JUX Journal of User Experience

Vol. 19, Issue 2, February 2024 pp. 109-114

Response to Commentary on "Talking About Thinking Aloud"

Liam O'Brien
MSc Student
City, University of London
Northampton Square
London EC1V 0HB
United Kingdom
liam.obrien@city.ac.uk

Stephanie Wilson
Professor of Human-Computer Interaction
City, University of London
Northampton Square
London EC1V 0HB
United Kingdom
s.m.wilson@city.ac.uk

We would like to thank the Editors for the opportunity to respond to the interesting points raised by Rolf Molich. We have a great deal of respect for Mr. Molich and his contributions to the field over many years, especially his work on the Comparative Usability Evaluation studies (CUE).

In our paper, "Talking About Thinking Aloud: Perspectives from Interactive Think-Aloud Practitioners," we describe a small-sample qualitative study exploring the attitudes of Interactive Think Aloud (ITA) practitioners. ITA has been extensively researched through lab-based and observational approaches, but our study builds on this work as the first to report in detail on the views of practitioners regarding their use of ITA. We adopted widely accepted best practice approaches for qualitative-depth research and fully acknowledged the limitations of our work.

Here we provide a comprehensive response to Mr. Molich's letter. We address key points raised under our three headings "Research Approach," "Study Presentation," and "Findings and Implications."

Research Approach

Mr. Molich raises two key concerns relating to our research approach. These arise from a misalignment in views about the nature of small-sample qualitative research.

Generalization Value

Mr. Molich questions whether our study has "strong generalization value," citing its small sample. However, we would note that qualitative researchers are not normally concerned with large samples and the statistical generalizability they can offer. The focus is instead on building a detailed and contextual understanding of a particular group, and generalizability is achieved through "transferability," which is the ability for conclusions to be applied by other researchers in other contexts (Polit & Beck, 2010). We appreciate that the small sample of UK practitioners has implications for our findings, and we acknowledged this in the "Limitations" section of the paper. However, we followed best practice for transferability, such as by providing "thick descriptions" of our findings (Barnes et al., 2005), and we believe our results are generalizable in that they are transferable to similar contexts (such as studies involving UX practitioners in western countries).

Many qualitative studies in the usability literature take a similar approach. For example, Mr. Molich's own CUE-10 study (Molich et al., 2020) also involved a small convenience sample of 14 UX practitioners with a particular sample bias, all recruited from UTEST (a private community for usability professionals, many who were Bentley University students and graduates, etc.). This paper was also published by *JUX*, and although the journal could have taken a narrow statistical definition of "generalization value," we think they made the right decision by taking a broader view in both cases.

Intercoder Reliability

We did not use multiple coders or assess Intercoder Reliability (ICR). Although this practice is common in some forms of highly-structured deductive qualitative analysis (such as coding usability problems), ICR is by no means ubiquitous in thematic analysis (O'Connor et al., 2020). We followed the classic Braun and Clarke method (2012), which is an approach for a solo researcher.

Braun and Clarke advise against assessing ICR on the grounds that it is incompatible with the interpretivist agenda of qualitative research (2013). Under this theoretical framework, ICR is seen as an inappropriate attempt to apply positivist ideas to the fundamentally interpretivist practice of qualitative analysis, in which "reflexivity and active personal engagement with the data are resources, not 'noise' to be minimized" (O'Connor et al., 2020). In our study, the coding framework evolved organically through a very broad inductive approach, and ICR may have prevented some of the more interesting findings from surfacing. This is obviously not the place to get into a detailed philosophical debate about the pros and cons of ICR in thematic analysis, but we stand by our decision not to use ICR as a legitimate theoretical choice.

Study Presentation

Mr. Molich also critiqued the way we defined key concepts and presented our study. We are happy to provide an explanation of these choices.

Defining Traditional Think-Aloud and Interactive Think-Aloud

Mr. Molich is dissatisfied with the way we defined Traditional Think-Aloud (TTA). In the introduction of the paper, we defined TTA according to the Ericsson and Simon model (1980), which forbids all interventions (including all listed in Tables 1 and 2). This is the most established TTA protocol, and it is traditionally a benchmark against which other think-aloud variants are compared. As Mr. Molich points out, some of the interventions we identified (such as clarifying the task) may be deemed acceptable by some moderators, but these are not permissible under TTA as we defined it. We could have used a more liberal definition of TTA, but we were interested in the wide spectrum of intervention types beyond the Ericsson and Simon protocol, and we did not want to prematurely judge the acceptability of any particular intervention types.

Mr. Molich is also critical of how we defined Interactive Think-Aloud (ITA). This is a difficult area because, as we mentioned in the paper, there is no established protocol or definition for ITA, and there is some inconsistency in the literature in how ITA is described. Although we did not explicitly set out to define an ITA protocol, we hope that by elucidating aspects of ITA practice, we have contributed toward that effort. In the "Recommendations for Further Work" section of our paper, our first recommendation is that researchers should define an ITA protocol.

Moreover, Mr. Molich seems uncomfortable with our presentation of ITA as a form of usability testing. But ITA studies are generally considered to be a form of usability testing. Boren and Ramey were among the first to identify ITA practice and present the approach as a "divergent" form of usability testing (2000). Others have followed suit in treating ITA as a form of usability testing (McDonald et al., 2016) even when they have been highly critical of ITA (Alhadreti & Mayhew, 2017). Our participants also referred to their ITA practice as "usability testing" most of the time.

Study Scope

Mr. Molich seems interested in lines of inquiry which were simply outside the scope of our study. For example, he asks for our participants' views on "the key differences between usability testing, semi-structured interviews, and design critiques." This was not what we set out to explore, and we did not collect enough data on these points to report on them.

Account of Method

Mr. Molich says we should have included the Discussion Guide in an appendix. This is not necessarily standard practice for *JUX*, and we were not asked to provide the Discussion Guide. Instead, we gave an account of the study design and interview format in the "Methods" section. Although, as Mr. Molich noted, we did forget to include the interview length (interviews were 1-hour long).

Conclusions and Tips for Usability Practitioners

Mr. Molich questions the provenance and presentation of our conclusion. Our conclusions represent the dataset (interview transcripts) as a whole. We did not report on how many participants agreed with our conclusions because we arrived at these conclusions after the data collection stage, and this kind of quantification would not be appropriate for a small-sample, qualitative, thematic analysis-based project anyway. One of our recommendations for further work was to conduct a survey, which would be a better way of quantifying these attitudes.

Mr. Molich seems to also be unsure about the provenance of our tips for usability practitioners. The *JUX* "Guidelines for Authors" state that, "The tips should come from the article you are preparing, typically from the method, findings, or conclusions." Ours come from the findings, and at the start of the "Tips" section we provided this clarification: "The following tips are based on our findings…"

Findings and Implications

Mr. Molich raises some concerns regarding our findings and their implications for usability testing as a method. We are glad to take this opportunity to further our case for usability researchers to think differently about ITA.

ITA and Opinion Elicitation

Mr. Molich's presentation of ITA as "an opinion-based method" is a mischaracterization. Our participants were wary about collecting opinions from users, especially redesign proposals (see "Results," "Theme 5"). Most intervention types we identified as intended to improve data usefulness (Table 2) are not aimed at gathering opinions. For example, participants said they intervene when the user is acting without a clear purpose in order to understand what the user is trying to do. This is not opinion elicitation, and the user's immediate response is likely to provide valid data on user cognition under the Ericsson and Simon model (1980).

In contrast to his view of ITA as an "opinion-based method," Mr. Molich describes results from TTA as "free from opinions," but this is also a mischaracterization. TTA has, in fact, been found to contain reflective data like opinions, recommendations, and explanations (Zhao & McDonald, 2010), so it seems users share their opinions whether or not moderators ask for them.

Mr. Molich claims that opinions are "dangerous," and we agree that user opinions can be misused. But attitudinal or opinion-based data can also be useful, have high validity, and complement other usability data (Følstad, 2017). Both TTA and ITA involve opinion data, and although ITA does offer the flexibility to ask the user’s opinion when required, opinion-elicitation was not a major feature of ITA for our participants.

Acceptability of ITA

Mr. Molich says that we do not address whether ITA is a “good way of doing things.” We disagree, as we do address this in our conclusion, although our advice is perhaps more nuanced than Mr. Molich would like: “ITA is not a replacement for TTA, but there are some circumstances (methodological, organizational, and practical) in which ITA may be a better choice for practitioners than the traditional approach.” (See the "Conclusion” section for a more detailed discussion.) While we agree that some ITA practitioners may not realize that their interventions can cause problems, our participants were generally aware of reactivity and took steps to mitigate it, and this finding was consistent with other studies (McDonald et al., 2012) (see “Theme 8—Managing Reactivity”). We would also add that some TTA practitioners could be overlooking the advantages of ITA, particularly in situations in which ITA may be especially beneficial, such as early-stage studies, test of highly technical or domain-specific systems, and studies focused on building a deeper or more contextual understanding of usability and usefulness.

ITA as a Hybrid Method

Although ITA is certainly more flexible than the traditional approach, we would not equate ITA with "talking to strangers in an elevator about problems they have with some system" as Mr. Molich does. ITA has much in common with TTA, such as the focus on the test system, the use of tasks, and the collection of behavioral data. But ITA does differ from TTA in that it incorporates a range of intervention types and involves the additional intentional collection of non-behavioral data. In our “Conclusion” section, we suggest that ITA could be positioned somewhere between TTA and semi-structured interviews.

Mr. Molich says that, under ITA, usability testing “loses its distinctive character.” However, method distinctiveness is not the goal. ITA is valued by practitioners as a hybrid method precisely because it allows them to collect different kinds of data within a single study. One could draw parallels with contextual inquiry, another hybrid method that combines techniques (interview, ethnography, and participatory design) to offer greater flexibility and build a deeper, more contextual understanding of usability. In fact, both ITA and contextual inquiry involve the addition of interview-like elements to a more traditional, mostly observational study.

Legitimizing ITA as User Experience Testing

We are keen to move the debate forward and have appreciated this opportunity to respond to Mr. Molich’s comments. We agree that terminology in this space could be a source of confusion and a barrier to progress. Given the differences between ITA and TTA, we propose something of a rebranding of ITA that moves away from the term *usability testing*. A new name will help foster an appreciation for ITA as a method in its own right with different goals, a different theoretical perspective, and involving the collection of different kinds of data. We propose *user experience testing*, as this maintains an association with usability testing but alludes to the often-broader objectives of ITA.

We hope this response will encourage usability researchers who criticize ITA as invalid usability testing to instead focus on building a better understanding of the approach and providing constructive advice. Helpful advice should be focused on how to maximize the advantages of *user experience testing* (improving data usefulness and dealing with practical challenges, for instance) while minimizing its disadvantages (mitigating reactivity and using data with variable validity).

References

Barnes, J., Conrad, K., Demont-Heinrich, C., Graziano, M., Kowalski, D., Neufeld, J., Zamora, J., & Palmquist, M. (2005). *Generalizability and transferability*. Writing@CSU. Colorado State University. https://writing.colostate.edu/guides/guide.cfm?guideid=65

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

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology* (Vol. 2, *Research designs: Quantitative, qualitative, neuropsychological, and biological*, pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. Sage Publications.

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

Følstad, A. (2017). Users design feedback in usability evaluation: A literature review. *Human-Centric Computing and Information Sciences*, *7*(1), 19. https://doi.org/10.1186/s13673-017-0100-y

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

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

Molich, R., Wilson, C., Barnum, C. M., Cooley, D., Krug, S., LaRoche, C., Martin, B. A., Patrowicz, J., & Traynor, B. (2020). How professionals moderate usability tests. *Journal of Usability Studies*, *15*(4), 184–209.

O’Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*. https://doi.org/10.1177/1609406919899220

Polit, D.F., & Beck, C.T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, *47*(11), 1451–1458. https://doi.org/10.1016/j.ijnurstu.2010.06.004

Zhao, T., & McDonald, S. (2010, October). Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, *New York*, *USA*, 581–590. Association for Computing Machinery (NordiCHI ’10). https://doi.org/10.1145/1868914.1868979

About the Authors

Liam O’Brien
O’Brien is a Senior User Experience Researcher at Independent Commodity Intelligence Services (ICIS). He has worked across the public and private sectors practicing and training others in a range of usability testing approaches.

Stephanie Wilson
Wilson is Professor of Human-Computer Interaction at City, University of London. She has over 20 years of experience in teaching, researching, and practicing UX design and evaluation. She leads on the MSc in Human-Computer Interaction Design.