Thinking Aloud: Reconciling Theory and Practice

—M. TED BOREN
AND JUDITH RAMEY,
ASSOCIATE MEMBER, IEEE

Abstract*—*Thinking-aloud protocols may be the most widely used method in usability testing, but the descriptions of this practice in the usability literature and the work habits of practitioners do not conform to the theoretical basis most often cited for it: Ericsson and Simon's seminal work PROTOCOL ANALYSIS: VERBAL REPORTS AS DATA [1]. After reviewing Ericsson and Simon's theoretical basis for thinking aloud, we review the ways in which actual usability practice diverges from this model. We then explore the concept of SPEECH GENRE as an alternative theoretical framework. We first consider uses of this new framework that are consistent with Simon and Ericsson's goal of eliciting a verbal report that is as undirected, undisturbed, and constant as possible. We then go on to consider how the proposed new approach might handle problems that arise in usability testing that appear to require interventions not supported in the older model.

Index Terms*—*Speech genres, thinking-aloud protocols, usability research, usability testing.

Manuscript received November 18, 1999; revised May 9, 2000.
M. T. Boren is with Microsoft Corporation, Redmond, WA 98052-6399 USA (email: tboren@microsoft.com).
J. Ramey is with the Department of Technical Communication, University of Washington, Seattle, WA 98195–2195 USA (email: jramey@u.washington.edu).
IEEE PII S 0361-1434(00)07486-5.

As a way to gain insight into test participants' thought processes, usability professionals often ask test participants to "think aloud." This practice plays an important role in usability data collection; Nielsen even goes so far as to say, "Thinking aloud may be the single most valuable usability engineering method" [2, p. 195]. Yet there is no detailed description in the usability literature of theoretically motivated rules of practice for thinking aloud; some sources cite a theory, but then suggest theoretically inconsistent procedures. Most others do not describe think-aloud practice at all. Nor has there been a concerted effort to examine whether verbalization models developed within cognitive psychology apply equally well in usability testing, despite calls some years ago for such an effort (e.g., [3]). Finally, there is ample anecdotal evidence that think-aloud procedures vary widely among practitioners (e.g., [3], [4]).

Without a firm theoretical grounding in and unified practice of this key technique, it could be argued that usability practice is not governed by a framework of principles of inquiry—that while usability professionals may contribute positively to a system's development, usability practices are not systematic or rigorous enough to merit the distinction of being called a **method** (a defined practice with clear rules for correct performance). Many usability practitioners would bristle at such claims, but until the field can firmly establish that its practice is reconciled with a sound theory that supports its research goals, such characterizations cannot be lightly dismissed. Furthermore, if practitioners do not uniformly apply the same techniques in conducting thinking-aloud protocols, it becomes difficult to compare results between studies, both in industry and in the published methodological research.

We suggest that while current think-aloud practice is indeed diverse and not well-motivated by theory, practitioners can reconcile theory and practice, both by re-evaluating and systematically extending theory and by changing practice to be theoretically consistent.

ERICSSON AND SIMON [1] AS THE THEORETICAL BASIS FOR THINKING ALOUD

Often in the published literature, no source is cited in support of using verbalization. For example, Desurvire, Kondziela, and Atwood [5] specify in detail the methods used for several evaluation techniques being compared, but do not do so for usability testing—even though it was the benchmark method for comparing the effectiveness of the other techniques. When a source is cited (e.g., by Denning, Hoeim, Simpson, Sullivan [6], or by Nielsen [2]), it is almost invariably the work of Ericsson and Simon [1]. To gain the proper perspective on this work, we should first briefly review what came before it.

Earlier Uses of Verbalization Among the first psychologists to use verbalization were W. James, W. Wundt, and others who used introspection to investigate psychological claims and theories of mind. In introspection, trained subjects attempted to report directly their own mental processes. The resulting findings were difficult to replicate and were therefore attacked by J. B. Watson and other behaviorists, leading to the demise of introspection as a psychological research technique (as discussed in [7]). However, verbalization began to make a comeback in the 1960s, as the field of cognitive psychology expanded. Nisbett and Wilson [8] gave a comprehensive and strenuous critique against the use of such verbal reports. They cited numerous studies suggesting that human beings have no direct access to mental processes and therefore cannot accurately report information about them, hence the opening phrase of their article's title: "Telling more than we know."

Ericsson and Simon's Model of Verbalization Ericsson and Simon [1] agreed with many of Nisbett and Wilson's concerns [8], but they pointed out that the critique was leveled primarily at specific types of verbalization. Other types, or "levels," of verbalization might rightfully be considered data, if they were valued only as indicators of what information was heeded and in what order, a sort of time stamp of the contents of short-term memory (STM).

Ericsson and Simon defined three levels of decreasingly reliable verbalization, each characterized by the amount of interference caused by nontask-related processing:

Level 1 Verbalizations are those that need not be transformed before being verbalized during task performance: For example, subjects who verbalize sequences of numbers while solving a math problem are producing Level 1 data because numbers can be verbalized in the same form as they were originally encoded in STM. This is the most reliable sort of verbalization under Ericsson and Simon's model but is often difficult to obtain.

Level 2 Verbalizations are those that must be transformed before being verbalized during task performance: Images or abstract concepts, for example, must be transformed into words before they can be verbalized; as long as this transformation into words is the only mediating cognitive process between STM and verbalization, such verbalizations are Level 2 data. This is also considered reliable data under Ericsson and Simon's model.

Level 3 Verbalizations are those that require additional cognitive processing beyond that required for task performance or verbalization: Examples of additional cognitive processes include filtering processes (e.g., "verbalize only information related to topic X"), making inferences about the subjects' own cognition, and information retrieved from long-term memory at the researcher's request. Also, any outside influence, including any comment or prompt from the researcher, turns subsequent verbalizations into Level 3 data because the normal flow of information in STM during the task has been altered. Nisbett and Wilson and Ericsson and Simon argue primarily against using this kind of data.

Additionally, some kinds of verbalization cannot be considered **data** in any sense under this model because they "raise difficult problems of interpretation and analysis" [1, p. 223]—stream of consciousness, daydreams, feelings, and value judgments, for example.

To summarize, under the Ericsson and Simon model, Levels 1 and 2 data can be confidently relied upon, for if collected properly, they reveal what information the subject heeded and in what order. Unlike introspection and other techniques which value the actual content of verbalizations, Ericsson and Simon's model values "hard" (attentive, sequential) verbal data, which can be used to validate hypothesized cognitive models; in their model, under no circumstances are verbalizations to be valued for their subjective content.

Applying Ericsson and Simon in Usability Testing Despite the fact that Ericsson and Simon's model was developed within a relatively narrow problem space within the field of cognitive psychology, it has since provided the rationale for collecting verbal data in many other fields, including cartography (e.g., Suchan and Brewer [9]), various engineering fields (e.g., Sanderson

[10]), reading comprehension research (e.g., Pressley and Afflerbach [7]), and usability testing (e.g., Denning et al.,[6] and Nielsen [2]). Some of these fields have research goals similar to those of cognitive psychology; they seek to understand how cognitive processes function in specific domains. Other fields, including usability research, have an interest in cognition but have their primary focus elsewhere. In the case of usability research, the primary concern is to support the development of usable systems by identifying system deficiencies; building robust models of human cognition is not a central concern. Furthermore, much of the cognition that is of interest is manifest not only through verbalizations but also through interactions with the system; this provides a useful context for interpreting and determining the validity of verbal data, as Nielsen has suggested [2]. Finally, the system being tested in usability testing is likely to display much more variability of behavior (including highly unexpected behavior) than a test apparatus used in cognitive research; such systems are more complex, are usually under development, and contain many experimental unknowns that are difficult to control. Such differences in research goals and context convinced Deffner [3] that a more complete review of the generalizability of Ericsson and Simon's theory to human factors and usability research was long overdue. Unfortunately, the colloquy that Deffner subsequently coordinated only scratched the surface of this complex topic, and little has since been added to it.

Despite the potential challenges to applying Ericsson and Simon's model to usability testing, it is currently the only generally cited theory for eliciting verbalizations in usability tests, so we should examine its implications for usability research. It would require a longer work to fully explore such implications, but the following key points from Ericsson and Simon may suffice for our purposes:

Collect and Analyze Only "Hard" Verbal Data: Procedures used to collect verbal data must withstand Nisbett and Wilson's critique of verbal data [8]. Participant introspection, inference, or opinion must not be valued or actively elicited. The only data considered must be what the participant attends to and in what order, in other words, Level 1 and 2 data as described earlier. Information that would not normally be present in STM during a task (Level 3 data or nondata) should be considered suspect or irrelevant.

Give Detailed Initial Instructions for Thinking Aloud: Among other things, researchers should distinguish between explanation and thinking aloud, encourage the participant to speak constantly "as if alone in the room" without regard for coherency, and inform the participant that reminders will be given if he or she falls silent. Participants should also practice thinking aloud before the test begins.

Remind Participants to Think Aloud: Reminders should come after a predetermined period of silence (perhaps 15–60 seconds) and should be as short and nondirective as possible. Reminders should also not encourage a sense of personal contact or heighten awareness of the researcher's presence. A successful reminder should result in the immediate resumption of thinking aloud, without pause for reflection or retrospection. "Keep talking" is Ericsson and Simon's recommended reminder.

Otherwise, Do Not Intervene: After a task begins, the only interaction should be the reminder to think aloud, as needed. Any other interactions, including "neutral" questions or comments, taint subsequent performance and verbalization by redirecting attention.

Practitioners relying on Ericsson and Simon should base their think-aloud methodology on these core principles enumerated in *Protocol Analysis* [1]. (Interestingly, these principles are not fully articulated in the published literature; two unpublished sources that give more detail on how verbal protocols might be conducted under Ericsson and Simon's model are a master's thesis by Boren [11] and a memo by McClintock [12], which is included with permission in the appendix to Boren's thesis.)

Theory and Practice Are Out of Sync

Over a decade ago, Deffner noted that evaluating the effectiveness of verbal protocols in human factors and usability testing was difficult because of the "great diversity in procedures employed by various researchers," different instructions to participants, different degrees of intervention, and different research motivations [3, p. 1263]. Nor has the situation changed much in the intervening years; Tamler cites widely varying opinions among prominent practitioners on the topic of tester intervention, ranging from completely "passive, unobtrusive observation" to "pragmatic flexibility" to "full partnership" with users [4, p. 11]. If Ericsson and Simon were being consistently applied, this level of variation should not exist. To gain a better understanding of the sources of these apparent discrepancies, we conducted a literature review heavily supplemented by a series of field observations, interviews, and artifact collections aimed at broadening our understanding of how verbal protocols were collected and subsequently analyzed.

Background on Our Exploratory Field Research Our exploratory field research is reported in Boren's thesis at the University

of Washington [11], where Ramey was serving as Boren's graduate advisor. Since access to theses is somewhat limited, we here give some background on that research to give context to the findings cited.

After performing nine preliminary observations to help refine our scope and data-collection tools, we conducted seven primary observations at two software companies, one with 2–5 usability professionals and another with more than 20 usability professionals. All but one of the usability professionals observed in these primary observations held advanced degrees in psychology or technical communication, and all but one had at least four years experience conducting usability tests. Many of these professionals were regarded by their peers as among the most methodologically rigorous at the company. All findings reported in this paper come from these primary observations.

During each test session, Boren took notes and made an audio recording of the interactions between usability professionals and their test participants. After each session, an interview was conducted to gather additional information about the usability professional's methods. Afterwards, a word-for-word transcript was made of the test session, including such features as overlapping speech, approximate length of pauses, and inflection. These transcripts were then analyzed to produce the findings reported below.

Examples of Discrepancies Between Theory and Practice Our field observations, combined with a review of the usability guidebooks and literature, provide many clear examples of differences between practice and theory as articulated by Ericsson and Simon.

Usability Practitioners Often Do Not Give Think-Aloud Instructions in the Prescribed Manner: Some of the guidebooks do not suggest any explicit instructions to give participants (Nielsen [2]; Rubin [13]), though they do suggest that practitioners should model thinking aloud (Nielsen and Rubin as well as Dumas and Redish [14]). Dumas and Redish give the most complete guidance, including instructions quoted from Ericsson and Simon, a detailed description of modeling, and detailed suggestions for practice. Others give more general accounts of usability testing practices (e.g., [15]).

In the field, however, we observed wide variations in how participants are instructed to think aloud [11]. Some of this variation was due to time constraints in three of the sessions, but there were also serious discrepancies between instructions in the remaining four sessions and Ericsson and Simon's suggested instructions. For example, several usability engineers did not explain the difference between thinking aloud and explanation, and some even requested explanations of behavior outright. Some published research also includes explicit instructions to report explanations (e.g., [16]); this seems likely to account for the reported discrepancies between such research and that of others (e.g., [17]) and provides a good example of why a consistently applied method is important.

Finally, in our field observations, only one usability specialist had the participant practice thinking aloud, and only one other substituted modeling for practice. This is surprising given the emphasis placed on extensive, even repeated, practice by Ericsson and Simon and the guidebooks cited earlier.

Usability Practitioners Often Do Not Give Reminders in the Prescribed Manner: Dumas and Redish, for example, cite Ericsson and Simon yet suggest reminders that are inconsistent with Ericsson and Simon's theory. Perhaps the most thoroughly incompatible reminder they suggest is, *John, could you tell us why you pressed the enter key?* [14, p. 281]. This reminder is long compared to the recommended *Keep talking,* redirects attention to *the enter key*, establishes personal contact between "John" and "us," requires a pause in the task flow, elicits retrospection, and requests an explanation of behavior. In a similar vein, Nielsen suggests saying, *What do you think this message means?* as a way to get a silent participant to keep talking [2, p. 196]. While these questions may certainly be of interest to the usability professional, such reminders simply cannot be offered under the Ericsson and Simon model.

We also found discrepancies among the prompts used in our field observations [11]. Reminders ranged from short to long, from directive to nondirective, from personal to impersonal, and everywhere in between. Here is a sampling (some of these reminders are taken directly from the transcripts and are not included in [11]):

- *OK*
- *Keep talking.*
- *Any thoughts about that*?
- *Don't forget to tell me what you're thinking*?
- *Is there anything special you're looking for…*?
- *Yeah, even as you're reading, just read that out loud so I can hear you better.*

Not only is the phrasing of actual reminders often at odds with Ericsson and Simon, but so is their timing. The published usability literature and guidebooks say virtually nothing about how long to wait before prompting, but Ericsson and Simon imply that experimenters should prompt after a predetermined interval of between 15 and 60 seconds, depending on the research goals. In field observations, however, we

noted that most of the usability professionals observed had no regular prompting interval [11]. Rather, there was wide variation within test sessions; sometimes a practitioner would prompt very quickly while in other cases he or she would wait much longer, or never prompt at all. The average disparity between the quickest prompt and the longest delay without prompting was about 16 seconds; in other words, in the typical test session, a usability professional might prompt as quickly as 5 seconds in one instance, but go as long as 21 seconds without prompting in another instance. Since the median delay before prompting was only 10 seconds, this 16 second disparity is huge in proportion to the total time participants were silent. Just as interesting, many usability professionals chose never to prompt at all; of 23 pauses extending 15 seconds or longer, only 4 were ended with a prompt; the rest were allowed to continue until participants resumed thinking aloud on their own.

These results suggest that rather than having a set waiting period before prompting, as suggested by Ericsson and Simon, usability practitioners were influenced by contextual factors. Such factors include placement in the task (usability professionals tended to prompt more during the first third of a task); the participant's state of mind (frustrated participants were sometimes "left alone"); and the usability practitioner's assumptions (if the current activity didn't seem problematic or if a problem's source was presumed known, participants were often "left alone") [11].

Usability Practitioners Often Intervene in Theoretically Inconsistent Ways: The guidebooks consistently recommend caution in intervening (e.g., [13], [14], [2]), but even their qualified exceptions constitute serious challenges to the Ericsson and Simon model. For example, they suggest that usability professionals can use neutral "probes" to gather more information during diagnostic usability tests [13], [14]. But even neutral questions are not allowed under Ericsson and Simon's model, because they redirect attention and interrupt the task flow; even if the data gleaned is valuable and reliable, any subsequent verbalizations must be considered Level 3 data. The same is true of Nielsen's reminder cited earlier and the "answer a question with a question" strategy many of the guidebooks suggest [2], [14]; even if these approaches produce useful and reliable information for the moment, they pose a threat to subsequent verbalizations under Ericsson and Simon's model. The guidebooks cited also give advice on how to intervene when the test situation requires it (the participant is stuck or the system crashes, for example). Such situations have few corollaries in cognitive psychology, so Ericsson and Simon's suggested procedures understandably do not address such issues.

Likewise, in the field we found that usability engineers frequently intervened in ways outside the limits that Ericsson and Simon define [11]. In fact, of roughly 125 interventions observed across all seven sessions, only 16 of them were reminders to think aloud. About half of these interventions were initiated by the usability engineer; the other half were initiated by test participants or by system breakdowns. Usability engineers intervened to probe a particular area of the software, to help a participant who was "stuck," to request clarification of a comment, to clarify task instructions for a participant, to help participants get around software problems, and so on. The complexity of the testing environment and practitioners' research goals compelled them to intervene, even when doing so would produce suspect or irrelevant data under Ericsson and Simon's model. This was true even for usability professionals who consciously based their method on Ericsson and Simon.

Usability Practitioners Often Do Not Use Verbalizations as "Hard" Data: Many published researchers seem to value verbal data in ways completely at odds with Ericsson and Simon's model. For example, many prefer explanations, coherency of verbalization, and participants' design or revision ideas over strictly procedural information—suggesting that Level 3 data is of greater importance than Level 1 and 2 data [17]–[19], [14]. Further, Dumas and Redish explicitly categorize verbalizations as subjective. During practice, they also suggest that usability professionals should reinforce any evaluative statements that participants make, even though evaluation is at best Level 3 data.

Likewise, Rubin suggests that an advantage of having participants think aloud is that "you are able to capture preference and performance data simultaneously" [13, p. 218]. Similarly, Dumas and Redish suggest questions probing how participants feel, while Ericsson and Simon explicitly exclude feelings from consideration as data. And in the field we found that usability engineers often asked participants to tell "what you like, what you don't like," and that during data analysis they seldom relied on what Ericsson and Simon would consider "hard" verbal data to build or validate cognitive models, instead using the continuous verbalization to help alert them to where trouble was occurring and gain some coarse level of understanding of participants' goals and motivations during the tasks [11]. Only rarely did practitioners say they analyzed verbalizations closely, and then only for particularly problematic segments of action; in other words, there was no protocol analysis.

Finally, usability engineers often noted that the experience of observing a usability test or at least of reading quotes from participants was often the most powerful tool for convincing team members that a problem exists; this sentiment is echoed by Desurvire et al.[5] and Rubin [13], even though such a use of verbalization would not constitute "hard" proof.

Sources of Disparity Between Theory and Practice Despite the frequent citation of Ericsson and Simon [1] justifying the collection of verbal data, most usability practitioners do not apparently subscribe to Ericsson and Simon's model in actual practice. Published usability research, usability handbooks, and field observations all suggest that theory and practice are out of sync. Perhaps just as problematic, the discrepancies are not uniform between different usability professionals. Why such profound theoretical divergence and methodological variation?

Perhaps, like Nielsen [2], [20], some practitioners feel that Ericsson and Simon's procedures are more complex than their needs require, so they simplify them in some way. (Nielsen endorses a "simplified think aloud method" over "traditional think aloud methodology" based on Ericsson and Simon, but defines simplified thinking aloud in very loose terms and gives no corresponding theoretical support or citations for the simplifications and other procedures he suggests.)

Others may simply want to collect data that is not collectable under Ericsson and Simon's model. And many likely find that Ericsson and Simon do not adequately address some of the contingencies they regularly face in usability sessions—computer system crashes, software bugs, and incomplete prototypes that require frequent and sometimes extended intervention, all of which interrupt the monolog required by Ericsson and Simon's model; research questions that are task- and system-focused instead of cognition-focused; complex tasks and interfaces that tax cognitive resources and cause participants to fall silent. These and other common contingencies are fundamentally hostile to the collection of a **pure** verbal protocol by Ericsson and Simon's definitions, so usability practitioners are forced to improvise. But the resulting procedural deviations are rarely informed by a theoretical framework, Ericsson and Simon's or any other. Lacking the guidance of unifying principles, such changes vary greatly in degree and kind, not only from Ericsson and Simon's model, but from each other.

Reconciling Theory and Practice This general lack of methodological consistency and theoretical justification is a deficiency in the field. Given the current situation, practitioners cannot give a coherent rationale for doing what they do. If usability testing is to be considered a discipline, the methods employed by its practitioners must be theoretically motivated and systematically applied. Otherwise, how can we compare or replicate studies, vouch for the validity of the results, or teach a standard of practice to newcomers to the field?

Practitioners could reasonably reconcile theory and practice in the following ways.

Stop Collecting Verbal Protocols; Gather Data Strictly on Observable Performance: This eliminates the discrepancy, but it also eliminates a valuable information source. Further, it does not address situations requiring some sort of communication (system crashes or stymied participants, for example).

Start Strictly Applying Ericsson and Simon's Theory: This means much less intervention and fairly clear guidelines for practice. However, the practicality of this approach is challenged by complex interfaces, self-conscious participants, tight deadlines, and incomplete or buggy software or prototypes. Further, practitioners who value Level 3 data would either have to do without it, or collect it only through time-consuming and somewhat less reliable retrospective reports (e.g., [17]).

Explore Alternative Theoretical Positions, such as Speech Communication: Such alternatives might better inform data collection procedures and help practitioners manage the inevitable contact with participants, not only when they fall silent, but throughout the many challenging interactions that arise during a test. Nisbett and Wilson's [8] critique must still be addressed or shown not to be applicable to the research questions of interest.

While arguments could be made for any of these approaches, we believe that the third has great potential for a providing practical, theoretically sound basis for usability practice. The remainder of this article proposes that theories of speech communication be further investigated and applied to usability testing.

Proposal: Explore Speech Communication as a Theoretical Alternative

Ericsson and Simon assume that verbal processes can be completely divorced from their original purpose, which is to communicate. Proponents of constructive interaction have criticized this assumption and suggested that speech (any speech) is best treated as communication (e.g., O'Malley et al.[21]). Some have even applied conversation analysis and communication theory to the analysis of transcripts from constructive interaction sessions (those involving at least two people) [22]. Similarly, we propose that speech communication theory be explored as a theoretical basis for thinking aloud in usability tests,

focusing on researcher–participant communication rather than participant–participant communication (as in constructive interaction).

The "Fiction" of the Passive Listener Ericsson and Simon take the stance that the listener is entirely passive, or even nonexistent—that subjects are able to temporarily suspend their awareness of any listener, so that they can comply with the request to "think aloud as if you were alone in this room." But in a usability test, there is always a listener (primarily the usability practitioner) and a speaker (primarily the participant). The issue of how (much) to intervene [4] is largely one of defining these roles of listener and speaker.

According to speech communication theory, any time words are spoken knowingly for another's benefit, the roles of speaker and listener exist; both parties are aware of and are reactive to each other. Listener and speaker may "accomplish asymmetrical communicative roles" with the speaker doing most of the talking and the listener responding either much or little, but the listener **must** necessarily respond [23, p. 157]. Bahktin, for example, argues strongly against theoretical representations of "the passive listener," one whose sole function is to absorb information from the speaker [24, p. 68]. He calls such a construct "a fiction," claiming that whenever a listener perceives and understands speech, he or she "simultaneously takes an active, responsive attitude toward it. He either agrees or disagrees with it (completely or partially), augments it, applies it, prepares for its execution, and so on.... Any understanding of live speech, a live utterance, is inherently responsive, although the degree of this activity varies extremely" [24] In a similar vein, Goodwin calls for "systematic study of the actions of hearers," to support the perspective that "talk is not simply a form of action" performed by the speaker alone, "but a mode of interaction" between speakers and hearers [25, p. 205].

Ericsson and Simon's implied position is that listeners can choose not to share their response, thus eliminating any influence on the speaker. Goodwin, however, cites research showing that "speakers may modify the emerging structure of the sentence they are producing in response to what the recipient is doing (or not doing)" [25, p. 206]. In other words, speakers cannot ignore listeners, even silent ones. Speakers expect that listeners will react to what they say, and that listeners' actions (or inactions) are reflective of that response. Bahktin suggests that "the speaker himself is oriented precisely toward such an actively responsive understanding. He does not expect understanding that, so to speak, only duplicates his own idea in someone else's mind. Rather, he expects response, agreement, sympathy, objection, execution, and so forth" [24, p. 68]. This, says Bahktin, is true whether the listener's response is immediate, delayed, or never shared. Assuming otherwise turns the listener into a "fiction," a representation "not accompanied by any indication of the great complexity of the actual phenomenon," a distortion that removes "precisely [speech communication's] most essential aspects" [24, p. 70].

If we accept this view of human communication, the issue in usability testing becomes not one of disappearing from participants, but rather one of creating a highly asymmetrical speaker/listener relationship, one which maximizes the speakership of the participant and minimizes the speakership of the usability practitioner. This relationship and the context in which it can exist may set the stage for a speech genre conducive to collecting usability data.

Speech Genres According to Bahktin, "each sphere in which language is used develops its own **relatively stable types** of... utterances" called SPEECH GENRES [24, p. 60]. These speech genres "reflect the specific conditions and goals" of the area of human activity in which they occur. The word choice, phrasing, and above all the compositional structure of utterances are all shaped by the "particular sphere of communication," including the roles assumed by the parties, the physical context, and the timing of the interaction. In other words, people employ different speech genres depending on the *who*, *what*, *where*, *when*, and *why* of the communication.

Who is being spoken to, and what is their perceived role? What is being discussed? Where and when is the communication taking place? Why is the communication necessary? What is the goal? The combination of these factors invokes a set of rules, a speech genre, that determines what responses are appropriate. If the parties do not know or choose not to abide by the rules of the genre, communication is impeded. Goodwin, for example, cites the case of a tribal leader who became so frustrated with an ethnographer's inability to "listen with the appropriate responses" to an official speech, that he finally sent for a native "'listener' before he could talk at all" [25, p. 206]. In this case, the ethnographer knew the mechanics of the language, but was so unfamiliar with the rules of the speech genre that communication was impossible.

In the context of usability testing, such dramatic failures of communication are unlikely to occur, but awareness of speech genre is nevertheless important. To a great degree, the usability practitioner is the one who sets up the test conditions and shapes the resulting speech genre by implicitly or explicitly defining the who, what, where, when, and why

of the interaction. In turn, the speech genre invoked determines the responses available, both to usability practitioners and to participants. For example, usability practitioners can invoke less useful speech genres by setting themselves up as the expert and the participant as the subject; this may invite a range of unproductive responses from participants, including those marked by self-consciousness, apology, or tentativeness. Or usability practitioners may downplay their own expertise and affirm the participant's contribution; this approach may help participants focus less on the novelty and potential pressure of the situation and more on immersing themselves in the task.

Once the stage is set and a productive speech genre has been invoked, usability practitioners are responsible for channeling the communication toward the goals of the study. Failing to consciously monitor interaction styles and speech genre may be tantamount to not caring what kind of participant is completing the tasks. Usability professionals should consciously ask themselves who they want working with their products—a self-conscious, apologetic subject, or an inquisitive work domain expert immersed in the task at hand—and then work to create an environment in which that kind of participant can exist.

Proposed Areas of Investigation for Speech Communication and Usability Testing

If we accept this view of verbalization as speech communication, how should we conduct usability test sessions? This section outlines four areas of proposed methodological change. Each requires further thought and research, but they also seem to offer the possibility of a reconciled theory and practice. They are listed here in increasing order of conflict with Ericsson and Simon's model:

- First, we describe how usability practitioners could use speech communication theory to help set the stage for productive interaction.
- Second, we suggest modifications to data-collection methods which would take into account the social nature of speech, but which work toward the same goal as Ericsson and Simon's procedures: eliciting a verbal report that is as undirected, undisturbed, and constant as possible.
- Third, we consider interventions often required by the contingencies of usability testing, but for which Ericsson and Simon give little guidance.
- Finally, we discuss proactive interventions that elicit information that is not allowable under Ericsson and Simon's model.

The first modifications suggested above depart somewhat from Ericsson and Simon's prescriptions; each subsequent modification suggests departures that are more and more pronounced, thus raising more complex theoretical challenges. Rigorous examination and testing of these proposed modifications is both welcome and expected.

Setting the Stage The guidebooks discuss how important it is to put participants at ease (e.g., [2], [13]); some of them even devote whole sections to the topic (e.g., [14]). Speech communication theory can give theoretical motivation and guidance on this topic by helping usability practitioners create an environment in which participants retain primary speakership while completing tasks as naturally as possible, even in patently unnatural situations. Practitioners can begin by consciously considering three important elements that will fundamentally shape how test interactions unfold: practitioner and participant roles, physical environment, and timing of the interactions.

Working under a speech communication model, usability professionals should explicitly define the roles and status of three parties:

The Product Being Tested (Software Interface, Document, etc.) is the Subject: Participants must understand that usability testing is about discovering whether the products tested work well for the people who need to use them; the product, not the person, is the object of study. (Note how different this social dynamic is from much research in cognitive psychology, where the **person** is the subject of the experiment and any test apparatus is simply an object to be manipulated by the subject.)

The Participant is the Work Domain Expert and Primary Speaker: The participant should be cast as an important contributor, work domain expert, or valued customer. Representing users as experts and researchers as apprentices is commonly done in fieldwork, where users are truly in their own element. But the same concept can apply to a lesser degree even in the lab; Tamler [4] hints at this when he suggests that the "full partner" model of contextual inquiry may also have a place in lab testing. It is important, however, that in accepting this role, participants realize that it is their experience in the work domain, not any insights they might have as an amateur interface critic, that is their peculiar contribution; they can approach the tasks in a way the designers cannot.

The Usability Practitioner is the Learner and Primary Listener: The main roles for the usability practitioner are as interested learner and listener. Usability professionals may also need to take on other, sometimes more authoritative roles, particularly at the beginning of usability

sessions. For example, a usability practitioner plays host while introducing the participant to the facilities and seeing to the participant's needs. Usability practitioners may also sometimes take the role of technical expert to troubleshoot software or guide the participant through an incomplete prototype. But these secondary roles should be "played small," and need no special introduction or emphasis.

These roles need to be both defined explicitly at the beginning of a test, then maintained implicitly throughout. Anything that subordinates the participant, shifts attention or speakership to the usability practitioner, or otherwise interferes with these primary roles should be avoided. Later sections will discuss in more detail how this might be done.

Defining roles sets up the "who, what, and why" of the communication. We should also address the "where and when;" physical environment and timing are subtle but powerful factors influencing a participant's perceived role. One-way mirrors, microphones, speakers, and video cameras all accentuate a participant's awareness of being observed—despite reassurances that the **system** is being tested. Usability practitioners should therefore make observational technology as unobtrusive as possible. (It is questionable whether state-of-the-art facilities actually provide sufficient benefits to justify their cost. Ericsson and Simon simply suggest that researchers remain out of sight but in the same room as the participant to avoid intimidating observational tools. Also, lighting requirements make it very difficult to view tasks through a one-way mirror; directly reading from a computer is nearly impossible.) They can also adjust the furniture and computer equipment and provide simple amenities—anything to help participants become immersed in the role of work domain expert instead of test subject. Likewise, the time of day, length of time into the test, and past experience with usability testing may all affect how participants verbalize. Usability practitioners must be sensitive to these temporal and environmental factors and make appropriate adjustments to achieve the goals of the study.

Using the Nature of Speech to Keep Verbal Reports Undirected, Undisturbed, and Constant Once the stage is set, usability practitioners should maintain conditions that help participants complete tasks with a minimum of direction or distraction, while verbalizing as constantly as possible. Ericsson and Simon maintain that this is best done by carefully instructing participants at the beginning of the test session, then cutting off any further communication, except for reminding participants to keep talking if they fall silent; their claim is that this approach minimizes contact and therefore interference. But a speech communication perspective may well interpret silence interspersed with commands to keep talking as a more **abrasive** form of contact. Ericsson and Simon's goal of eliciting verbalizations that are as undirected, undisturbed, and constant as possible may be better achieved by applying speech theory to account for the ways in which human beings naturally communicate.

Use Acknowledgment Tokens Continuously: We have already discussed the centrality of the active listener in speech communication theory. Usability practitioners adhering to a speech communication model must therefore decide how to acknowledge—or not acknowledge—the participant. There is a surprisingly rich body of literature on acknowledgment, particularly on the role of "acknowledgment tokens" (for example, *OK*, *yeah,* or *mm hm*). Careful use of these token responses can help the usability practitioner provide the response expected of engaged listeners, while still "lying low" and promoting the participant's speakership.

For example, research suggests that speakers need acknowledgment from listeners at certain fairly predictable transitions in their own talk. One kind of acknowledgment token (the CONTINUER) shows that the recipient is "prepared for movement to a new unit" of thought, and that the speaker can then "feel free to begin the next unit," having received "a signal to continue" [25, p. 208]. For example, we have observed usability sessions in which the participant essentially refuses to move on until being acknowledged—even when the practitioner running the study had hitherto adhered fairly closely to Ericsson and Simon's recommendations [11]. Paradoxically, silence from the usability professional seemed distracting; without acknowledgment, the participant apparently felt the need to "check the connection." While counter to accepted experimental procedures, participants may need a continuous yet unobtrusive stream of acknowledgment to keep focused on their tasks while producing fluent, uninterrupted verbalizations.

The Choice of Token Affects the Available Responses: But what constitutes an acknowledgment token, and are some better than others? Language researchers used to conceive of acknowledgment as occupying a "back channel" of communication through which feedback flowed from the listener to the speaker, indicating that the channel was open and communication was being received. According to this back-channel model, "all nonspeakership is alike" and all forms of back-channel acknowledgment "have an equal (and minimal) impact on an encounter's course"

[23, p. 161]. Some usability practitioners have adopted this back-channel concept, suggesting that practitioners provide "uncommitted sounds like 'uh huh' to acknowledge comments from the user and to keep the user going" [4, p. 11]. The implication is that it does not matter much what the "sound" is as long as it is "uncommitted."

More recent research shows that not all such sounds act strictly as acknowledgment tokens, nor are such sounds "uncommitted." Goodwin [25, p. 207] found that while *uh-huh* often functions as a continuer, *Oh* assesses "what was said as something remarkable;" *Oh* is an "assessment" rather than a "continuer." Similarly, some acknowledgment tokens turn into affirmations depending on context. For example, *OK* or *yeah*, while often used to acknowledge the speaker, also carry stronger connotations of agreement than other tokens [26], [27]. This seems especially true of *yeah* when the primary speaker has just asked for agreement [27].

Further, there are systematic differences in how acknowledgment tokens affect speakership changes [23]. By combining techniques of distributional analysis (to see how frequently different tokens are used) and conversation analysis (to account for the context of each occurrence instead of lumping all instances of a token into one bucket), Drummond and Hopper showed that some acknowledgment tokens are more likely to be followed by a change in speakership. For example, in the following list of acknowledgment tokens, those listed first exhibited lower speakership incipiency and those listed later exhibiting increasingly higher speakership incipiency:

- *Mm hmm*
- *Uh huh*
- *Yeah*
- *Oh*
- *OK*

In other words, the research showed that *Mm hm* is used primarily to encourage "the teller to continue" [23, p. 163]; listeners who said *uh huh* or *mm hm* took over speakership in only 4–5% of occurrences. By contrast, listeners who said *yeah* took over speakership 45% of the time; listeners who used *Oh* and *OK* took speakership even more frequently. Condon [26] similarly noted that *OK* seems to occur prominently at changes of speakership.

Condon further notes that the effect of acknowledgment tokens seems "to be derived as much from context and intonation" as from the sound itself [26, p. 73]. More specifically, Drummond and Hopper note that "ending a statement with a stretched, interrogative intonation contour" creates a "slot" for an acknowledgment token [28, p. 204]. If that slot is filled by a continuer that also displays an interrogative intonation, another "slot" is opened up for the previous speaker to retain speakership, as in the following example [28]:

D: *...I got three extra points so I made a ninety-six on my test?*
M: *Uh-huh?*
D: *And then I got extra points for ...*

Using the right combination of intonation and token, the listener (M) uses just two syllables to acknowledge the speaker (D), while at the same time smoothly ceding back speakership to D.

From this research in speech communication, we might conclude that *mm hm* or *uh-huh* followed by an interrogative intonation would be the most appropriate acknowledgment token during usability tests. These tokens rarely occur at changes in speakership, and when given an interrogative intonation, they may seamlessly invite the participant to "please continue"—all without introducing any task-relevant content, diverting attention to anything in particular, or requiring further elaboration on the participant's part. What might be perceived as a filler word or uncommitted sound [4] is actually not "uncommitted" at all; acknowledgment tokens may subtly but surely confirm that "the connection is good" and that the listener has received the message, thus freeing the participant to move on.

Frequency of Acknowledgment: It is difficult to prescribe how often usability practitioners ought to acknowledge participants because the best cues will likely come from the participants themselves. If participants seem to be making a transition [25] or if they have "opened a slot" for response by using an interrogative intonation themselves [28], these may be particularly appropriate times to acknowledge participants. Acknowledgments should follow the flow of the current communication; they should not be forced but should rather be a fairly constant feature of the communication backdrop. (Test administrators, however, should be careful not to use an acknowledgment when the participants may take it as an affirmation that what they said is actually true.) However, maintaining this kind of contact is difficult over an intercom system, particularly if other observers are present. This may be an argument in favor of separate viewing areas, one for test administrators and another for observers, or for the test administrator remaining out of sight but in the same room as the participant.

Arguments Against Continuous Acknowledgment: Practitioners such as those cited by Tamler [3] may welcome the theoretical support and guidance provided by speech communication research. But others will doubtless be concerned that any response at all, no matter how minimal, has the potential to influence participants.

From a speech communication perspective, there are two main responses to this objection. The first is to return to the premise stated earlier that when speech is involved, silence is in fact a sort of response with attendant effects on participant behavior—quite likely greater effects than an unobtrusive acknowledgment token. The token is a natural, nonintrusive response; silence is an unnatural, potentially intrusive response. The only way around this is to claim that there is no listener as far as the speaker is concerned, which brings us back to the foundational premise that all live speech recognizes and anticipates an active, responsive listener [24]. In situations where speakers know their speech is heard, the researcher's silence indicates not the absence of a listener, but the presence of an uninterested or (worse yet) an aloof or condescending listener.

Another response to this objection is that even within the framework of cognitive psychology such tokens should have a minimal impact on task performance. For example, token-passing may well be pre-attentive; most people do not even realize the extent of token passing that occurs during normal speech. Further, token passing should introduce no new content into the speaker's short-term memory. Nor should tokens encourage the participant to attend to specific interface elements because tokens carry practically no content and should not redirect attention. A remaining concern may be that processing the token requires processing capacity and may therefore change task performance. But there is support for the idea that token passing requires little to no processing; for example, speakers regularly and predictably overlap the acknowledgment tokens of listeners, moving on with their own speech before the token has been completed [25], even in usability tests [11]. This suggests that participants may spend few cognitive resources on processing acknowledgment tokens.

Finally, it is interesting to note that Ericsson and Simon actually suggest that instead of saying *Keep talking* after a subject falls silent, a researcher might choose to play a "reminder tone" to keep the reminder short and completely contentless. *Mm hm*? has all the virtues of such a reminder tone plus the advantage of being far less conspicuous. The difference is that an acknowledgment token is given periodically, as the flow of communication demands, not just when participants fall silent, and tokens flow naturally into the communication milieu instead of intruding on it.

Remind the Participant to Keep Talking: Ericsson and Simon note that when the initial instructions are given properly, participants rarely stop thinking aloud. However, several factors affect the likelihood that a participant will stop verbalizing. For example, children under 10 or people with poor verbal skills may forget to think aloud more frequently [6] [12]. Also, task complexity can sometimes overwhelm the capacity of short-term memory (STM), increasing the likelihood that participants stop thinking aloud [6]. This seems particularly likely in diagnostic usability tests where participants are often performing tasks for the first time with unfamiliar tools. The use of acknowledgment tokens may partly diminish the need for reminders, but it seems not likely to eliminate it because of the complexity of the tasks and interfaces involved.

Reminders, when needed, should be unobtrusive. As mentioned earlier, Ericsson and Simon's preferred prompt is *Keep talking.* But our field observations, reminders to keep talking often resulted in apologies, interrupting the task flow and flustering the user [11]. While more research would be needed to determine causality, a speech communication approach might suggest that using an imperative (particularly an abrupt one like *keep talking*) casts the researcher into a more controlling, authoritative role and the participant into a subordinate role, a subordinate who has "messed up." An appropriately subdued command with a positive intonation, such as *Please remember to think aloud?*, may be able to elicit more satisfactory responses, but personal experience and observation indicate that saying those words without putting oneself in a superior, order-giving stance is very difficult.

Instead, usability practitioners might try reminders that reinforce the researcher's role as an engaged listener and encourage participants to resume speaking without making them overly conscious that they even stopped. A first step might be to extend an acknowledgment token such as *Mm hm*? even though there is nothing to acknowledge. If this does not work, the usability practitioner might progress to something more overt yet similarly content-free, such as *And now…*? We have tried these approaches with apparent success in our own usability work. Successful reminders would be those which result in an immediate resumption of concurrent verbalization without backtracking, apology, or explanation. Additional research into what reminders produce such results would be extremely valuable. Such research should consider not only the words used, but the intonation as well. In particular, it would be worthwhile to see if an interrogative intonation affects how a reminder is perceived. It would also be interesting to compare *Mm hm?* to the "reminder tone" suggested by Ericsson and Simon.

Handling the Contingencies of Usability Testing that REQUIRE Interaction Between Usability Practitioner and Participant Usability

practitioners running diagnostic tests must deal with a higher level of uncertainty and complexity than researchers conducting more formal experiments. For example, when cognitive psychologists have a human subject think aloud as they solve the Towers of Hanoi puzzle, the researchers know how the puzzle works and have defined ahead of time what its role in the test will be. In experimental terms, what is being studied is human cognition; the puzzle is simply the "test apparatus." In a diagnostic usability test, what is being studied is **the system itself**, not the cognition of the test participant. While both the system and the participant represent unknown quantities, the deficiencies within the system are the prime focus of the inquiry.

Furthermore, the systems being tested are usually "works in progress," resulting in anything from minor bugs to system crashes and from unfinished interface text to missing functionality. To further complicate matters, the entire interface is generally not being tested all at once, yet usability practitioners cannot usually "turn off" the parts that are not being tested; this presents additional opportunities for participants to go astray or encounter unexpected difficulties. Finally, usability participants are usually role-playing in order to complete the test scenarios; this often results in unexpected actions or questions from participants about exactly what the role or task entails. Not surprisingly, Ericsson and Simon do not say much about such situations because they do not need to deal with them. But how should usability practitioners proceed in such cases? They must generally say something—but what and how?

What follows is only a first take on how speech communication theory might help in these situations; we hope others will challenge, extend, and clarify these ideas with further library and laboratory research into speech communication.

The Trouble with Maintaining Monolog During a Usability Test: Before examining specific cases, we should identify a fundamental difficulty with using the Ericsson and Simon model for verbalizations during usability tests—the difficulty of maintaining monolog in the face of intermittent physical or verbal contact.

For example, when a system crashes, the usability practitioner must step in and get the test back on track. While Ericsson and Simon do not address the issue directly, we might assume one of two courses of action would be acceptable under their model:

(1) The researcher silently fixes the problem, signaling afterward for the participant to continue. In many cases participants would need some kind of guidance as to where to pick up in order to avoid crashing the system again, so this may be an impractical solution. (It would also seem rather awkward socially, but this is probably not a compelling argument in terms of the Ericsson and Simon model.)

(2) The researcher explicitly signals that the test procedure is being suspended, then steps the user through the adjustments, then indicates how to proceed with the test. Next, the researcher would need to signal that the interruption had ended and the user was to resume thinking aloud, perhaps with "Now continue with the task, thinking aloud as you do" [12]. This is, in effect, an indication for the participant to end dialog and resume monolog.

Either solution presents serious problems if we adhere to Ericsson and Simon's model, which assumes that participants can carry on a monolog, forgetting momentarily that anyone else can hear them. But both of these solutions bring the researcher and participant into direct verbal or physical contact during the test itself.

Nor is a system crash the only situation requiring at least minimal intervention. Even if a participant could imagine initially that he or she were "alone in this room," the combination of bugs, crashes, missing functionality, and participant uncertainty about the meanings or goals of tasks almost guarantees that the practitioner and participant will have to interact at least a few times during the course of a diagnostic usability test. Can users switch between dialog and monolog at will? Can they recreate the illusion that they are again "alone in this room?"

Rather than bearing the weight of these assumptions, a speech communication approach can suggest ways to manage the transition between a highly asymmetrical dialog (before the intervention), to a more equally interactive dialog (during the intervention), back to a highly asymmetrical dialog (after the intervention). There is never any question of monolog, since we assume instead that there is always both a speaker and a listener. This approach is elaborated in the rest of this section in addressing specific circumstances that require interaction between practitioner and participant.

The System Crashes, a Serious Bug is Encountered, or the Prototype is Incomplete: From a speech communication perspective, the first concern when a serious system deficiency is encountered should be to maintain the role of the user as work domain expert and the role of the interface as test subject. This can be complicated somewhat by the usability practitioner necessarily slipping into the more technical role of troubleshooter (and possibly the reassuring role of apologetic host). Nevertheless, the participant needs

to be reassured that he or she did not cause the problem, and that the trouble lies with the product being tested. If the problem was heretofore undiscovered, this may even be an opportunity to emphasize the participant's role as a contributor. These assurances should not be overdone and may vary depending on participant demeanor, but they may be necessary to maintain the primary roles set up at the beginning of the session. Without them, some participants are likely to feel self-conscious or responsible for "wasting everybody's time" or "breaking the product."

Second, the usability practitioner has to decide how to get the user back to a point where he or she can continue. If fixing the problem will take more than a few steps, or if the steps needed to troubleshoot the problem will expose unexplored areas of the product, it is probably best to have the participant take a break, stand up for a stretch, or do something else that takes him or her away from the screen long enough for the practitioner to make the required changes.

Finally, the practitioner must indicate where the participant should begin, as mildly as possible to downplay the usability practitioner's control of the situation (for example, *Can we try doing thus and such at this point instead?*). As soon as possible, the practitioner slips back into the listener/learner role by acknowledging the participant's speakership as discussed earlier.

If managed properly, the entire sequence can be viewed not as a switch from dialog to monolog to dialog, but rather as a series of (hopefully) smooth transitions between the different speech genres required by the changing roles—movements from one kind of dialog to another, and back again.

The Participant Avoids Vital Functionality: It is not uncommon for tasks to be "over before they begin;" participants may think they are finished prematurely, get "stuck" before completing the task, or find a creative way to accomplish the task. In some cases this may not be problematic and may even answer the practitioner's research question. But if an important part of the task remains unfinished, critical parts of the system may remain untested. If human problem-solving was the focus, these situations would not present any difficulty; these cases would simply represent a failure of some degree or an interesting, unanticipated solution. But for usability practitioners charged with evaluating an interface, the situation is quite different; if they do not collect data on key system components, they have in effect not done their job.

But how should the usability practitioner restart the participant? A speech communication approach can suggest how to shift roles slightly to accommodate different situations:

- **The participant thinks the task is completed, but it is not.** The usability practitioner should take care that the participant does not feel embarrassed or foolish so that the participant can continue to play the role of work domain expert instead of "subject." If it is simply a case of not having understood the instructions, having the participant reread the task may suffice. We have found it beneficial to conclude each task with a statement that begins with *You'll be done when*.... This helps avoid misunderstanding and (if formatted distinctly) provides a useful review in situations where participants think they are done prematurely. If the participant mistakenly thinks the task is complete, the situation is trickier. Any intervention clearly disturbs the participant's "normal" task flow, but nonintervention here may leave critical functionality untested. Further research should be done to determine the best approach in such cases.

- **The participant sidesteps the functionality of interest.** Participants may find a creative, unanticipated approach to a task, leaving an important part of the product unexplored. Pretesting task lists, of course, is an important part of preventing such situations, but even thoroughly piloted task lists can be thwarted by participant creativity or system deficiency. Such detours are not exactly "mistakes" or "misunderstandings," so it may be relatively easy to maintain the appropriate roles of work domain expert and learner. For example, the practitioner could say something like, *That's a perfectly reasonable way to approach this task; you've helped me understand one way to do this. Could you also [approach it some other way]?* As in the case of participants stopping prematurely, further research should be done into how such prompts should be phrased to minimize the effect on how the participant continues with the task.

- **The participant is stuck.** This is situation is difficult because the participant is likely frustrated or self-conscious; he or she may feel stupid or inadequate, despite reassurances that the system is being tested. The participant's inability to complete the task may be all the data the practitioner needs; the practitioner can simply acknowledge that the participant has helped identify and diagnose a usability problem. But if later tasks rely on the participant continuing, or if the participant has not yet reached the key portion

of the task, how can they be encouraged to continue without influencing later task performance too greatly?

In each case, the question of whether to "stand pat" or to encourage the participant to continue depends on what is more important—maintaining normal task flow or probing the area that would otherwise be missed. Regardless, maintaining the appropriate roles can be tricky, particularly when the participant is stuck or thinks he or she is done prematurely. The usability practitioner should re-affirm the participant's contribution, place any blame on the interface (the object of study), and maintain the researcher's role as learner by emphasizing on "how valuable this has been," "the amount we're learning," and so on.

Other Contingencies Requiring Communication: Many other situations may also require more direct communication with participants. Handling these in any detail is beyond the scope of this paper, but we can suggest a few in which speech communication seems likely to shed a good deal of light:

- **The participant asks a question answered in the task list.** This is faulty role-playing, not faulty task performance; the participant should simply be encouraged to reread the relevant material.
- **The participant asks a question about the task that is not answered in the task list.** Should the usability practitioner ignore it, verbally refuse to answer it, or answer it? If the question is answered, how can it be answered in a way that has the least impact on how the participant proceeds? Ericsson and Simon's model is not well equipped to answer these questions; speech communication may be.
- **The participant asks something that suggests that he or she is approaching the task in an unexpected way.** Turning the question back to the user (for example, *What did you expect to happen?*) may be the most instructive option. Ignoring the question (or to a lesser degree, verbally refusing to answer it) violates communication rules and upsets the balance of roles. Since some kind of response is required, probing the roots of the problem may at least help explain the confusion while keeping the usability practitioner firmly in the learner role. Speech communication research might suggest which responses can achieve this.
- **The participant is unusually "chatty."** No matter how clearly the goals and procedures for a test are explained, some participants will want to strike up a conversation in the middle of a task. Usability practitioners need to turn these diversions back on point without being rude. Specific guidelines are difficult to formulate because different participants may react quite differently, but asking them to restate their current goal may often be effective: *OK—so right now you're working on…?* (OK may actually be the best token in this case, since the usability practitioner is taking a momentarily more directive role.) This approach tries to emphasize the practitioner's interest in the participant's approach to the work, instead of the practitioner's role as test administrator.

Proactively Eliciting Additional Information Heretofore we have discussed two main kinds of contact with participants during test sessions: the constant, background contact of acknowledgment and the reactive contact necessitated by circumstance. The first is directed at achieving the same ends pursued by Ericsson and Simon: a concurrent verbal protocol that is as undisturbed, undirected, and constant as possible. The second is for the most part unavoidable, though there is still room for debate about the type of contact that would be most appropriate in some cases.

But usability practitioners routinely interfere with task performance or verbalization in a third way that neither serves to promote an undisturbed protocol, nor is absolutely required by the contingencies of the testing situation. Instead, they intervene simply because they want to proactively probe for additional diagnostic information (as in Dumas and Redish [14] and Rubin [13]). For example, a participant may make an unclear comment about a key area of the product, or the practitioner may want to elicit specific information at a key point. Intervening in these cases is more controversial than interventions discussed earlier because the Ericsson and Simon model requires silence. From the perspective of formal experimentation, collecting the data from the current participant via additional invention would inescapably tarnish the results; instead, another participant could be recruited and another session run until the research questions has been satisfactorily answered.

Most usability practitioners, however, probably lack the resources to adopt this approach. Probing may contaminate the data to some degree, but if the intrusion can be minimized, many usability practitioners may feel that gathering the additional information is worthwhile, even if it means sacrificing some degree of methodological purity. Some may take the argument a step further, claiming that if "done right," interventions do not contaminate the data of interest. Dumas and Redish, as wll as Rubin, for example, have entire sections devoted to neutral probing without any theoretical reconciliation with the Ericsson and Simon model,

which would prohibit such probes, regardless of their neutrality. Can we then replace the strictures of Simon and Ericsson with a more accommodating theoretical framework?

The remainder of this section suggests how an understanding of the rules of speech communication can better inform practice in these situations than can Ericsson and Simon's model, if usability professionals choose to probe for diagnostic information. However, further research should be done to determine the extent to which such probes elicit reliable information and the extent to which they affect subsequent verbalization and task performance.

Clarifying an Unclear Comment: Participants regularly say things that are difficult to understand. This is to be expected. In fact, participants should be told not to try to make every utterance completely coherent so as to remain focused on the task at hand. But when participants say something unclear at a particularly inopportune moment (perhaps the participant has come to a screen of particular interest or has just unexpectedly failed a crucial task), usability practitioners may want to clarify the comment. Likewise, if the test session or series is drawing to a close, they may need to nail down a particularly elusive issue while they still have time.

Before continuing, we should note that the concern with asking for clarification immediately after an unclear comment is primarily a concern with affecting future performance, not with getting an inaccurate clarification. Careful prompting for clarification after an unclear comment is very similar to asking for an immediate retrospective report. Under Ericsson and Simon's model, chances are good that such reports actually represent what was in short-term memory during task performance. The greater difficulty from Ericsson and Simon's perspective is that the task flow has been interrupted and attention has been redirected; subsequent task performance may therefore be affected. One of the chief considerations, therefore, is which portion of the test is most important; if the current portion is the most important, then perhaps this justifies collecting additional data at the expense of less critical research questions later on. If the main test questions lie further down the path, perhaps restraint is in order.

If practitioners choose not to prompt, another option is to use video-cued retrospective report to "unpack" the participant's thoughts afterwards [17], [29]. Watching the video may provide participants with useful cues as to what they were thinking, but additional research may be useful to ascertain the effectiveness of this technique. The existing research (e.g., [17]) suggests that the video-cued reports elicit "Level 3" data; this may or may not be acceptable, depending on what sort of data the practitioner values. Also, such a procedure takes a great deal of time and creates a sizable interval between the actual task and the retrospective account.

If practitioners do prompt for clarification during the test, they should strive for prompts that do not influence the participant's response. While saying anything at all in these situations takes us beyond the confines of Ericsson and Simon's model, their concern that whatever verbalizations are elicited be as undirected as possible remains. In that spirit, any questions that are posed should be as short and nondirective as possible. Depending on the situation, a single word with the proper inflection may suffice. For example, if the practitioner wants to clarify a particular comment, he or she could repeat part of the comment with an interrogative inflection:

Participant: *That was odd...*

Practitioner: *Odd?*

This compact probe introduces no new content and redirects attention only insofar as participants "unpack" a thought they have already had. It also has the virtue of being extremely short while following natural rules of speech. And even from Ericsson and Simon's perspective, the immediate result should be good, clean data, essentially an immediate retrospective report. The impact on later tasks caused by this redoubled attention is more difficult to ascertain, but at least this approach limits the practitioner to words and concepts already in the participant's short-term memory. Further research would be useful in determining the extent to which such prompts affect later performance.

Probing for More Information: Moving further away from Ericsson and Simon's position, usability practitioners may sometimes want to be more direct in their questioning. Because such direct questions have a greater potential for interfering with subsequent task performance, they should be asked after the task or even in a post-test debriefing. Because of their length, such questions also have the potential to bias participants to respond in one way or another. Dumas and Redish give a great deal of sensible advice for avoiding bias in questions. They cite, for example, "Coleman's Rule" (for M. Coleman "who first brought it to [their] attention"): "A good way to know whether you may be biasing the participant is to **examine how you use adjectives and adverbs in your questions**" ([14, p. 298]; emphasis in the original). For example, *Was that task easy or hard? Why?* may be preferable to *How easy was that task?*

Regardless of the timing or wording of more directive prompts, however,

the fundamental objection of Ericsson and Simon remains: such prompts elicit Level 3 data—inferences, self-explanations of behavior, information retrieved from long-term memory (which is prone to the vagaries of encoding and retrieval processes), and so forth. Many fields routinely accept such information as data; cognitive psychology does not. If a practitioner has a research question that is fundamentally a question of cognition (of inputs, outputs, and their sequence), then Ericsson and Simon's concerns should probably take precedence; participants cannot provide information about their own cognition at a level reliable enough to validate a cognitive model.

But other areas are often of concern to a design team, such as the user's expectations, explanations, prior experience, likes and dislikes, feelings, work practice, design ideas, and so forth. Many published researchers explicitly valued such information over more procedural information [17]–[19], [14]. The field of usability testing needs to come squarely to terms with this issue; if this data is of interest, then we must look further for a theory that can inform its collection because Ericsson and Simon do not provide it.

CONCLUSION

This article has raised a number of issues associated with the widespread claim that the use of thinking aloud in usability practice is derived from and grounded in the work of Ericsson and Simon. In underlining the differences between what practitioners do and what Ericsson and Simon's theory would allow, we hope to initiate vigorous discussion of the appropriateness of both our field's methods and their theoretical grounding. We also hope that our exploration of the alternative theoretical framework of speech communication, and of the methods that are consistent with it, will itself be useful and will also illuminate the discussion of the relationship of theory and method.

However, we must sound a note of warning. A potentially negative impact of proposing speech communication as a theoretical supplement or alternative to Ericsson and Simon may be that some usability practitioners construe this as a "talkative" approach because they believe usability tests should be conducted more like interviews. This is most definitely not our intent. Rather than providing an excuse for less rigor or fewer rules, we hope to have shown that in some cases speech communication can provide different rules for better achieving Ericsson and Simon's ends. In other cases, we hope to have shown that Ericsson and Simon's model does not address some key issues in the context of usability testing, and that some other approach must be adopted if those issues are of any importance.

Most importantly, we hope usability professionals will strive for a practice that is theoretically informed in deed, not merely in word. To be successful, that theory should be sensitive both to the specific context and constraints of usability testing and to the nature of verbalization. We feel that speech communication offers such an approach, either as a supplement to Ericsson and Simon's model or (with much additional research) as a replacement for it. With this further research and professional discourse, the reconciliation of theory and practice in the field of usability testing can become a reality.

REFERENCES

[1] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MTT Press, 1984.
[2] J. Nielsen, *Usability Engineering*. Cambridge, MA: AP Professional, 1993.
[3] G. Deffner, "Verbal protocols as a research tool in human factors," in *Proc. Human Factors Soc. 34th Annu. Meet.*, Orlando, FL, 1990, pp. 1263–1264.
[4] H. Tamler, "How (much) to intervene in a usability testing session," *Common Ground*, vol. 8, no. 3, pp. 11–15, 1998.
[5] H. Desurvire, J. Kondziela, and M. E. Atwood, "What is gained and what is lost when using methods other than empirical testing," in *Proc. HCI'92 Conf.*, York, U.K., 1992, pp. 89–102.
[6] S. Denning, D. Hoiem, M. Simpson, and K. Sullivan, "The value of thinking-aloud protocols in industry: A case study at Microsoft Corporation," in *Proc. Human Factors Soc. 34th Annu. Meet.*, Orlando, FL, 1990, pp. 1285–1289.
[7] M. Pressley and P. Afflerbach, *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. Hillsdale, NJ: Erlbaum, 1995.
[8] R. E. Nisbett and T. D. Wilson, "Telling more than we know: Verbal reports on mental processes," *Psychol. Rev.*, vol. 84, no. 3, pp. 231–241, 1977.

[9] T. A. Suchan and C. A. Brewer, "Qualitative methods for research on mapmaking and map use," *Prof. Geographer*, vol. 52, no. 1, pp. 145–154, 2000.
[10] P. M. Sanderson, "Verbal protocol analysis in three experimental domains using SHAPA," in *Proc. Human Factors Soc. 34th Annu. Meet.*, vol. 2, Orlando, FL, 1990, pp. 1280–1284.
[11] M. T. Boren, "Conducting Verbal Protocols in Usability Testing: Theory and Practice," Ph.D. dissertation, Univ. Washington, Seattle, WA, 1999.
[12] M. McClintock, Procedures for Obtaining Think Aloud Protocols, in Appendix I (included by permission) of M. T. Boren, Conducting Verbal Protocols in Usability Testing: Theory and Practice, Ph.D. dissertation, Univ. Washington, Seattle, WA, 1999.
[13] J. Rubin, *Handbook of Usability Testing*. New York: Wiley, 1994.
[14] J. S. Dumas and J. C. Redish, *A Practical Guide to Usability Testing*. Norwood, NJ: Ablex, 1994.
[15] D. J. Mayhew, *The Usability Engineering Lifecycle: A Practitioner's Handbook for UserInterface Design*. New York: Morgan Kaufman, 1999.
[16] R. B. Wright and S. A. Converse, "Method bias and concurrent verbal protocol in software usability testing," in *Proc. Human Factors Soc. 36th Annu. Meet.*, Monterey, CA, 1992, pp. 1220–1224.
[17] V. Bowers and H. Snyder, "Concurrent versus retrospective verbal protocols for comparing window usability," in *Proc. Human Factors Soc. 34th Annu. Meet.*, Orlando, FL, 1990, pp. 1270–1274.
[18] G. S. Hackman and D. W. Biers, "Team usability testing: Are two heads better than one?," in *Proc. Human Factors Soc. 36th Annu. Meet.*, Monterey, CA, 1992, pp. 1205–1209.
[19] M. Sienot, "Pretesting web sites: A comparison between the think-aloud and the plus-minus method," *J. Business and Tech. Commun.*, vol. 11, pp. 469–483, 1997.
[20] J. Nielsen, "Estimating the number of subjects needed for a thinking aloud test," *Int. J. Human-Computer Studies*, vol. 41, pp. 385–397, 1994.
[21] C. O'Malley, S. Draper, and M. Riley, "Constructive interaction: A method for studying human-computer-human interaction," in *Proc. IFIP Conf. Human-Computer Interaction: Interact'84*, Amsterdam, The Netherlands, 1984, pp. 269–274.
[22] S. A. Douglas, "Conversation analysis and human-computer interaction design," in *The Social and Interactional Dimensions of Human-Computer Interfaces*, P. J. Thomas, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1995, pp. 184–203.
[23] K. Drummond and R. Hopper, "Back channels revisited: Acknowledgment tokens and speakership incipiency," *Res. Language and Social Interaction*, vol. 26, no. 2, pp. 157–177, 1993.
[24] M. M. Bakhtin, "The problem of speech genres," in *Speech Genres & Other Late Essays*, C. Emersen and M. Holmquist, Eds. Austin, TX: Univ. Texas Press, 1986, V. W. McGee, Transl., pp. 60–102.
[25] C. Goodwin, "Between and within: Alternative sequential treatments of continuers and assessments," *Human Studies*, vol. 9, pp. 205–217, 1986.
[26] S. Condon, "The discourse functions of OK," *Semiotica*, vol. 60, no. 1/2, pp. 73–101, 1986.
[27] L. Hirschmann, "Female-male differences in conversational interaction," *Language in Society*, vol. 23, no. 3, pp. 427–442, 1994.
[28] K. Drummond and R. Hopper, "Some uses of yeah," *Res. Language and Social Interaction*, vol. 26, no. 2, pp. 203–212, 1993.
[29] J. Ramey, A. Rowberg, and C. Robinson, "Adaptation of an ethnographic method for investigation of the task domain in diagnostic radiology," in *Field Methods Casebook for Software Design*. New York: Wiley, 1996.

M. Ted Boren completed his Master of Science degree in technical communication at the University of Washington in 1999, emphasizing information design, software user interface design, and usability testing. He is currently employed at Microsoft (Redmond, WA) as usability engineer for geography software. Previously, he earned a bachelor's degree in English from Brigham Young University and worked as

atechnical writer and team lead for the Department of Technical Information Services at Ameritech Library Services (Provo, UT).He is a member of UPA and ACM SIGCHI.

Judith Ramey (A'83) is Professor and Chair, Department of Technical Communication, and Adjunct Professor, industrial engineering, at the University of Washington. She is also Founder and Director of UWTC's Laboratory for Usability Testing and Evaluation. Her research interests are usability research, human–computer interaction, and communication design. In 1989, she was guest editor of a special issue of this journal (December 1989) devoted to usability testing. She is a Fellow of the Society for Technical Communication, a member of the Human Factors and Ergonomics Society, and a member of the Usability Professionals' Association.