Journal of Usability Studies
Vol. 15, Issue 4, August 2020, pp. 184–209

How Professionals Moderate Usability Tests

Rolf Molich
DialogDesign
Skovkrogen 3
3660 Stenlose, Denmark
molich@dialogdesign.dk

Chauncey Wilson
WilDesign Consulting
Wayland, MA, USA
chauncey.wilson@gmail.com

Carol M. Barnum
UX Firm
Atlanta, GA
carolbarnum@uxfirm.com

Danielle Cooley
DGCooley & Co.
9730 E. Watson Rd. #100
St. Louis, MO 63126
danielle@dgcooley.com

Steve Krug
Advanced Common Sense
Chestnut Hill, MA
skrug@sensible.com

Chris LaRoche
P.O. Box 398094
Cambridge, MA 02139 USA
c.laroche@northeastern.edu

Beth A. Martin
Maryland Institute College of Art (MICA)
1300 W Mount Royal Ave.
Baltimore, MD 21217
bmartin01@mica.edu

Jonathan Patrowicz
Jonnypats.com
Maynard, MA 01754
Jonmpat@gmail.com

Brian Traynor
Information Design
Mount Royal University
Calgary, Alberta
btraynor@mtroyal.ca

Abstract

This paper reports how 15 experienced usability professionals and one team of two graduate students moderated usability tests. The purpose of the study is to investigate the approaches to moderation used by experienced professionals. Based on this work, we present our analysis of some of the characteristics that distinguish good from poor moderation.

In this study, each moderator independently moderated three think-aloud usability test sessions of Ryanair.com, the website of a low-fare European airline. All moderators used the same six usability test tasks. The test sessions were video recorded so that both the participant and moderator were visible.

Key observations were identified by asking other study participants to review a random video from each moderator. Each video was reviewed by five to seven study participants. With this approach, the data (not a single person, author, organizer, or moderator) determines what the key observations are.

This study documents a wide difference in moderation approaches. In this paper, we discuss several common issues in usability test moderation, including time management, prompts and interventions, moderator interaction styles, and the provision of positive participant feedback during sessions.

This study is the tenth in a series of Comparative Usability Evaluation studies.

Keywords

moderation, usability test, think-aloud, facilitation, comparative usability evaluation

Introduction

The results of usability testing are highly dependent on the skills and abilities of the moderator (Barnum, 2011; Dumas & Loring, 2008; Dumas & Redish, 1999). For example, moderators can influence results explicitly by asking biased questions and providing premature assists, and cognitive biases like the confirmation bias can subtly influence verbal behaviors and task performance.

While some moderation techniques are learned in formal classes on usability evaluation methods, many UX practitioners learn about moderation through trial and error and from literature that highlights best practices for moderators (Barnum, 2011; Dumas & Loring, 2008; Dumas & Redish, 1999; Krug, 2010, 2014; Rubin & Chisnell, 2008).

Dumas and Loring (2008), for example, highlighted the following 10 Golden Rules for interacting with test participants:

1. Decide how to interact based on the purpose of the test.
2. Protect the test participants' rights.
3. Remember your responsibility to future users.
4. Respect the test participants as experts, but remain in charge.
5. Be professional, which includes being genuine.
6. Let the test participants speak!
7. Remember that your intuition can hurt and help you.
8. Be unbiased.
9. Don't give away information inadvertently.
10. Watch yourself to keep sharp.

Each of these rules were broken down into sub rules with explanations to help apply the rules in actual testing. For example, Rule 6, "Let the test participants speak!" has four sub rules:

a. Speakership
b. Appropriate interruptions
c. Judicious speaking
d. Silent communication

While Dumas and Loring (2008) and others focused on core principles of moderating a usability test, Tedesco and Tranquada (2014) expanded on core principles and examined how to handle "unexpected, tricky, and sticky situations" that occur during moderated sessions. They organized advice based on frequency of occurrence. Table 1 presents a sample of frequent, occasional, and rare situations from Tedesco and Tranquada:

Table 1. Sample of Tedesco and Tranquada's Test Participant Situations (2014)

Frequent	Occasional	Rare
Test participant is not thinking aloud.	Test participant struggles excessively with a task.	Test participant starts to look ill or otherwise unwell.
Test participant is not able to complete a necessary task.	Test participant seems annoyed at your neutrality.	Test participant becomes agitated by a product's usability issues.
Test participant is reluctant to say anything negative.	Test participant curses or makes inappropriate comments.	Test participant does something awkward or uncomfortable.
Test participant starts going on a tangent.		Test participant flirts with you.

Books and articles that contain best practices and troubleshooting guidelines for moderation are generally based on the experience of senior usability professionals rather than on empirical evidence of the effects of different moderation approaches. In a 2000 paper, Boren and Ramey (2000) noted that the work of Ericsson and Simon (1993) on verbal protocols was sometimes given as a theoretical framework for usability practice, but that the formal rules were seldom followed by actual practitioners (Boren & Ramey, 2000). The Ericsson and Simon framework for think-aloud usability testing has several key rules:

- Collect and analyze only "hard" verbal data. Don't ask for opinions about what a person "likes" or ask "what do you think about…".
- Give detailed instructions for thinking aloud.
- Remind test participants to think aloud. The basic reminder is "Keep talking."
- Otherwise, do not intervene.

In practice, few practitioners follow the strict think-aloud protocol rules of Ericsson and Simon. Boren and Ramey (2000) noted that usability moderators are not consistent in their think-aloud instructions, their procedures for intervention, or their procedures for giving reminders. Boren and Ramey offer an alternative theoretical framework: speech communication theory.

Both Rubin and Chisnell (2008) and Dumas and Redish (1999) suggested that moderators can use "neutral probes," but in the Ericsson and Simon (1993) framework, even neutral probes are not allowed because they disrupt the work of the test participant and affect the quality of the data.

Hertzum and Kristoffersen (2018) examined 12 think-aloud usability sessions to understand what types of verbalizations were made by moderators before, during, and after a think-aloud usability test. They found that moderators "talked quite a lot" before, during, and after the tasks. Affirmations such as "Mm hm," "Okay," and "Uh-huh" were the most common moderator verbalizations followed by instructions and prompts for reflection. Several practical implications emerged from their work:

- Moderators should not talk so much.
- Moderators should strive to keep affirmations neutral. It appears that little is gained by extending these neutral affirmations with words such as "good," "great," and "wow."
- Moderators should keep a balance between negative and positive verbalizations.
- Moderators should avoid closed questions, which may also be leading. A possible example is "Are you confused by that label?"
- Moderators should avoid asking for answers to hypothetical situations, for example, "What do you think would happen if you …?"
- Moderators should consider that their verbalizations can influence test participants and there is not a good way to evaluate this influence on the study results.

CUE-10 examines moderator verbalizations and behaviors, and test participants' responses in detail. Like Boren and Ramey (2000), we found a wide diversity in actual moderation practices and different interpretations of some existing best practices. For example, some moderators provided detailed instructions on how to think aloud, while others provided only perfunctory instructions. And moderators sometimes used neutral affirmations like "Yeah...," but other times used positive affirmations like "Great!"

This paper provides data on real-world moderator practices and provides suggestions on how to make moderation more effective.

About CUE

This study is the tenth in a series of Comparative Usability Evaluation (CUE) studies conducted from 1998 to 2018. The essential characteristic of a CUE study is that a group of practicing usability professionals agree to evaluate the same product or service and discuss their confidential evaluation results at a workshop.

Previous CUE studies have focused mainly on qualitative and quantitative usability evaluation methods, such as think-aloud testing, expert reviews, and heuristic inspections. An overview of the ten CUE studies and their results is available in a retrospective by Molich (2018).

Goals of CUE-10

The following are the main goals of CUE-10:

- Provide CUE-10 participants with the opportunity to compare their moderation skills to those of their peers and to learn from the differences.
- Examine the techniques used by practitioners while moderating think-aloud usability test sessions.
- Discuss and identify best practices and common issues in moderating usability test sessions.

The following are the key questions that CUE-10 set out to address:

- What are the key issues that distinguish good from poor moderation?
- How can we avoid moderation errors?
- What is good moderation practice?

Method

The following sections present how the organizers of CUE-10 prepared for the study and decided what the usability test sessions would look like, as well as determining what the deliverables would be for the study. This section also describes a discussion of the workshop to present preliminary results, review of videos, and the CUE-10 participants.

Preparations

Fifteen usability professionals and two graduate students participated in CUE-10. They were recruited by posting a call for participation on UTEST, which is a private community for usability professionals, and through a call to participants in previous CUE studies.

The organizers (the two first authors) selected the website used in CUE-10 and coordinated the CUE-10 activities. They did not conduct any usability test sessions. The website was chosen because it is relevant to Americans traveling in Europe and its basic functionality is well-known. Also, the first author knew from previous university courses that it contained interesting usability challenges.

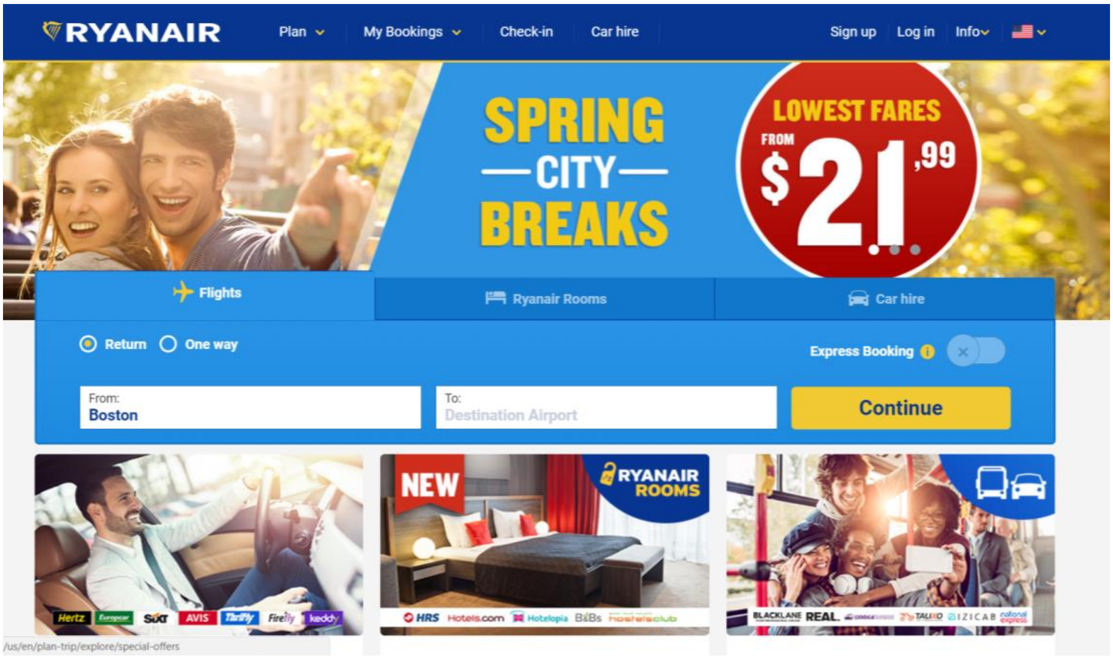

Figure 1. The Ryanair.com home page as it appeared in February–April 2018 when most of the CUE-10 tests took place.

Usability Test Sessions

From February to April 2018, moderators independently moderated usability test sessions of the website Ryanair.com (see Figure 1). Each of the 16 moderators moderated three usability test sessions. See Figure 2.

The full set of instructions for CUE-10 are available online (Molich, 2019).

There were two primary instructions for the test:

- Limit each session to at most 40 minutes, including briefing, interview of test participant, task solution, and debriefing.
- Record all test sessions. The videos must show both the test participant and the moderator in addition to the website.

A total of 48 sessions were recorded: 47 were conducted in person and one was remote, in which the test participant was not visible.

The test sessions all used the same six test tasks:

1. Book a round-trip flight from Madrid to Dublin.
2. Locate and understand rules for carry-on baggage.
3. Find the lowest-priced ticket from London to Copenhagen.
4. Change a flight.
5. Book a multi-leg flight from Copenhagen to Cagliari.
6. Check-in for a flight.

The full task descriptions are available online (Molich, 2019).

Figure 2 shows the roles of CUE-10 participants. CUE-10 had 19 participants of which 16 were moderators and two were organizers. Each of the 15 participating usability professionals and one student moderated usability test sessions. The second student did not moderate.

Sixteen people reviewed the videos: two organizers, 12 moderators, and the two students, who submitted independent reviews. The thirteen videos produced by the 12 moderators who took part in the review and the student team were reviewed. Three moderators did not participate in the review due to time constraints. Their videos were not further reviewed.

Moderators recruited their own test participants. Test participants typically consisted of family members, friends, co-workers, and students—that is, a convenience sample. One moderator (Moderator Q) did their customary professional recruiting process with paid test participants who were selected from a large pool screened for eligibility, gender, ethnicity, age, and travel experience.

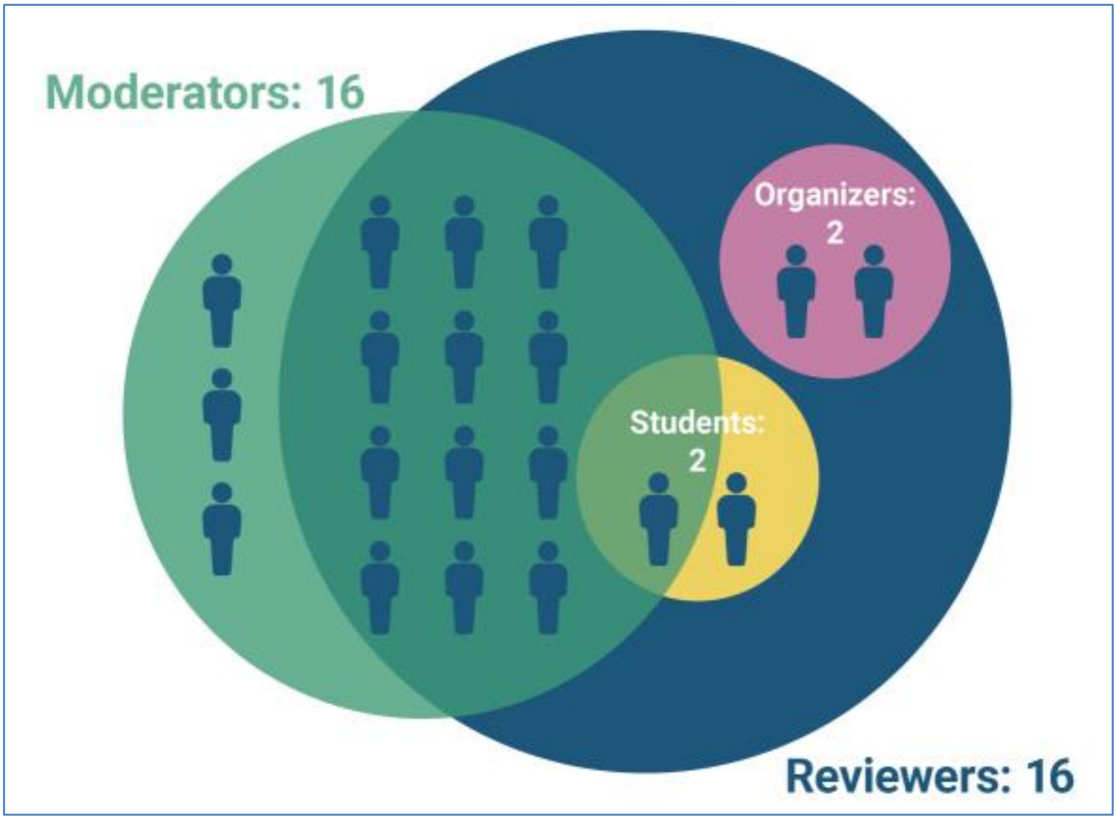

Figure 2. Diagram showing the roles of CUE-10 participants.

Deliverables

After the three test sessions were completed, each moderator submitted the following:

- three raw videos, one from each usability test session. Each video showed the test participant's screen, with the test participant and the moderator displayed as a picture-in-picture.
- the moderator's guide (usability test script).
- an anonymous usability test report where the moderator is identified only by a letter designation (A through Q). The order A through Q does not match the list of moderators alphabetically by either first name or last name.

The anonymous reports from 15 of the 16 moderators are available online (Molich, 2019). One moderator did not want their report published.

The focus of CUE-10 was on moderation rather than the usability test reports. Two moderators did, however, find the CUE-10 usability test reports particularly interesting and published an analysis of them (Cooley & Barnum, 2019).

Because the moderators and test participants are clearly visible on the videos, all moderators agreed at the start of the study that the videos would not be made publicly available. After the analysis had been completed, six moderators and their test participants gave permission to make some or all of their videos publicly available.

Workshop

All 19 CUE-10 participants (see Figure 2) met for a full-day workshop in conjunction with the UXPA Boston Conference in May 2018. The purpose of the workshop was to do a preliminary analysis of the results.

For the workshop, moderators were assigned to four groups of four moderators each. In preparation for the workshop, each moderator reviewed all three videos from the other three members of their group (nine videos total) and the corresponding reports.

Precise review criteria were deliberately not prescribed to avoid biasing the moderators.

At the workshop, each group spent some time discussing the observations they made while reviewing each other's contributions. Each group was asked to reach consensus on 10 to 20 recommendations on critical issues for moderation based on the contributions. Towards the end of the workshop, all CUE-10 participants participated in a general discussion of the four group lists and attempted to merge them into a common list.

Following the workshop, preliminary results from CUE-10 were presented in a panel discussion at the UXPA Boston 2018 conference, and again in a panel at UXPA 2019 in Scottsdale, Arizona.

Review of Videos

CUE-10 is grounded on data from the videos and the usability test reports that moderators submitted. Experience from other sources—for example, previous or later usability tests—are not relevant for the analysis of CUE-10 results and this article.

Ultimately, it was decided that reviewing all 48 videos thoroughly would be unrealistically time consuming. We therefore decided to focus on one video per moderator.

The organizers randomly selected the second or third video produced by each moderator to ensure that the moderator had gained some familiarity with Ryanair's website and the test tasks.

At this time, three of the 16 moderators concluded their CUE participation due to time constraints. The remaining 13 videos were further reviewed. Thirteen moderators and one co-moderator moved on to the next step—review. To simplify the description, both the moderator and the co-moderator are referred to as moderators in the rest of this article.

CUE-10 participants

The CUE-10 moderators were essentially a convenience sample. They were not recruited specifically to provide best practices. Nevertheless, there was a good mixture of qualifications between CUE-10 moderators. Although the range of experience was wide, it could well be wider if a more systematic random sample of participants over the world was taken, but such a comprehensive systematic study would most likely be cost prohibitive.

Table 2. The 19 CUE-10 Participants

Name	Affiliation[b]	First UTEST	# of UTEST[c]	Influence[e]
Andy Hollenhorst	Sallie Mae	2013	250[d]	Bentley: E Rosenzweig
Avram Baskin		2002	-	Demetrios Karis
Beth A. Martin		1999	30	Bob Bailey; HFI courses
Brian Traynor	Mount Royal University	1995	40	Tullis; Dumas books
Cameron Cross[a]	Bentley University	2017	12	Bentley: E Rosenzweig
Carol Barnum	UX Firm	1994	1,000+ [d]	Dumas & Redish book
Chris LaRoche	Northeastern University	1999	250	C Wilson; Pearrow book
Chauncey Wilson[f]	WilDesign Consulting	1980	1,000+ [d]	Dumas & Redish book
Danielle Cooley	DGCooley & Co.	2000	150	Bentley: J Dumas
Devlin McDonough	itsLearning	2016	~40	Krug; Dumas & Loring books
Elizabeth Rosenzweig	Bentley University	1986	110	Nielsen book; many others
Jen McGinn	CA Technologies	2004	50	Bentley: Dumas; Dumas book
Jonathan Patrowicz	MathWorks	2010	150	Krug; Rubin; Dumas books
Kanika Ahirwar[a]	Bentley University	2017	10	Bentley: E Rosenzweig
Mike Ryan	Liberty Mutual Insurance	2008	450 [d]	Bentley: G Almquist
Rolf Molich[f]	DialogDesign	1984	100	Works by J Raskin, C Lewis
Steve Krug	Advanced Common Sense	1988	20	Nielsen; Rubin books
Susan Mercer	Insulet Corp. & Bentley University	2010	300 [d]	Dumas & Loring book
Thanh Nguyen	Mad*Pow	2015	10	Bentley: L Dmitrieva; B Virzi

[a] Indicates graduate student moderator.

[b] A CUE-10 participant's affiliation is listed only if the employer allowed the use of company facilities or work time to carry out part of CUE-10.

[c] Shows the CUE-10 participants' responses to the question "About how many think-aloud usability tests have you facilitated?"

[d] Indicates individual sessions, as opposed to studies.

[e] Shows the CUE-10 participants' responses to the question "What most influenced your approach to usability testing?" Some of the responses have been shortened.

[f] Rolf Molich and Chauncey Wilson organized CUE-10.

What we present is essentially 13 different approaches to moderation. The benefit of CUE-10 is that in this area of moderation, we can comment on what appear to be the strengths and weaknesses within each test session and present them as takeaways.

We are aware of the following biases:

- The CUE-10 participants include many Bentley University students and graduates. On the other hand, the Bentley participants represent a span of graduation years of Bentley students. Danielle Cooley's Bentley experience from 2002 likely bears little resemblance to Cameron Cross's and Kanika Ahirwar's 2018 Bentley experience.

- Many participants were recruited via UTEST. This is potentially limiting and self-selecting as some groups or demographics of UX professionals do not know about this group.
- The moderators often had some familiarity with the test participants which may have affected the interactions during the usability sessions.

Key Insights

In this section we discuss how the key observations were determined, then we give examples of specific observations for each of these key observation areas. The examples come from the analyzed videos.

Deriving the Key Observations

Systematic reviews of selected test sessions (13 videos) were undertaken by 12 moderators, two students, and two organizers as shown in Figure 2. The two organizers each reviewed all 13 videos. Fourteen reviewers each completed reviews of four videos making notes of any observations that they considered interesting. The organizers deliberately left it up to the moderators to define what they considered "interesting observations" to reduce bias. One of the organizers later reported that he picked observations that opposed or supported a best practice.

Each video was thus reviewed by three to five moderators and the two organizers. The process used for reviewing the videos is described in the Appendix.

The first author combined the observations into key observations. The key observations were examined by all reviewers. Some reviewers submitted several observations that were considered different aspects of the same key observations.

Table 3. Areas of Key Observations

Key observation	#reviewers	#observations
Building trust and rapport	7	8
Managing time	7	8
Giving tasks to test participants	5	5
Asking test participants for their opinions	5	5
Structuring usability test sessions	4	5
Giving prompts, probes, and assists	4	5
Preparing for the sessions	3	5
Complimenting test participants	3	4
Making different observations	4	4
Being effective as a moderator	3	3

Building Trust and Rapport

Building trust and rapport begins during recruiting and continues through the entire test session. In this study, moderators used a variety of techniques to build trust and rapport.

The following were the approaches that built trust and rapport that were used throughout the study:

- Let the test participant know that it is important to find out what does not work well. It is okay if things go wrong or are difficult to perform.
    - "So if you are confused by something or find yourself unable to complete a task, that's a big red flag to us that there's something about the site that needs to be changed and that there are likely a lot of other people who are going to have trouble with those same things." (Moderator F)
- Ask background information about the test participant, such as whether they are already familiar with the site or not, or what type of other prior experience might affect the way they interact with the design.

  - “Tell me a little bit about what you do.” (Moderator C)
  - “Are you familiar with Ryanair?” (Moderator M)
- Encourage the test participant if they self-blame.
  - When a test participant stated that she thought she failed a task, the moderator reassured her and told her there were no right or wrong answers and that all feedback helped to improve the site.
    Test participant: “I failed this one.”
    Moderator: “No. There is no failure in this one. Because knowing this difficulty helps us understand where the usability problems are.” (Moderator K)
- Compliment the test participants in non-trivial ways. More about this in the “Complimenting Test Participants" section.
- Inform the test participant that the moderator will act as neutral observer.
  - Example of how the moderator’s role can be defined: “My name is ... . I am going to be walking you through this session today. ... If you have any questions as we go along, just ask them. I may not be able to answer them right away, since we’re interested in how people do when they don’t have someone sitting next to them to help, but if you still have any questions when we’re done, I’ll try to answer them then.” (Moderator L)
  - Three moderators did not define the role of the moderator. (Moderators D, J, Q)

Some approaches hampered the development of trust and rapport. For example, throughout a few sessions, the test participant asked the moderator questions and was met with silence that felt awkward as an observer of the video and presumably awkward for the test participant. The following are two examples from the videos:

- The test participant got a flight and asked, "Is this correct?" There was no response from the moderator. (Moderator D)
- The moderator monotonously read the briefing instructions from a prepared script, creating the impression that the moderator was cold and detached. (Moderator D)

Managing Time

In a usability test session, the moderator has an implicit agreement with the test participant that they will not exceed the agreed length of the test session. The CUE-10 Rules (Molich, 2019) said, “Each video must not be longer than 40 minutes even if this means that the test participant does not get the chance to complete all test tasks.” None of the moderators who exceeded the time limit asked for permission.

The website and the tasks, particularly Task 5, were deliberately chosen so time management might be an issue. A test session that included all six tasks would almost definitely take substantially more than the prescribed 40 minutes. Table 4 shows that four of the 13 reviewed videos (Moderators C, E, K, O) addressed all six tasks within the 40-minute limit. Table 4 also indicates the verbal style of the moderator: Six moderators had a “talkative” style and seven a “quiet” style.

Table 4. Basic data for the 13 Reviewed Test Sessions

Observation / Video	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**L**	**M**	**O**	**P**	**Q**
Session length in minutes	33	41	29	39	39	45	39[a]	37	40	36	38	39 [a]	43
Last task given to test participant	6	6	6	4	4	6	4	6	5	3	6	4	5
Style: Talkative, Non-talkative [b]	T	T	N	N	N	N	N	T	T	N	T	N	T
Number of observations in the reviews, step 3	95	79	75	51	84	84	55	75	114	47	71	87	83

Note: Three moderators (A, B, N) did not participate in the review due to time constraints. Their test sessions were not further reviewed.

[a] Video recordings that appear to have been abbreviated for time.

[b] Talkative style resembles a conversation between the moderator and the test participant; non-talkative style is when moderators limit themselves to confirmations and asking clarifying questions.

The moderators who addressed all six tasks reasonably within 40 minutes used the following techniques:

- Focused on the given task.
  - Test participant: "I don't know where that airport [Stansted] is."
    Moderator: "You are just looking for the lowest fare [so please don't worry about where Stansted is.]" (Moderator L)
  - The test participant wanted to find information about a flight change by using chat.
    Moderator: "Let's not get into chat, we don't have time, it takes forever." (Moderator L)
- Gave hints when the usability problem was clear to the moderator.
  - Moderator: "I am going to give you a hint. Try scrolling. Just in the interest of time." (Moderator L)
  - The test participant had struggled for two minutes to find a direct flight from Copenhagen to Cagliari in Task 5.
    Moderator: "There are no direct flights." The hint directed the test participant to look at two-leg options. (Moderator D)
- Stopped a task when the usability problem was clear to the moderator.
  - The moderator stopped Task 5 (flight to Cagliari) after realizing that the test participant would not be able to solve the task despite the assistance that he just received. (Moderator Q)
  - The test participant was considering various options like seat reservation and baggage options very carefully, which takes quite some time. To cut through this, the moderator terminated Task 1 even though the test participant had not reached check-out. (Moderator M)

The moderators who had problems addressing all six tasks within 40 minutes allowed for the following situations:

- Allowed or encouraged test participants to stray from the given task.
  - The test participant wanted to contact customer service by chat or telephone. The moderator allowed the test participant to continue. (Moderator J)
  - The test participant found "fare finder," which finds cheap flights for dates and destinations determined by Ryanair. The test participant kept on exploring it for 6 minutes, then took himself back to the home page. The moderator allowed this fare finder tangent that was not a reasonable part of any of the given tasks. (Moderator G)
- Encouraged test participants to continue elaborating on topics that were not related to the given task.

    - The test participant was talking about how other sites autocomplete names. Moderator: “Do you like that or do you not like that?” (Moderator O)
    - The test participant scanned the list of other frequently asked questions for about 90 seconds. The test participant talked a lot about the special rules for carrying wedding dresses, footballs, and so on onto a flight. The moderator did not intervene. Considerable time was spent on wondering about the many interesting, subtle questions in Ryanair’s FAQ, but there were no findings. (Moderator F)
- Solicited opinions from the test participant.
    - The test participant was looking at the summary page in Task 1. Moderator: "What are your thoughts when you see this page here?" (Moderator O)
    - At the end of each task the moderator asked: "On a scale of 1 to 5 where 1 means very difficult and 5 means very easy, how would you rate the task?" No follow-up questions were asked. (Moderator Q)

The moderator also has an obligation towards the client to use the time efficiently. Six of the 48 sessions were shorter than 32 minutes. Two of these six sessions (E1 and C3) lasted only 24 and 25 minutes, respectively. It could be argued that these moderators did not explore the user interface with the test participant as thoroughly as time permitted or failed to debrief the test participant to better understand problems.

Giving Tasks to Test Participants

Moderators varied in how they presented tasks to test participants. Some moderators handed the tasks to test participants and asked them to read the task aloud. Other moderators read the task to the test participants.

The approaches that we observed were the following:

- Test participant read the test task out loud after receiving it in writing from the moderator. Nine moderators used this approach (code PL in Table 5).
- Test participant read the test task silently after receiving it in writing from the moderator. One moderator used this approach (code PS in Table 5).
- Moderator read the test task out loud, then handed it to the test participant in writing without asking them to read the task or rephrase it in their own words. Two moderators used this approach (code MP in Table 5).
- Moderator read the test task out loud, but did not hand it to the test participant in writing. One moderator used this approach. This moderator subsequently said that they were not aware of the alternatives and that they had now switched to approach PL (code ML in Table 5).

Table 5. How Moderators Provided Tasks to Test Participants

Video	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**L**	**M**	**O**	**P**	**Q**
Providing tasks to test participant	PL	PL	MP	PL	PL	ML	PS	PL	MP	PL	PL	PL	PL

Note: PL=Participant read task out loud; PS=Participant read task silently; ML=Moderator read task out loud and kept task; MP=Moderator read task out loud and handed task to participant. The codes are explained in detail above.

The main disadvantage of approaches PS, MP, and ML is that the moderator did not confirm that the participant understood the test task. Approach ML has the additional disadvantage that test participants sometimes have to ask the moderator to repeat details while they attempt to complete the task. Some of the moderators who used the PS, MP, and ML approaches said that they were not aware of the PL approach and the tradeoffs.

The following is further advice from the reviewed test sessions:

- Present test participants with one task at a time and provide them with a copy of the task.
- Do not tell the test participant how many tasks you have prepared, and do not hand out all tasks to the test participant at the start of the session (only Moderator F handed out all tasks at once). Doing so will interfere with your option to skip a task, may overwhelm the test participant, or may result in feelings of failure.

Asking Test Participants for Their Opinions

A usability test shows what representative users are able to accomplish with the interactive system when they carry out representative tasks. Eliciting personal opinions from users, or discussing them, is not part of a usability test (UXB, 2018).

The following are examples of how moderators addressed opinions in their briefings:

- Moderator: "As you use the site I'm going to ask you to as much as possible to try to think out loud, to say what you are looking at, what you are trying to do, what you are thinking, This would be a big help to us." (Moderator L)
- Moderator: "As you navigate these tasks, I would like you to think aloud. And that would mean that you would talk about what you are looking at, what you are thinking and what you feel and any of that." (Moderator E)
- Moderator: "... so not just [tell me] what you are doing, like 'I'm entering a city' and 'I am looking for this,' but also how you feel about it, like 'that was easy,' 'that was hard,' or 'I'm a little confused here'."
  Test participant: "Can I make suggestions like 'It might be better if it was this way or that way'?"
  Moderator: "Sure. Not so much design suggestions, but 'From my experience, I expect to be able to do this' or 'I expected it to work this way.'" (Moderator Q)
  Authors' comment: "Sure" is inadvisable because it invites design suggestions, but the rest of this instruction is advisable.

Moderators pointed out that the second and third example mention spontaneous feelings while the first one does not. Most moderators included some mention of reporting "feelings" during their briefing.

Once the tasks began, a few moderators encouraged deviations from simple think-aloud by soliciting opinions.

- Whenever a new page was displayed, the moderator immediately asked, "What do you think of this?" (Moderator O)
- Moderator interrupted test participant during solution of Task 1: "What are your thoughts on when you went back to change [the information about 'From' and 'To' airport], would you expect the details you entered to stay there or be cleared to start again?" (Moderator O)

At least one moderator allowed or even encouraged test participants to engage in design discussions:

- Moderator asked the test participant, "If you were to put some information on this page [Upsell page, "Recommended for you"], what would you expect to see?" The focus deviated from the original task, and this diversion took considerable time. (Moderator P)
- The test participant complained about the length of the pages displayed during check-out: "You have to go to the bottom of the page to see the check-out button." The test participant noticed the check-out button in the upper-right corner but she feared that others might not notice it. (Moderator P)
  Authors' comment: This is an example of a test participant who speaks on behalf of other users. The test participant has no problems, but fears that others might have problems. The moderator can safely ignore such comments and may even tell the test participant that they should not report on whether others would be influenced by an issue.

Table 6. Data for Opinions

Observation / Video	C	D	E	F	G	H	J	K	L	M	O	P	Q
Encourages reflections during think-aloud in briefing	No	No	No	No	No	No	X[a]	No	No	No	No	No	No
Probes for test participant's opinions during task solution: "What do you think of [design feature]?"	No	No	No	No	No	No	No	No	No	No	Yes	Yes	No
Allows or encourages test participant to get into design discussions	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	No

[a] Briefing not included on video.

Structuring Usability Test Sessions

During the initial workshop, the moderators agreed that a usability test session should have the following phases (UXQB, 2018):

1. Briefing: The moderator presents information to the test participant. The briefing is important in setting the expectations correctly with the test participants, which includes explaining their role in the session as well as the role of the moderator. The test participant is informed about the following:
   - purpose of the usability test
   - the procedure
   - their role and contribution, including "We are not testing you."
   - the approximate length of the test session
   - their rights: privacy, data handling, stopping
2. Pre-task interview: The test participant gives information to the moderator. The interview helps the moderator gather background information about the test participant, such as experience with the subject matter area, education, work, familiarity with the product, and other prior experience that might affect the way they interact with the design. Another purpose for this is to provide context for the observers. The moderator might already have most of this information from recruiting, but the observers might walk in not knowing the test participant's background. The pre-task interview is also a helpful way to build rapport. A few moderators said that they do not usually discuss the test participant's prior experience with the domain or tasks because they prefer to have this information surface naturally while the test participant works on the tasks.
3. Task moderation: During moderation, the test participant performs each task, one at a time. The moderator listens, may take notes, and lets the test participant speak. The moderator may provide assistance when needed and keeps the test participant focused on the task and time.
4. Debrief after each task has been completed: A few moderators conducted a short debrief after each task had been completed as shown in Table 7.
5. Final debriefing: The test participant answers questions about their user experience and general impression of the interactive system after the test tasks have been completed. Typical questions are "What did you like most?" and "What is most in need of improvement?" Also, the moderator can ask any probing or open-ended questions about specific things that happened during the test.

Table 7 shows which of these phases were included in the various usability test sessions. All test sessions included task moderation.

Table 7. Data for How Moderators Structured Their Test Sessions

Observation / Video	C	D	E	F	G	H	J	K	L	M	O	P	Q
Briefing	Yes	Yes	Yes	Yes	Yes	Yes	No[a]	Yes	Yes	Yes	Yes	Yes	Yes
Says "We are not evaluating you" or similar	Yes	Yes	Yes	Yes	Yes	Yes	No[a]	Yes	Yes	Yes	No	No	No
Interviews test participant about personal background	Yes	No	No	No	No	Yes	No[a]	No	Yes	No	No	No	No
Interviews test participant about knowledge of travel by air	Yes	No	No	No	Yes	Yes	No[a]	No	No	Yes	Yes	No	Yes
Interviews test participant of knowledge of Ryanair [b]	T1	No	Yes	Yes	Yes	Yes	No[a]	No	No	Yes	Yes	No	Yes
Takes notes during test [c]	No	CS	No	Yes	Yes	No	CS	No	CS	CS	Yes	CS	NT
Debrief after each task	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes
Time used for final debriefing [d]	0	0	1:40	1:40	1:20	2:00	0	0	1:50	2:50	3:00	0	3:15

[a] Video starts with Task 1.
[b] T1 means "during Task 1."
[c] CS means "Can't see." NT: Used a notetaker.
[d] 0 means no debriefing on video.

Some moderators skipped the pre-task interview, which meant that video reviewers had no insight into the participant's background. One moderator asked pre-task interview questions during Task 1 (Moderator C).

Most moderators did not debrief test participants after each task had been completed. They argued that doing a task debriefing might bias the subsequent tasks. It also takes time to debrief after each task which might reduce the number of tasks completed, which is a loss of information. The moderators who debriefed after each task said that clients like this feedback in reports.

Some moderators skipped the debrief after all tasks had been completed. This meant that it was hard to understand what had pleased or bothered the test participant most. This might be an issue of time management.

Giving Prompts, Probes, and Assists

The reviews of the videos showed a need for establishing agreement on what qualifies as good and poor prompts, probes, and assists.

All moderators used these three words consistently and in accordance with the following definitions:

- A **prompt** is an act of encouraging a hesitating test participant.
  A good prompt lets the test participant know that the moderator is engaged and listening. It is equally important that the moderator not bias the test participant.
- A **probe** (intervention) is a question to a test participant during task solution.
- An **assist** is an act of helping a test participant. An assist may be given to help a struggling test participant move on or to skip a well-known or unimportant usability problem.

Example of a **prompt** that moderators recommended:

- "What are you thinking?"
  Authors' comment: Moderators considered this an all-purpose and unbiased prompt, which helps keep the test participant thinking aloud. Some argued that overuse of the prompt could pull the test participant out of the task.
- Contextual example: The test participant was scrolling around the page with various options for the flight. She said, "I wonder if I should…"

Moderator asked, "What were you thinking?" (Moderator K)
The test participant replied that she was wondering if she should get a car.

Examples of poor or missing **prompts**:

- Moderator: "Remember to think out loud." (Moderator E)
  Authors' comment: This is a poor prompt because the test participant was talking all the time and did not need a reminder.
- The test participant hesitated for a moment before she clicked on x to close the "Change flights" pop-up. Moderator: "Yes, you can go ahead and close that." (Moderator L)
  Authors' comment: The prompt was unnecessary.

Examples of **probes** that moderators recommended:

- Test participant: "I'll pick this [option]."
  Moderator: "And why is that?" (Moderator C)
- Test participant: "Is that the return [flight]?"
  Moderator: "What do you think?" (Moderator C)

Examples of poor or missing **probes**:

- Moderator asked, "Did you find the European time format confusing?" (Moderator D)
  Authors' comment: This is a poor probe because it is a closed and leading question.
- The test participant said, "Oops," and it was not clear why. (Moderator D)
  Authors' comment: To clarify the statement, the moderator should have asked what was meant by the "Oops."
- The test participant spent some time looking at the FAQ with the header "Can checked baggage allowances be pooled?" (Moderator H)
  Authors' comment: It was not clear how this was relevant to the task, and the moderator should have intervened (e.g., "I noticed that you hesitated while looking at the FAQ. What are you thinking?")

Examples of **assists** that moderators recommended:

- In Task 3, the test participant only found prices for flights from Luton to Copenhagen. Moderator: "Are there any other flights from London to Copenhagen on that day?" (Moderator F)
  Test participant: "I do not believe there were. There was just this one."
  Authors' comment: The moderator's question was a subtle assist. In the interest of time, the moderator accepted this answer to the moderator's assist even though the answer was not correct.
- Moderator immediately after starting Task 2: "So we can start on [Task 2]. So if you could just click the Ryanair logo on the top left - that will bring you back to the home page." (Moderator J)
  Authors' comment: This was a recommended assist because the affordance of the logo was not important in this test.

Examples of poor or missing **assists**:

- The test participant was filling in the "fly out" date and hesitated. The moderator almost immediately said, "Ryanair is an Irish airline so they are using the European date format." (Moderator C)
  Authors' comment: This assist was premature, as the test participant had not yet articulated confusion about the date format.
- The test participant after working on Task 3 for 3 minutes said, "So I guess this is as far as I can go. I can't fly to Copenhagen." The moderator immediately said, "From Gatwick ..." Test participant said, "From Gatwick ... aah ... OK," then moved on to check for flights from Stansted and Luton to Copenhagen. (Moderator Q)
  Authors' comment: This assist was unjustified because the test participant had not given up and had only worked on the task for a short time.

Table 8. Summary of Examples of Prompts, Probes, and Assists

	Recommended	Not recommended
Prompt	“What are you thinking?”	“Remember to think out loud” when the test participant is already thinking out loud.
Probe	“Why is that?”	The test participant says “Oops” and the moderator does not probe.
Assist	“Are there any other options available?”	"Did you find <feature> confusing?"

Preparing for the Sessions

Some moderators noticed incidents in the videos they reviewed that could be easily avoided if the moderators had been more familiar with the test tasks or conducted pilot sessions.

Examples:

- Several moderators looked surprised when they saw the “this change cannot be made online” message in Task 4, but this message occurs every time, and the organizers had provided alternative solutions for this task. (Moderator C, D, Q)
- The test participant thought that Task 4 was solved, even though they had not answered the given question, “Is it possible to change the flight?” The moderator accepted the incomplete answer even though only 21 of the 40 minutes of the test session had passed. (Moderator K)

Complimenting Test Participants

Many moderators demonstrated elegant ways of complimenting their test participants. The compliments made the test participants feel relaxed and appreciated and built rapport without creating an environment in which the test participants were altering their feedback in the hope of obtaining the moderator’s approval.

The following are examples of elegant and non-trivial compliments used by moderators after task completion:

- “That's great feedback to get. You have done a great job thinking out loud, by the way. Just what I am looking for!” (Moderator M)
- “Thanks for persisting through that task." (Moderator M; Task 2 caused problems for this test participant.)
- Test participant: "Oh, that's it?" Moderator: "Well, you made it [the task solution] look easy, good job." (Moderator C)

Laughing with the test participant can sometimes be a good, indirect way of complimenting the test participant:

- Website confirmed log in, the test participant responded by saying, “Thank you, that was sweet of you, computer!” The moderator laughed with the test participant, which was appropriate. (Moderator C)
- Both the test participant and the moderator chuckled because of the long list that appeared when the test participant clicked destination "Anywhere." (Moderator G)

Reinforcing test participants for providing excellent feedback seemed like a more effective approach than trivial compliments like “You did great on that task” or “Great!”

There were some situations where compliments caused friction with the test participant.

- During a task, trivial or routine sounding compliments like "Great" or “OK” may cause problems.
    - The test participant ignored an upgrade ad and said, "I'll ignore that." The test participant clicked on the Continue button. Moderator said, "good." (Moderator M)
      Authors’ comment: "Good" indicates approval which the person would not have working alone. Positive or negative feedback during task solution may be

  - *interpreted by test participants as meaning they are on the right track or the wrong track and cause them to change the path they are on.*
  - While the test participant was working on a task, a moderator told the test participant that "you are doing great." (Moderator K)
    Authors' comment: If the test participant does not complete the task after being complimented, they may feel a loss of self-esteem.
- At the end of a task, trivial or routine sounding compliments never felt like hints. (Moderator K, L, Q)

Table 9. Data for Compliments to the Test Participant

Observation / Video	C	D	E	F	G	H	J	K	L	M	O	P	Q
Compliments to test participant during the first task, which was presumably where affirmation was most needed	No	No	Yes[a]	No	No	No	No	Yes	No	Yes	No	No	Yes
Compliments to test participant at end of session	Yes	No	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes

[a] The moderator's accentuation indicates that they were unenthusiastic about the compliments.

Making Different Observations

Even though all moderators thoroughly reviewed the assigned videos, many were surprised by the number of observations that they agreed with, but had missed during their reviews. Some moderators said that they had one to three "oops-moments" per video where they thought, "Why didn't I make this obvious observation?"

Moderators agreed on many observations even when they had not originally reported them. There were, however, disagreements. The disagreements were of two types:

- Some considered some observations reported by others to be "of less importance." Meaning, they were correct, but trivial and not worth reporting.
  - "The test participant is given paper to take notes if they need to" (Moderator D)
    Four moderators considered this observation important, while three considered it to be of less importance.
  - "The moderator notes that he wants the test participant to be 'open and honest.' It is an error to mention 'honesty' in a briefing as this implies an assumption on the part of the moderator." (Moderator K)
    Three moderators found this observation important, while two considered it to be of less importance.
- Some differed in their coding of selected observations.
  - "Moderator: Do you think Ryanair doesn't go there [from Naples to Cagliari]? Is that why you're going to go by train from this point - under water? The "under water" suggestion is inappropriate, even for fun." (Moderator Q)
    Five moderators coded this as ME (Moderator Error), while one coded it as GI (Good intervention) and one coded it as AS+ (Assist is justified).
  - "In task 1, the test participant chooses two seats that are not adjacent. One is a middle seat, the other is an aisle seat but on the other side of the aisle. The moderator does not ask why." (Moderator O)
    Three moderators coded this as MI (Missing intervention) and added that it begs a probe; three said it was of less interest.

Discussion

The following sections present a discussion on what the moderators learned, what the key observations were that helped to distinguish poor moderation from good, and how to avoid moderation errors. We also discuss some of the moderators' comments about the study tasks.

What the Moderators Learned

After the workshop, we asked all moderators, "Going forward, are you going to do anything differently as a result of your participation in CUE-10?" Some of the answers were the following (ranked by the most common answers first):

1. Less chat with participants. (Moderators G, O, Q)
2. I will watch my videos more often to check myself. (Moderators B, G)
3. I will stop asking participants to tell me what they are doing when I can see it. (Moderator Q)
4. Wait 5 seconds silently before offering aid. (Moderator A)
5. I will be more aware of my commentary during sessions. (Moderator C)
6. I will be more intentional about the way I both phrase questions and assists as well as the tone I use. (Moderator E)
7. I will redirect people a bit sooner. I will ask for "candid comments" instead of "direct and honest feedback." (Moderator F)
8. Take fewer notes during the moderation portion, rely on my observers and videos. (Moderator G)
9. Always stick to a script. (Moderator N)
10. Nothing specific. (Moderators L, M, P)

One of the moderators who answered that they would change nothing specific added: "I would be more convinced to change some aspects of my approach if I had stronger evidence that I should be doing something differently (certain types of moderation techniques yield greater/better insights), but I don't think we have that evidence quite yet."

None of the moderators asked to see the comments provided by the reviewers of their videos.

What Are the Key Observations that Distinguish Good from Poor Moderation?

All agreed that the moderation in the 13 test sessions reviewed differed considerably.

The moderation errors that reviewers repeatedly noted in this study were the following:

- poor time management, including continuing a task after the usability problem was clear and letting a test participant stray from the test task
- moderation style, in particular talking too much
- using inappropriate or unnecessary prompts, probes, and assists

To the best of our knowledge, no objective measure of moderator effectiveness or efficiency currently exists so we did not attempt to compare these aspects of moderation.

Most CUE-10 participants agreed that there were no obvious differences between moderation by the 12 professionals and the two graduate students in our study. A few CUE-10 participants thought that professional moderators were more effective than graduate student moderators, but when they were challenged, they could not offer any observations to substantiate their claim.

How Can We Avoid Moderation Errors?

Several CUE-10 participants said that they saw the incredible learning value of listening to the other moderators in the workshop and participating in the review. They suggested that moderators should take a critical look at themselves every now and again. They also suggested that moderators should review their own moderated sessions and show some of their moderation videos to independent peers and ask for constructive feedback.

A few CUE-10 participants may not agree with this. One broke off their review of other participants' videos after classifying most of the observations that they had not reported themselves as "Noticed it. Agree. But NOT IMPORTANT ENOUGH to mention."

Humility and openness to criticism are important preconditions for improving the quality of moderated test sessions.

Comments on Tasks

After review, results started to emerge, and some moderators said that they thought that some of the tasks were to blame for the critical remarks that came up.

One moderator said: "Like most other participants, I pilot tested and saw a problem with this [task 4]. But I think we just assumed that since [the organizers] had specified this task [task 4] (and the Cagliari task [task 5]) you [the organizers] had reasons of your own, even if we didn't agree with them and would not specify tasks that way."

A second moderator said: "This was an unusual situation as we did not create the tasks ourselves (which I know I do in usability tests). I also as a moderator did not feel it was in my power or ability to question the tasks – I thought they were created for specific reasons. Even though I did a pilot test and foresaw issues, I did not feel I could modify or bring the attention to this as I assumed the tasks were 'baked in' already."

A third moderator countered: "If a moderator thought that the error message prevented that task [task 4] from ever being completed successfully, or if they felt a task wasn't worded appropriately, they should have flagged that when the organizers first provided the tasks. Not having looked through or run through the tasks themselves, or through pilot studies, shows a lack of preparedness. Some blamed the tasks at the end of the study, but the issue is more evidence of a lack of preparedness on the part of the moderators."

In hindsight, the organizers should have encouraged the moderators even more to comment on the tasks as they were developed.

Tips for Usability Practitioners: What Is Good Moderation Practice?

The following advice is based on what the moderators considered most noteworthy based on their experience in this evaluation and their analysis of the results:

- Manage the limited time you have with the test participant well by focusing on the given tasks, stopping test participants if they explore beyond the given structured tasks, and proceeding to the next task when the usability problem is clear.
- Be prepared! Conduct one or more pilot sessions to get familiar with the product, tasks, and the moderator's guide. Learn about issues that could come up so you do not get surprised. Understand what errors and bugs you might encounter.
- Build trust and rapport during recruiting and throughout the entire test session. Get background information about the test participant. Reinforce the value of the test activity when the test participant self-blames. Compliment the test participants in non-trivial ways.
- Say as little as possible while the participant is carrying out tasks.
- Do not encourage opinions in briefings or during moderation. Say "What are you thinking?" when the test participant is quiet, instead of intruding with a probing and sometimes leading questions.
- Structure usability test sessions: briefing, interview test participant, moderation, and debriefing.
- Speak up if you think there is a problem with the brief or the tasks provided to you by the client.
- Learn from your peers. Several moderators said that they saw the incredible learning value of listening to the other moderators in the workshop.
- Take a critical look at yourself every now and again. Review your own moderated sessions. Show some of your moderation videos to independent peers and ask for honest feedback.

Acknowledgements

Thanks to Elizabeth Rosenzweig and Bentley University for letting us use a room for our workshop at Bentley University. Also, thanks to Aivy Phan and Cheryl Tullis Sirois for creating the graphic in Figure 2.

We thank the anonymous reviewers for their insightful comments on this paper.

References

Barnum, C. M. (2011). *Usability testing essentials: Ready, set…test!* Morgan Kaufmann.

Boren, M., & Ramey, J., (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261–278.

Cooley, D., & Barnum, C. (2019) How do other people do it? A comparative review of usability study reports. *User Experience Magazine*, *19*(1). Retrieved from http://uxpamagazine.org/how-do-other-people-do-it-a-comparative-review-of-usability-study-reports/

Dumas, J., & Loring, B. (2008). *Moderating usability tests: Principles & practices for interacting*. Morgan Kaufmann.

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

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised edition). The MIT Press.

Hertzum, M. & Kristoffersen, K. B. (2018). What do usability test moderators say?: 'mm hm', 'uh-huh', and beyond. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (NordiCHI '18; pp. 364–375). Oslo, Norway: ACM.

Krug, S. (2010). *Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems*. New Riders.

Krug, S. (2014). *Don't make me think revisited: A common sense approach to web and mobile usability*. New Riders.

Molich, R. (2018) Are usability evaluations reproducible? *Interactions,* XXV.6, 82–85.

Molich, R. (2019) CUE-10 – Usability test moderation. Retrieved from http://www.dialogdesign.dk/cue-10/

Rubin, J. & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd Edition). John Wiley & Sons, Inc.

Tedesco, D., & Tranquada, F. (2014). *The moderator's survival guide: Handling common situations in user research*. Morgan Kaufmann.

UXQB (2018) CPUX-F – Curriculum and glossary for the certified professional for usability and user experience, foundation level. Retrieved on January 3, 2020, from https://uxqb.org/en/documents/

About the Authors

Rolf Molich
Mr. Molich manages DialogDesign, a small Danish usability consultancy. He conceived and coordinated the Comparative Usability Evaluation (CUE) studies where more than 140 usability professionals evaluated the same applications. He is the co-inventor of the heuristic inspection method (with Jakob Nielsen). He received the UXPA Lifetime Achievement Award in 2014.

Chauncey Wilson
Mr. Wilson is a UX Consultant. He has published several books and dozens of articles in UX journals and magazines. He was an adjunct professor at Bentley University and has presented often at UXPA, HFES, STC, APA, and CHI conferences. He received the UXPA Lifetime Achievement Award in 2015.

Carol Barnum
Dr. Barnum is co-founder of UX Firm, a UX research consulting company. She conducts research for clients in all sectors, using a variety of tools and techniques. She is the author of 6 books including *Usability Testing Essentials: Ready, Set…Test!* (2nd edition.)

Danielle Cooley
Ms. Cooley has been working in design research and strategy for 20 years and teaches Content Strategy at Northeastern University's College of Professional Studies. She runs a bespoke consulting practice, providing services for such organizations as Hyundai, Graco, Equifax, Ascension, and The Federal Reserve Bank of St. Louis. She sometimes tweets @dgcooley.

Steve Krug
Mr. Krug is best known as the author of *Don't Make Me Think: A Common Sense Approach to Web Usability* and the usability testing handbook *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems*. He has taken part in three CUE studies and has always learned a great deal from them.

Chris LaRoche
Mr. LaRoche is a senior user experience consultant at the Massachusetts Institute of Technology, focusing on researching and evaluating the accessibility & usability of Web sites. He is also a Senior Lecturer at the College of Professional Studies (CPS) at Northeastern University, where he has taught for several decades.

Beth A. Martin
Ms. Martin began her public sector career in the federal government reference work for human centered design at Usability.gov. In addition to her current role as UX Lead, she has served as an adjunct professor at the University of Virginia, Fairmont State University and at the Maryland Institute College of Art.

Jonathan Patrowicz
Mr. Patrowicz is a Senior UX Analyst at a Natick, MA based software company. He is involved in both UX design and research, with a penchant for usability testing. He also serves as the Volunteer Chair for the Boston UXPA and has been involved in organizing the last four annual conferences. Patrowicz received his master’s from Bentley University in 2013.

Brian Traynor
Mr. Traynor is Chair and Associate Professor in Information Design. His research interests include Job Performance and Information Comprehension, Design Teaching Methods, and User Attribution of Blame. He spent 20 years in Telecom and IT Technical Communications before returning to an academic environment.

Appendix: Review of Videos

The 13 videos were reviewed in the following manner:

1. The organizers developed an initial set of codes for classifying observations on the videos.
2. The organizers and two moderators performed a pilot review of four videos. They each coded a list of observations that they considered interesting. The organizers deliberately left it up to the experienced moderators to define what they considered "interesting observations" to reduce bias.
3. The organizers and the two moderators discussed and revised the codes. Finally, they updated their list of observations to reflect the revised codes. The final list of codes is shown in Table A.1.
4. Fourteen moderators each reviewed four of the 13 videos using Table A.1 to code observations. They each coded a list of observations that they considered interesting. Again, the meaning of "interesting" was left to the moderators to reduce bias.
5. There were 14 rather than 13 moderators in this step because the two graduate students carried out separate reviews in this phase. In addition, the two organizers each independently reviewed all 13 videos. Due to time constraints, not all moderators were able to review all four videos assigned to them. Each video was reviewed by five to seven people.
6. The first author merged the observations from the 14 moderators plus two organizers into 13 spreadsheets, one for each video. Each spreadsheet contained from 50 to 110 coded observations. All observations were time stamped, except for a few that were marked "General observation." Observations were ordered by video time. Moderators were anonymous in the spreadsheets.
7. The consolidated spreadsheet was returned to the relevant moderators for a second review. Each moderator was asked to comment on and code the observations that they had not reported, but had been reported by other moderators.
8. Moderators were also asked to review their previously submitted observations. They were allowed to change their observations based on the consolidated information.
9. Based on their experience from the reviews, each of the 14 moderators plus the two organizers each listed the six to eight most important insights that they had gained from CUE-10.
10. The insights that were brought up most often in Step 6 are discussed in the "Key Insights" section.

Table A.1. Codes for and Examples of Observations on Videos

This table uses the abbreviation TP for Test Participant and M for Moderator.

Description	Code	Explanation	Example from CUE-10
Assist - justified	AS+	A timely and well-phrased nudge that helps the TP to continue with the task after the usability problem is clear	TP wants to take train from Naples to Cagliari, which is not possible. M: "What if you wanted to take ... make it two flights, what would you do?" (C)
Assist - not justified	AS-	Premature, excessive, or otherwise unnecessary assistance from the M to the TP	M explains that Ryanair "security fast track" is the same as US "pre-check." (C)
Incorrect assist or probe	ASI	Incorrect information provided by the M in an assist or a probe, or assist happens too quickly	TP types in Madrid; M asks her what she is thinking. It felt like the probe came too quickly—only a few seconds after TP typed in Madrid into the text field. If M had waited 10 seconds or so, this would have been a good probe. (L)

Description	Code	Explanation	Example from CUE-10
Good intervention or probe	GI	A timely and well-phrased question that asks the TP to clarify an important issue	M: "Can you go back to [the previous page?] to show what you had in mind?" It may be quite clear but M obviously wants to be sure. (P)
Missing intervention or probe	MI	The M should have intervened to ask the TP to clarify an important issue	TP says, "I really don't like these 'list things.'" M just says OK. It wasn't clear what the TP didn't like about the list. M could have followed up. (L)
Unnecessary intervention or probe	UI	The intervention or probe is unnecessary small talk, provides unnecessary information, asks for unnecessary information, or attempts to direct the TP in a new direction while the TP is working on a test task	M reminded TP to think aloud. TP has been thinking aloud quite well throughout the study, including at the moment M reminded him. (E)
Positive statement	POS	A positive statement by the M intended to encourage the TP	"That's great feedback to get. You have done a great job thinking out loud, by the way. Just what I am looking for!" (M)
Unnecessary positive statement	UPOS	Unnecessary, exaggerated, or unfeeling positive statements reduce the effect of positive statements when they are really needed	TP selects flight information and clicks on Let's Go, and M says "You're doing great." This task is not yet complete and telling the TP that she is "doing great" is helping with the task. (K)
Negative statements	NEG	A statement from the M that could be interpreted as condescending by the TP	M asks TP "Do you know what we are doing?" This struck me as perhaps condescending. To verify understanding is OK, but maybe something like "Is the task clear?" (P)
Stressful situation - Moderator	SM	A situation where the M is not sufficiently in control of the situation to answer a reasonable question from the TP	The TP has entered the email address correctly in order to log in to the account. She asks for the password. Twelve seconds pass while the M is looking in his papers for the password. M: "I can give you a password or you can ... Let's do the confirmation." (K)
Stressful situation – Test participant	SP	A situation where the TP's utterings or body language indicates that they are stressed	TP: "Gosh, I am getting all frustrated with this now." (K)
Good time management	TPOS	An intervention by the M that helps to ensure the best utilization of the limited time available for solving the given test tasks	M: "And so what would you normally do now? Would you ..." TP: "I'd phone them." TP and M discuss the problems for another 30 seconds, which is OK. Then M correctly moves on to next task without asking the TP to locate the phone number. (K)

Description	Code	Explanation	Example from CUE-10
Poor time management	TNEG	An intervention by the M that directs the TP in a direction that is not helpful towards understanding the usability issues related to the task set	M asks the TP to click again on the link for "Fees," which was broken. M says, "Try clicking on the table just one more time—I don't know why that didn't…. Ha, OK." There was no reason to test a link that was already shown to be broken to satisfy M's curiosity. (L)
Moderator error	ME	The M makes an error in moderation. Includes "Things you should never say"	1. M starts video recording before TP has signed release form. (D) 2. TP talks about Ryanair nickel and diming people. M says, "That's the reputation they have." M referred to the company's reputation. This doesn't seem appropriate. (J)
Configuration error	CE	The system is incorrectly set up, usually due to a M error	HP Printer Assistant pops up when trying to open Ryanair, and an assistant comes in to correct. (Q)
Interesting usability issue	USA	A usability problem	TP notes that the text for the confirmation number is very small. (P)
Personal note	Pers	A note made for the reviewer's personal needs	Overall, the M let the TP investigate the Web site and the tasks for long periods of time in the session without intervention. (G)
Other – explain	O	An observation that does not fit any of the above categories	TP checked Stansted first, noted the lowest price for [departure from] that airport. It appears that M did not provide paper and pen for the TP to record information as needed, for example, the lowest price at each airport. (H)