessential practices.

Think-aloud demonstrations are not part of Ericsson and Simon's approach, however, they are recommended by a number of texts on usability testing [17], [21], [23]. However more than half of our sample, rarely or never gave a demonstration. For those who more frequently used demonstrations, there was great diversity in practice including: video clips of previous studies, or providing live critical evaluations of another

product, as well as neutral examples. Several respondents commented that users do not need demonstration, and others that they were given only when a user seemed confused or unsure.

Interventions: Intervening beyond the use of think-aloud reminders has been reported by several authors [8]–[10]. The evidence from our survey is of divergent practice: varying from no probing during tasks but following up on issues after completion, to probing interventions where "interesting" issues emerge. Overall, only a third of our sample said that they tried to avoid interventions, in line with Ericsson and Simon's advice, 53% took a more flexible approach responding to differences in participants and situations. Only 14% indicated that they always intervened. Interestingly, analysis of the data against respondents' experience revealed that those with substantial experience were least likely to intervene, whereas those with least experience were most likely to do so. This is an interesting finding but the reasons underlying it are unknown and would require investigation beyond the scope of this survey.

Compliance With Ericsson and Simon: Overall, only 6% ($n = 13/201$) of participants who had used the concurrent method answered in a way that was in strict compliance with all three of Ericsson and Simon's recommendations. If we remove the requirement for think-aloud practice, this increases to 25% ($n = 51$). This finding corroborates those of Boren and Ramey in their initial investigation of the gap between the theory and practice of using think-aloud methods.

Evaluator-Participant Interactions: As noted earlier, only 33% of our respondents limited interventions to maintaining the think-aloud. One of the main reasons that others adopted a more flexible approach is related to individual differences in test users. Choosing to intervene during a think-aloud session will depend upon a number of situational factors, including time pressures, the nature of the test, the characteristics of the user, or the clients requirements. The need for flexibility appeared to be the key for many of our respondents.

Our qualitative data suggest that a one-size-fits-all approach to gathering verbal data is not always appropriate. Evaluators frequently have to tailor their approach to suit the characteristics of the individual user or the situation in which they are testing. Not all users perform equally, some think aloud with ease, while others struggle and evaluators intervene to capture the necessary data. This has resonance with the reaction to cultural factors identified by [10] and [24]. However, it may also be related to the infrequency with which think-aloud practice is used, leaving test users unprepared for the process: this may be a fruitful area for further investigation. A number of respondents used interventions to reduce test anxiety and create a more relaxed environment for the user to work in. This raises an interesting question about the impact of methodological processes on the test user's experience. Does strictly adhering to the classic technique inadvertently reinforce the notion that it is the user who is being tested rather than the product?

Some field studies have uncovered the use of evaluator interventions that would most certainly threaten the validity and reliability of the resulting data [9], [10]. In these studies, there is evidence of evaluators seeking specific responses to interfaces and directing test users where such responses were not forthcoming or asking participants questions that went beyond their user experience. Qualitative data in the survey provide some limited confirmatory evidence of such practices (for instance, the statement ... *would intervene if the participant was not providing the right kind of data*). By contrast, many of our respondents were clear about the need to avoid steering users and frequently discussed strategies to limit bias in tests such as only questioning users after the test, and then only following up on issues already mentioned by the users.

The intervention types that were used most were those that would either lead to a better understanding of a verbalization, for example, when a participant makes an unexplained interjection or to clarify the evaluator's understanding of a verbal utterance or an action performed that was not accompanied by a verbalization. Ericsson and Simon note that concurrent verbal protocols are often disjointed and fragmented; the process of using probes may help evaluators to understand verbal data as it occurs. Indeed, in line with [9] and [20], our results suggest that evaluators do not frequently transcribe verbal data, and rely mainly on their test notes for purposes of analysis, suggesting that there is a need for fast turnaround of results. Interventions that make utterances more readily understood are likely to aid this process. However, research suggests that the use of a relaxed approach, which includes interventions, can compromise the reliability of the resulting data. Our qualitative data indicate that respondents are well aware of this, particularly when intervening

during tests, and take steps to reduce this impact. For some respondents, it seems that the reliability and validity considerations which affect the decision of whether to intervene, are pragmatically balanced against the need to capture useful data in the time available.

Limitations It could be argued that a questionnaire aimed at gathering data about think-aloud methods would confirm the extensity of their use, as respondents are likely to be interested in, and, therefore, using the techniques. While we do not dispute this, we would argue that the value in our data lies in the insights it provides in terms of the nature of the methods used and respondents' reasons for adopting these approaches.

Respondents involved with the practice of usability indicated that they used a more flexible approach to gathering think-aloud data, and used more intervention types than those in academia/research. This may be due to respondents in the academic/research community engaging in tests that were more summative in nature, such as controlled experiments. Indeed, when asked about test type, individuals from the academic/research group reported engaging in more summative and research studies than formative usability evaluations. A limitation of the study is that we did not drill down further to explore how interventions were modified further by test type. However, since evaluator interventions have been linked to changes in behavior at the interface [1] and enhanced task performance [4], an argument could be advanced that interventions, regardless of whether the test is formative or summative in nature, should be avoided as they could lead to interface problems being missed at either stage.

The main part of the survey focused on the methodological application of the concurrent method; while we did explore the other types of methods participants used and their preferences in this respect, we did not consider these methods in detail or explore any methodological differences in their application.

Suggestions for Future Research Our results have confirmed some of the findings of earlier field investigations on the use of think-aloud methods within usability testing. Evaluators do appear to be moving away from Ericsson and Simon's classic approach to a more relaxed way of conducting think-aloud tests, and we have found evidence that only a small proportion of our sample is adhering to Ericsson and Simon's recommendations. While the use of evaluator interventions has been investigated, some of the other areas of divergent practice warrant further investigation, in particular, the use of instructions that require specific types of verbalizations from users, such as explanations or directions to comment on aspects of the user's experience during interaction.

The popularity of the concurrent method and its apparent suitability to the context of testing which may be resource poor, perhaps provides a *prima facie* case for the continued focus on concurrent think-alouds in usability research so that their effectiveness can be optimized.

ACKNOWLEDGMENTS

The authors would like to thank everyone who took the time to complete the questionnaire. The authors also thank Ted Boren, Erik Frokjær, Emilija Stojmenova, and Jennifer Romano (UPA-DC Vice President), who assisted them greatly in distributing the survey. They also thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] M. Hertzum, K. D. Hansen, and H. H. K. Andersen, "Scrutinising usability evaluation: Does thinking aloud affect behavior and mental workload?," *Behav. Inf. Technol.*, vol. 28, no. 2, pp. 165–181, 2009.
[2] L. Cooke, "Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach," *IEEE Trans. Prof. Commun.*, vol. 53, no. 3, pp. 202–215, Sep. 2010.
[3] T. Zhao and S. McDonald, "Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods," in *Proc. 6th Nordic Conf. Human-Comput. Interaction: Extending Boundaries*, 2010, pp. 581–590.
[4] E. L. Olmsted-Hawala, E. D. Murphy, S. Hawala, and K. T. Ashenfelter, "Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability," in *Proc. 28th Int. Conf. Human Factors Comput. Syst.*, 2010, pp. 2381–2390.
[5] J. Nielsen, *Usability Engineering,* 1st ed. San Mateo, CA: Morgan Kaufmann, 1993.
[6] J. Gulliksen, I. Boivie, J. Persson, A. Hektor, and L. Herulf, "Making a difference—A survey of the usability profession in Sweden," in *Proc. 3rd Nordic Conf. Human-Comput. Interact.*, 2004, pp. 207–215.
[7] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data Revised Edition.* Cambridge, MA: MIT Press, 1993.

[8] M. T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *IEEE Trans. Prof. Commun.*, vol. 43, no. 3, pp. 261–277, Sep. 2000.
[9] M. Nørgaard and K. Hornbæk, "What do usability evaluators do in practice?: An explorative study of think-aloud testing," in *Proc. 6th Conf. Design. Interact. Syst.*, 2006, pp. 209–218.
[10] Q. Shi, "A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests," in *Proc. 5th Nordic Conf. Human-Comput. Interact.: Building Bridges*, 2008, pp. 344–352.
[11] A. Lund, "Post-modern usability," *J. Usability Studies*, vol. 2, no. 1, pp. 1–6, 2006.
[12] S. Rosenbaum, "The future of usability evaluation: Increasing impact on value," in *Maturing Usability: Quality in Software, Interaction and Value*, E. Law, E. Hvannberg, and G. Cockton, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 344–378.
[13] B. Still and M. J. Albers, "Editorial: Technical communication and usability studies," *IEEE Trans. Prof. Commun.*, vol. 53, no. 3, pp. 189–190, Sep. 2010.
[14] V. A. Bowers and H. L. Snyder, "Concurrent versus retrospective verbal protocols for comparing windows usability," in *Proc. 34th Annu. Human Factor Soc. Meeting*, 1990, pp. 1270–1274.
[15] M. J. van den Haak, M. D. T. de Jong, and P. J. Schellens, "Evaluation of an informational web site: Three variants of the think-aloud method compared," *Tech. Commun.*, vol. 54, no. 1, pp. 58–71, 2007.
[16] N. Miyake, "Constructive interactionand the iterative process of understanding," *Cognit. Sci.*, vol. 10, pp. 151–177, 1986.
[17] C. M. Barnum, *Usability Testing Essentials: Ready, Set . . .Test!*. San Mateo, CA: Morgan Kaufmann, 2010.
[18] M. J. van den Haak, M. D. T. de Jong, and P. J. Schellens, "Constructive interaction: An analysis of verbal interaction in a usability setting," *IEEE Trans. Prof. Commun.*, vol. 49, no. 4, pp. 311–324, Dec. 2006.
[19] T. Grossman, G. Fitzmaurice, and R. Attar, "A survey of software learn ability: Metrics, methodologies and guidelines," in *Proc. 27th Int. Conf. Human Factors Comput. Syst.*, 2009, pp. 649–658.
[20] A. Følstad, E. L.-C. Law, and K. Hornbæk, "Analysis in usability evaluations: An exploratory study," in *Proc. 6th Nordic Conf. Human-Comput. Interact.: Extend. Boundaries*, 2010, pp. 647–650.
[21] J. S. Dumas and J. Redish, *A Practical Guide to Usability Testing Revised Edition*. Exeter, UK: Intellect, 1999.
[22] J. Rubin and D. Chisnell, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 2nd ed. Hoboken, NJ: Wiley, 2008.
[23] J. S. Dumas and B. A. Loring, *Moderating Usability Tests: Principles and Practices for Interacting*. San Mateo, CA: Morgan Kaufmann, 2008.
[24] T. Clemmensen, M. Hertzum, K. Hornbæk, Q. Shi, and P. Yammiyava, "Cultural cognition in usability evaluation," *Interact. Comput.*, vol. 21, no. 3, pp. 212–220, 2009.
[25] H. Tamler, "How (much) to intervene in a usability testing session," *Common Ground*, vol. 8, no. 3, pp. 11–15, 1998.
[26] S. Makri, A. Blandford, and A. L. Cox, "This is what I'm doing and why: Methodological reflections on a naturalistic think-aloud study of interactive information behavior," *Inf. Process. Manage.*, vol. 47, pp. 336–348, 2011.
[27] I. Bark, A. Følstad, and J. Gulliksen, "Use and usefulness of HCI methods: Results from an exploratory study among Nordic HCI practitioners," in *People Comput. XIX*, 2005, pp. 201–218.
[28] K. Vredenburg, J.-Y. Mao, P. W. Smith, and T. Carey, "A survey of user-centred design practice," in *Proc. 20th Int. Conf. Human Factors Comput. Syst.*, 2002, vol. 4, no. 1, pp. 471–478.
[29] A. Følstad and K. Hornbæk, "Work-domain knowledge in usability evaluation: Experiences with cooperative usability testing," *J. Syst. Softw.*, pp. 2019–2030, 2010.
[30] P. K. Chilana, J. O. Wobbrock, and A. J. Ko, "Understanding usability practices in complex domains," in *Proc. 28th Int. Conf. Human Factors Comput. Syst.*, 2010, pp. 2337–2346.

Sharon McDonald is a Reader in Human-Computer Interaction in the Faculty of Applied Sciences, University of Sunderland, Sunderland, UK. Her current research interests are in the area of usability evaluation, specifically focusing on the use of think-aloud methods within usability testing and the development of methods for conducting field-based usability evaluation.

Helen M. Edwards is Professor of Software Engineering in the Faculty of Applied Sciences, University of Sunderland, Sunderland, UK. Her current research interests are in the area of socio-technical systems, specifically focusing on field studies of the user's experience of technology in use.

Tingting Zhao (M'11) received the M.A. degree in Business Studies and the M.Sc. degree in IT Application Development from the University of Sunderland, Sunderland, UK, in 2005 and 2007, respectively, where she is currently pursuing the Ph.D. degree in the Faculty of Applied Sciences. Her research interests involve usability evaluation and user experience, and her doctoral research focuses on the use of the concurrent think-aloud method within usability testing.