

Messy Methods and Moderation in User Tests

Lene Nielsen

Section for Digitalization, Democracy and Governance
IT University Copenhagen
Copenhagen, Denmark
lene@itu.dk

Bernard J. Jansen

Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
bjansen@hbku.edu.qa

Sara Marie Ertner

Section for Digitalization, Democracy and Governance
IT University Copenhagen
Copenhagen, Denmark
saramarie@itu.dk

Joni Salminen

University of Vaasa
Vaasa, Finland
jonisalm@uvasa.fi

Abstract

In this paper, we argue that practices of moderating usability and user trials are never as neutral, formal or procedural as most of the methods literature suggests. Borrowing from science and technology studies scholar John Law's notion of 'messy methods', we propose that moderating usability trials is always a 'messy affair' that involves unruly assemblages and requires the more or less ad hoc and in situ social and material management of rowdy relationships between different entities, such as human bodies and a vast amount of material objects. Based on in-depth and ethnographically oriented interviews with usability trial moderators, we analyse the skilled social, creative and embodied practices of moderators managing the realities of user test situations that are more complex than the literature renders visible. Our study shows that moderators employ a rich repertoire of techniques to master the situation. This points to the need for a revitalisation of the user test method to acknowledge the relationship between test person and moderator to obtain richer material. The paper contributes methodologically to human-computer interaction (HCI) and user test studies by developing the notion of messy moderation and messy methods in HCI. This holds significant implications for our view of user study methods applied and developed in HCI, as it invites sensitivity and reflexivity into how our methods contribute to creating the realities they aim to study.

CCS Concepts

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI design and evaluation methods**;

Keywords

User studies, UX, Moderation, Messy methods

ACM Reference Format:

Lene Nielsen, Sara Marie Ertner, Bernard J. Jansen, and Joni Salminen. 2026. Messy Methods and Moderation in User Tests. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3772363.3798368>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI EA '26, Barcelona, Spain*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2281-3/2026/04
<https://doi.org/10.1145/3772363.3798368>

'26), April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3772363.3798368>

1 Introduction

User experience (UX) design can be viewed as a scientific practice that seeks to supply clear and definite knowledge about user perspectives and preferences, thus ordering the world into categories of users with specific needs and static preferences. From this perspective, usability testing is one method for retrieving such knowledge. The science and technology studies scholar, John Law, has famously argued that the world is largely messy and research methods tend to leave out a lot of this mess [14]. Mess refers to the nature of reality as being vague, unspecific, slippery, emotional and unordered. Hence, our methods often leave out a lot of complexity because 'the world' rarely behaves in ways that fit neatly into our scientific instruments or conforms to the ways in which we investigate it. Inspired by his notion of messy methods, we ask "what mess is left out in user test methods, and what are the implications of this for methodological development and knowledge enhancement in technology production?"

The case focused on here employs a usability test protocol on an artificial intelligence (AI) chatbot (specifically, CIPHERBOT, a state-of-the-art AI educational assistant for professors and students) for learning [11] that allows students to ask questions and get answers from class materials. The test protocol has a five-step procedure: (1) a pre-survey; (2) a task with a random assignment of creating a social media post based on either visual (text) or audio output and information; (3) a post-test survey; (4) the same task performed with a new setup; and (5) a survey of the whole learning experience. During and after completing the test sessions, we interviewed the eight moderators on their experiences with moderating.

The data collection partly followed the Thayer and Dugan [18] model, involving a pre-experience interview, a post-task questionnaire, performance data in the form of a think-aloud protocol and evaluator observations, immediate reflections, a post-test questionnaire and a post-experience survey.

In the following, we use the concept of mess as a lens for understanding and point to how we can become better at making sense of user study situations. We focus on the research question: *How are subjects, moderators, the moderation situation and process performed in practice?* This is juxtaposed with the theoretical literature on usability testing to reveal contrasts and differences.

2 Related literature

2.1 User tests and moderation

One of the most prominent methods used in UX design is testing of both the usability and the user experience. The purpose of user-testing is to evaluate the design of information technology (IT) systems. Viewing the literature on user tests through the lens of mess, it is evident that the literature speaks of how to avoid the mess through the methods and views the world as ordered and systematic. Based on this literature, we describe how to perform moderation, the relationship between the moderator and the test subject, and the moderation situation and process performed in practice. In the following, we refer to the one in charge of the test protocol as the moderator, and we refer to the literature on concurrent usability tests with a relaxed protocol and user tests that accepts minimal interaction between the test subject and the moderator [3, 9].

User-testing employs resources, such as test tasks, users and measuring instruments, to expose usability problems in their use. The reasoning behind user-testing is that, based on a limited number of test participants, it is possible to generalise the test to accommodate different users, tasks and contexts [2]. In general, the literature on moderation covers techniques for probing without leading participants, how to remain neutral during sessions, when (and how) to intervene, and how to manage difficult participants [3, 4, 12, 15, 16]. Both the test questions and moderator verbalisations can influence the test participants and impact a study's results [10].

When it comes to the interactions between the moderator and the test subject, the perspective in the literature is mainly directed towards the moderator and their moderation. When it comes to the participant, Boren and Ramey [1] suggested viewing the test subject as a domain expert and the moderator as an active listener, and usability sessions as interactive and conversational. Thus, some moderator interventions are necessary and natural. The role of the moderator is to follow the script and get the test moving forwards in an orderly manner. The moderator must help the participants stay on track [9] but otherwise should take a neutral position towards the test participant to avoid bias and influence.

The literature describes how moderators need to be in control of themselves and the situation. Of the 12 characteristics of a good moderator mentioned by Molich et al. [15], several are concerned with being able to control oneself. The moderator should respect the test participants as experts but remain in charge [2]. The moderators also need to be watchful of themselves to remain sharp [2]. Dumas and Loring [3] instructed the moderator to be professional and genuine.

Because it can be difficult to understand what goes on in a test situation, and from observations and verbalisations pertaining to the test, the moderator needs a rich repertoire of constructs for thinking about how users experience products [9]. Apart from understanding the UX, the moderator needs to be able to observe the situation, including any non-verbal cues [1] expressed through body movements and postures that might express surprise, interest or an encounter with something unpleasant [5].

In summary, the description of usability and user tests is of a procedural, instrumental systematic methodology. To extract data, the moderator's role is to document and, from that documentation,

to extract results that are valid and applicable across different users and contexts. User-testing is methodological and moderation is a neutral role [8]. There is common agreement that if the tasks in the tests are performed under the same conditions, performance comparisons are legitimate and can be generalised to other users, tasks and contexts [2, 17].

2.2 The concept of messy methods

The world is largely messy is the starting point in science and technology studies scholar John Law's book *After Method* [13]. This book offers a critical inquiry into the common-sense realism inherent in most natural and social science methods. It is Law's main argument that most social and natural science methods produce a highly prescriptive version of the nature of the real, which charges that reality cannot be a mess. This stance is characterised by ideas about reality existing independently and prior to our knowledge of it. Reality is seen to exist 'out there' in definite and singular terms. These ideas together work to 'purify' the empirical world or distort reality into clarity, as Law [13] puts it, which makes contemporary methods extremely poor at understanding the messy nature of the world.

Law [13] proposed methodological reflexivity, which engages explicitly with the messy relationships between reality and our methods. In human-computer interaction (HCI), similar claims have been made by, for example, Frauenberger [6] who famously argued that HCI is entering into a fourth wave characterised as 'entanglement HCI'. Entanglement HCI rejects positivist understandings of an external world and theorises humans and objects as relational and co-evolving, thus producing realities through intra-actions. Our initial curiosity was awakened by the seeming disparity between the formal description of user trial moderation (see, e.g., [4, 8, 9]) and how moderators talk about and carry out moderation in practice.

To echo Law's question, "if much of the world is vague, diffuse or unspecific, slippery, emotional and does not really have a pattern at all" [13], p. 2], where does this leave user testing and moderation? If usability testing, users and moderators do not 'behave', much less 'exist', as the positivist underpinnings of the usability methodology prescribe, then what is being othered, excluded or deleted in our understandings of, and approaches to, performing moderation of user tests? And what are the implications of this othering? Inspired by Law's [13] notion of mess as both a methodological and ontological stance and call to reflexivity, we analysed how test subjects, moderators and user test situations are performed in practice, and how this is disconnected from the way they are described and categorised in the methods literature.

3 Method

During a four-day, large-scale user-testing of an AI chatbot for educational purposes, eight moderators performed 62 test sessions, lasting around one hour each, and one researcher interviewed each of the eight moderators. The semi-structured interviews took place in situ, either at the end of the day, after a series of moderations, or in the morning, looking back at the previous moderating sessions. Each interview lasted between 15 and 45 minutes. The study involved eight moderators, two female and six male, aged 24–64 years,

and one female interviewer, 67 years of age. Three interviewees were very experienced with more than hundred sessions, two had medium experience and three moderators were trying it for the first time. All interviews were recorded and transcribed verbatim.

The analysis of the interviews was carried out by systematically coding all collected materials. The codes were analysed according to qualitative grounded theory research [7], with the first coding resulting in 47 first-order concepts. Following the initial analysis, the aggregated dimensions were analysed using messy methods as a lens, where we looked at the entangled relationships between the subjects, material objects, and user test processes in practice. The lens produced three aggregated dimensions: embodied and social moderation, material and temporal arrangements, and managing progress and process.

4 Analysis

4.1 Embodied and social moderation

Most user-testing literature represents both the moderator and the test subject as rational individuals who form thoughts relatively independently and through disembodied and socially detached modes of being in the world. By contrast, our respondents talked about bodies, energies and affects as something profoundly important to the test situation and as entwined in various ways.

As a moderator, you are required to focus your awareness on the test subject during the entire session. This active and constant awareness requires the consumption of mental energy and, for the moderators, an adjustment of their own energy to the energy of the person doing the test. One moderator coined this as matching the energy of the other.

“I try to match the energy (...) matching the energy doesn’t mean that if they say a lot, you have to say a lot, but you have to listen actively (...) They are so emotionally intelligent that if you say ‘mm’ and there’s a slight difference in the tone or a 20-millisecond difference in when you say it, in this high-frequency communication with the person who is, like, using their cognitive capacity completely, they will notice that: ‘Okay, that guy didn’t completely follow my trail’. So that’s mentally taxing.” (M6)

Matching energies is a form of mental and energetic intra-action, where the moderator places themselves in the mind of the test-subject to ‘follow their trail’- to understand everything they say and be able to respond in a way that evinces active engagement. Matching energies points to the ways that moderators must both mentalise with the test subject and internalise aspects of the test subject, thereby making adjustments to their own mode of being present. In other words, they adapt parts of themselves to each individual test situation and test subject. This form of intra-action goes far beyond the instrumental activities of prompting and documenting so often referred to as part of the tasks of the moderator. According to our respondents, bodies play a prevalent role in moderation practices.

“I have to monitor their facial expression also. I have to look at their facial expression, their typing behaviour... I notice, like, if they’re really concentrat[ing] or not on the task. Or are they anxious, uncomfortable? So, if they are uncomfortable, I make them comfortable.” (M8)

In this quote, the moderator recounted how to pay attention to embodied expressions that reveal a lot about the test subject’s condition and engagement with the task. This requires looking at facial expressions, the hands’ interactions with the keyboard, and the embodied expression of affective states, such as discomfort, concentration or even anxiety. Another moderator explained how observations of yawning test participants and legs that move restlessly become signifiers that the person has lost interest, which is important data concerning the person’s experience of the interface they are interacting with.

The moderators are aware that the participants have an impact on their mood and, implicitly, on their behaviour and attitude towards the participant, even if they express that they try to maintain a neutral position. Other situations include enduring long stretches of silence (M6), boredom, appearing unthreatening to a sensitive test subject (M6), or downplaying an immediate expression of affect when they sense a positive feeling from a test subject (M3). Being neutral is not a default attitude for the moderators. It involves active bodywork to stay calm and appear untouched. Moderators are aware of their own bodily expressions of moods and affective states, using their bodies to align with those of the participants to convey interest and attention through embodied expressions. At the same time, the bodies of the test subjects are also important means of non-verbal communication and the interpretation of moods, behaviours and attitudes throughout the test.

Despite the absence of attention to social and interpersonal dimensions in the usability and user test literature, a usability test is a profound social activity. This is evident in our empirical material, where our respondents talked a lot about how they perceived the test subjects as ‘good’ or not. The social processes of valuation and judgement, the formation of antipathy/sympathy, and categorisation played a central role in the moderators’ perceptions of the test subjects, of their moderation and of user-testing. There is a notion of ‘a good participant’ among moderators. The moderators also have clear ideas of participants that they perceive as being difficult, those that are embarrassed by the situation, do not make an effort, or are very talkative and forget the task (M3 and M4).

The moderators’ perceptions of what kind of test subject they are facing shapes their own mood and affective attitude towards the test situation, as the next quote illustrates.

“I consider myself, in that sense, a feelings person. It affects me. It might put me a little down if I see that the other one is one of these cases (...) It’s going to take a lot of effort from me, and it’s going to take a lot of effort from the participant to complete the task.” (M3)

While moderators strive to be objective, as the methodology prescribes, moderators are human beings and tend to react towards other people. Recognising the social dynamics and their embodied expressions can be a way of dealing more professionally and consciously to resistance and barriers in the social situation. Acknowledging the social and bodily dimensions of these encounters opens us to viewing user-testing and moderation as a profoundly social affair where social dynamics and intra-actions, and embodied actions affect and shape the user test and moderation practice and affect the test situation and the results.

4.2 Material and temporal arrangements

The ideal user test is described in terms that omit practical and material restraints. In practice, being a moderator oftentimes extends beyond the moderation of the test procedure to also engage with specificities, such as room reservations, time consumption and physical objects that are present in or absent from the situation. Thus, engaging in material and creative practices is an important aspect of moderating user tests. Moderators make use of material objects, such as monitors and notebooks, and engage in material practices, such as face-reading, note-taking and note transcription. Such material conditions and practices are rarely, or never, described in usability and user test methods, but appear to bear significant importance in the moderating task.

Monitors are very present in the test situation, but sometimes they serve unexpected functions. Some moderators explain how monitors are used as ‘mirrors’ that reflect the facial expressions of the test subjects that are otherwise not necessarily easily available, especially because the participants’ faces are hidden, facing towards a screen and immersed in the task at hand. In these cases, the moderators’ ability to creatively use the materialities in the room, such as a computer monitor, comes in handy.

Note-taking is a profoundly material task, which can be quite demanding. It requires the operation of two simultaneous work processes of overseeing the test procedure and capturing activities in notes, and this is neither procedural nor seamless. The moderators talked about the intricacies of taking notes, the difficulties and importance of taking ‘good’ notes, and the material practice of hiding and transcribing notes.

“I found note-taking difficult today. I generally am a note-taker. And then yesterday, I had to spend an hour extra after everyone slept to transcribe those notes to a Word document. So, I thought I’ll give it a try today to write the notes.” (M1)

User test set-ups are often very complex materially and technically. In our case, the test consisted of technologies for tracking eye and mouse behaviours and making voice recordings. The test was followed by an onscreen survey. The moderator needs to be able to set up the equipment, manage the test flow and, preferably, also take notes.

The test set-up includes having several participants meet in specific time slots to perform the test within a specific amount of time. This means that the moderators must both manage the user trial in action while also paying attention to time schedules, room availabilities and the next participants in line.

“That’s the first thing I do, I see the other room’s schedule as well. Do they have someone coming in? So, if they have a participant coming and I have a participant coming, then probably, in that case, I try to rush things. Somehow tell them, okay, we have limited time or something, but if I don’t have a participant, then I just let them go through, like, a natural process.” (M4)

Not only does the moderator need to have an overview of what is going on in the user test situation, they also need to sense what is going on outside the test facility and be observant of both their own and their co-moderators’ schedules. These orientations are not external to the test because room availability may cause a test to be rushed to avoid delays and queues. That moderators may need to rush things to make everything run smoothly implies

that moderators need to be able to skilfully manage time, spatial capacities and restrictions.

4.3 Managing progress and process

Moderation is usually described as a non-intrusive activity, where moderators abstain from influencing the process as far as possible, even though user trials are not merely moderated as if they progressed automatically. They must be managed and steered through invisible strategies to secure progress and processes that stay on track. This may be achieved through encouragement, reminders, nudging and examples. Doing this requires skill and experience, and over time, experienced moderators may develop a ‘spider-sense’ of how and when this progress management works well. For less experienced moderators, it may elicit doubt and insecurity.

In practice, however, moderators engage in a wealth of progress management work to keep test participants on track, such as reassuring, helping or nudging them to keep the test moving forwards, or to speed it up when time is running out. Other times, they may read questions aloud to encourage responses to keep the session moving or use exact sentences or keywords from survey questions to prompt discussion. In these situations, moderators embark on progress management work, which reminds the test participant of the end goal and seeks to simplify the task to encourage them to keep going. Progress management work may also involve bringing test subjects who are too talkative back to the topic.

“In the first five minutes... he has talked about five different topics. So, I realised this is going to lead in a very different way (...) I’m going to try to keep on bringing him back to the same topic.” (M1)

Talking aloud while solving a problem does not come easy to all participants, and in such situations, the moderator needs to find a balance between gently guiding and pushing the participant too much. To keep the test participant talking, the moderator can apply different strategies, such as the example from an earlier test (M1) or asking about their thoughts on a particular response.

5 Conclusion

We set out to investigate how test subjects, moderators, the moderation situation, and process are performed in practice. We found that moderators employ a rich repertoire of techniques to master user test situations that appear more complex and ‘messy’ in practice than prescribed in the methods literature. Our findings illustrate this through analytical themes that show 1) how moderation emerges as a profoundly social and embodied practice, where energies, affects and human bodies intra-act and collectively shape the test situation and knowledge; 2) a process affected by material and temporal conditions that shape and restrain the test situation and its pace; 3) and a process that requires the active engagement of moderators to perform skilled progress management work. We argue for a view on user testing as a collective situation shaped through social, embodied, material and temporal intra-actions, and moderation as an active, engaged process of managing relations and co-creating knowledge.

This points to a need to revisit the usability and user test methods to, among other things, acknowledge the social and embodied relationship between the test subject and the moderator. Furthermore, this calls for intervention in the common orientation

towards (and fear of) ‘bias’, and the conjoint theory of test subjects’ perspectives as existing in ‘pure’ form and independently of test situations and their moderation. This means that paying attention to bodily movements, facial expressions and energies is important. To leave the objective scientific paradigm for a collaborative perspective, where materials and subjects act together, we suggest reframing the moderator/evaluator as a facilitator, a co-creator of knowledge with the participant, inspired by co-design.

References

- [1] Boren, Ted and Ramey, Judith. 2000. Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication* 43, 3: 261–278.
- [2] Gilbert Cockton. 2014. Usability Evaluation. *IxDF - Interaction Design Foundation*. Retrieved from <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/usability-evaluation>
- [3] Joseph S. Dumas and Beth A. Loring. 2008. *Moderating usability tests: Principles & practices for interacting*. Morgan Kaufmann.
- [4] Joseph S. Dumas and Janice Redish. 1999. *A practical guide to usability testing*. Intellect.
- [5] Joseph S. Dumas and Marilyn C. Salzman. 2006. Usability Assessment Methods. *Reviews of Human Factors and Ergonomics* 2, 1: 109–140. <https://doi.org/10.1177/1557234X0600200105>
- [6] Christopher Frauenberger. 2020. Entanglement HCI The Next Wave? *ACM Transactions on Computer-Human Interaction* 27, 1: 1–27. <https://doi.org/10.1145/3364998>
- [7] Dennis A. Gioia and Aimee L. Hamilton. 2013. Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods* 16, 1: 15–31. <https://doi.org/10.1177/1094428112452151>
- [8] Elizabeth Goodman, Mike Kuniavsky, and Andrea Moed. 2013. Observing the User Experience: A Practitioner’s Guide to User Research (Second Edition). *IEEE Transactions on Professional Communication*. <https://doi.org/10.1109/TPC.2013.2274110>
- [9] Morten Hertzum. 2020. *Usability Testing: A Practitioner’s Guide to Evaluating the User Experience*. <https://doi.org/10.2200/S00987ED1V01Y202001HCI045>
- [10] Morten Hertzum and Kristina Bonde Kristoffersen. 2018. What do usability test moderators say?: “mm hm”, “uh-huh”, and beyond. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, 364–375. <https://doi.org/10.1145/3240167.3240181>
- [11] Soon-Gyo Jung, Johanne Medina, Kholoud Aldous, Jinan Azem, Joni Salminen, and Bernard J. Jansen. 2025. CIPHERBOT: A Learning Platform for AI-Augmented Education. In *Proceedings of the Augmented Humans International Conference 2025*, 478–481. <https://doi.org/10.1145/3745900.3746106>
- [12] Emiel Kraemer and Nicole Ummelen. 2004. Thinking About Thinking Aloud: A Comparison of Two Verbal Protocols for Usability Testing. *IEEE Transactions on Professional Communication* 47, 2: 105–117. <https://doi.org/10.1109/TPC.2004.828205>
- [13] John Law. 2004. *After Method*. Routledge. <https://doi.org/10.4324/9780203481141>
- [14] John Law. 2007. Making a mess with method. In *The Sage handbook of social science methodology*, 595–606.
- [15] Rolf Molich, Chauncey Wilson, Carol M Barnum, Danielle Cooley, Steve Krug, Chris LaRoche, Beth A Martin, Jonathan Patrowicz, and Brian Traynor. 2020. How Professionals Moderate Usability Tests. *Journal of Usability Studies* 15, 4.
- [16] Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2381–2390. <https://doi.org/10.1145/1753326.1753685>
- [17] Pekka Reijonen and Kimmo Tarkkanen. 2015. Artifacts, Tools and Generalizing Usability Test Results. In *Nordic Contributions in IS Research*, Harri Oinas-Kukkonen, Netta Iivari, Kari Kuutti, Anssi Öörni and Mikko Rajanen (eds.). Springer International Publishing, Cham, 121–134. https://doi.org/10.1007/978-3-319-21783-3_9
- [18] Alexander Thayer and Therese E. Dugan. 2009. Achieving design enlightenment: Defining a new user experience measurement framework. In *2009 IEEE International Professional Communication Conference*, 1–10. <https://doi.org/10.1109/IPCC.2009.5208681>