

Remote Usability Testing Beyond Screen Sharing:

A Comparative Study of Assisted and Traditional Think-Aloud Methods

Frida Takman

DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2026

MASTER THESIS

ASSA ABLOY



Remote Usability Testing Beyond Screen Sharing:

A Comparative Study of Assisted and Traditional Think-Aloud Methods

Frida Takman



LUND
UNIVERSITY

Remote Usability Testing Beyond Screen Sharing: A Comparative Study of Assisted and Traditional Think-Aloud Methods

Frida Takman

May 25, 2026

Copyright © 2026 Frida Takman

Published by

Department of Design Sciences
Faculty of Engineering LTH, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden

Supervisor: Kirsten Rasmus-Gröhn

Co-supervisor: Roger Dreyer, Jonathan Vesterlund

Examiner: Mattias Wallergård

Abstract

When user interfaces are embedded in physical systems and cannot be accessed through conventional screen sharing, remote usability testing becomes challenging. This is increasingly relevant as products are developed for global markets, where remote evaluation is often necessary despite limited user interface access. This study investigates how alternative testing methods can be used and how they influence usability evaluation outcomes.

Four methods were compared in a controlled 2×2 study, varying in physical co-presence (on-site vs. remote) and interaction execution (direct think-aloud vs. proxy interaction), using a door control panel as a test case (n = 20). Participants completed scenario-based tasks, either through direct interaction or by instructing a moderator, while qualitative and quantitative data were collected, complemented by moderator reflections.

The results show that while task performance and perceived usability were similar across methods, clear differences emerged in interaction and identified usability problems. Direct interaction primarily revealed functional and physical issues, whereas proxy interaction highlighted problems related to interpretation and system logic, generating richer verbal data but increasing cognitive load.

The moderator played a central role, particularly in proxy and remote settings, where interaction became distributed between participant and moderator. This shifted the moderator from observer to active interpreter and executor, making usability issues partly dependent on how interaction was communicated. The findings show that methods enabling remote testing are not neutral substitutes for in-person testing, but reshape both interaction and outcomes.

This suggests that usability testing approaches should not be selected merely to enable access, but designed with an understanding of how they shape the insights produced.

Keywords: Usability Testing, Remote Usability Testing, Proxy Interaction, Think-Aloud Method, Moderator Role

Sammanfattning

För användargränssnitt som är inbyggda i fysiska system och inte kan nås genom traditionell skärmdelning blir användbarhetstestning på distans utmanande. Detta blir allt mer relevant i takt med att produkter utvecklas för globala marknader, där utvärdering på distans ofta är nödvändig trots begränsad åtkomst till användargränssnittet. Denna studie undersöker hur alternativa testmetoder kan användas och hur de påverkar resultaten av användbarhetsutvärderingar.

Fyra metoder jämfördes i en kontrollerad 2x2-studie, där fysisk närvaro (på plats respektive på distans) och hur interaktionen genomfördes (direkt think-aloud respektive proxyinteraktion) varierades, med en dörrkontrollpanel som testfall (n = 20). Deltagarna genomförde scenariobaserade uppgifter, antingen genom direkt interaktion eller genom att instruera en moderator, medan kvalitativ och kvantitativ data samlades in, kompletterad med moderatorreflektioner.

Resultaten visar att även om deltagarnas uppgiftsutförande och upplevda användbarhet var likartade mellan metoderna, framträdde tydliga skillnader i interaktion och identifierade användbarhetsproblem. Direkt interaktion synliggjorde främst funktionella och fysiska problem, medan proxyinteraktion lyfte fram problem kopplade till tolkning och systemlogik, och genererade rikare verbal data men ökade den kognitiva belastningen.

Moderatorn spelade en central roll, särskilt i proxy- och distansbaserade upplägg, där interaktionen blev distribuerad mellan deltagare och moderator. Detta innebär att moderatorn gick från observatör till aktiv tolkare och utförare, vilket gjorde att identifierade användbarhetsproblem delvis blev beroende av hur interaktionen kommunicerades. Resultaten visar att metoder som möjliggör distanstestning inte är neutrala ersättningar för tester på plats, utan formar både interaktion och resultat.

Detta indikerar att val av testmetod inte enbart bör göras för att möjliggöra åtkomst, utan bör utformas med en förståelse för hur de påverkar vilka insikter som genereras.

Nyckelord: Användbarhetstestning, Användbarhetstestning på distans, Proxyinteraktion, Think-aloud-metod, Moderators roll

Acknowledgements

I would like to extend a sincere thank you to my academic supervisor, Kirsten Rasmus-Gröhn, for her guidance, encouragement, and support throughout this project.

A big thank you also to my supervisors at ASSA ABLOY, Jonathan Versterlund and Roger Dreyer. Roger, thank you for trusting me with this project and giving me the freedom to shape it. Jonathan, thank you for always being available to help, whether creating the test setup or discussing details throughout the whole project.

Thank you to Michiel de Kok for your valuable support with testing logistics and for helping me connect with the right people.

I am very grateful to all participants who took part in the study, as well as the managers who helped make the recruitment possible.

Lastly, thank you to my family and friends for their encouragement, and especially to Mariia for being a constant source of support throughout this process, both in person and remotely.

Declaration of Generative AI Use

I, Frida Takman, hereby declare that I have used generative artificial intelligence tools (e.g., ChatGPT) solely as support in my writing process and data analysis. This includes restructuring text, suggesting alternative phrasing, improving clarity, and refining language.

Generative AI has primarily been used in parts of the data analysis related to moderator comments. In this context, the tool was used to increase efficiency by assisting in identifying key insights within specific questions and responses. Subsequently, it was used iteratively to condense the material into more comparable analyses.

I have continuously reviewed and verified all AI-generated output to ensure that it accurately reflects the data and does not introduce unsupported distortions or interpretations.

Generative AI was *not* used in the thematic analysis of finding usability problems. However, it was used in later stages to support the formulation and presentation of findings. This also applies to the quantitative analysis.

Generative AI has also been used to discuss theoretical aspects and support the interpretation of findings, primarily to clarify uncertainties. These interactions were not treated as authoritative sources, but rather as a reflective dialogue to explore alternative perspectives.

All ideas, results, and conclusions are my own. All AI-generated content has been reviewed, verified, and edited by me, and all sources have been read by me.

Contents

1	Introduction	9
1.1	Background	9
1.2	Purpose	10
1.3	Problem Description	11
1.4	Research Questions	12
1.4.1	Concise Analytical Clarification of Each Research Question	12
1.4.2	Theoretical Framing of the Research Questions	13
1.4.3	Clarification of "Usability" & "User Experience"	13
1.5	Related Work	13
1.6	Limitations	15
1.6.1	Methodological Limitations	15
1.6.2	Practical & Contextual Limitations	15
1.6.3	Participant-Related Limitations	15
1.6.4	Analytical Limitations	16
1.6.5	Single-Moderator Considerations	16
1.7	Relevance for Sustainable Development	16
1.8	Ethical Considerations	17
1.9	Outline of the Thesis	17
2	Theory & Literature study	18
2.1	Theoretical Background	18
2.1.1	Distributed Cognition	18
2.1.2	Situated Cognition	19
2.1.3	Embodied Interaction/Embodiment	19

2.2	Methodological Background	20
2.2.1	Usability Testing & Usability Problems	20
2.2.2	Comparing Test Methods	20
2.2.3	Think-Aloud Methods	21
2.2.4	Proxy Interaction	22
2.2.5	Remote Testing	23
2.3	Linking Theory to the Research Questions	24
3	Test Design	25
3.1	Research Approach & Methods	25
3.2	User Profile & Participants	27
3.3	Test Setup & Equipment	28
3.4	Pilot Study	30
3.5	Test Procedure	31
3.5.1	Phase 1: Preparation	31
3.5.2	Phase 2: Usability Test Execution	32
3.5.3	Phase 3: Moderator Reflection	33
3.6	Test Moderator Role	34
4	Data Collection & Analysis	35
4.1	Data Collection	35
4.2	Analytical Approach	36
4.3	Analysis Procedures	37
4.3.1	Qualitative Analysis – Usability Problems	37
4.3.2	Qualitative Analysis – Moderator Debriefing	42
4.3.3	Quantitative Analysis	44
5	Results	46
5.1	Summary of Method Differences	46
5.2	Usability Problems & Severity	47

5.3	Task Performance	48
5.4	Perceived Usability & Experience	50
5.5	Interaction Characteristics	51
5.6	Moderator-Oriented Adaptation	52
5.7	Coordination, Misunderstandings, & Context	54
5.8	Moderator Influence	57
5.9	Participant Connection	58
5.10	Debriefing & Post-Task Behavior	59
5.11	Summary of Key Results	60
6	Discussion	63
6.1	Process & Methods	63
6.1.1	Preparations	63
6.1.2	Participants	64
6.1.3	Order Effects & Data Analysis	64
6.1.4	Single-Moderator Limitation	65
6.2	Results	65
6.2.1	Task Performance, Perceived Usability & Experience	65
6.2.2	Usability Problems	66
6.2.3	Think-Aloud vs Instruction-Based Interaction	68
6.2.4	Remote vs On-Site: Control & Connection	69
6.2.5	The Moderator's Role in Balancing Support & Neutrality	70
6.3	Theoretical Interpretation	71
6.4	Relation to Research Questions	72
6.4.1	RQ1 – Reliability of Testing Methods	72
6.4.2	RQ2 – Proxy Interaction, User Behaviour, & User Experience	72
6.4.3	RQ3 – Moderator Role Across Methods	73
6.4.4	Overall Reflection	74
6.5	Relevance for Sustainable Development	74
6.6	Ethical Considerations	74

6.7 Further research	75
7 Takeaways for Practitioners	76
8 Conclusion	77
References	78
Appendix:	
A Script for the Usability Tests, incl. Tasks	81
B Debriefing Questions - Participant	89
C Debriefing Questions - Moderator	93
D Usability Problems with Severity Scores	99
E Data Analysis – Supporting Figures & Tables	106

1. Introduction

This chapter introduces the study and outlines its overall structure, beginning with the background and context of the project. It then defines the purpose of the study and presents the research questions that guide the investigation. Furthermore, the chapter situates the study within existing research on usability testing methods, outlines key limitations, and provides an overview of the thesis structure.

1.1 Background

Remote usability testing typically relies on direct user interaction. But what happens when the participant cannot interact with the product at all?

This challenge emerged during usability testing of a prototype under development at ASSA ABLOY, a global company specializing in access solutions such as automatic doors, locking systems, and entrance automation [1]. The product in focus was a control panel for sliding doors, also referred to as a door mode selector, used to configure operating modes and display system feedback such as error messages.

Unlike conventional digital user interfaces, the prototype was implemented directly in the physical device, meaning that the user interface could not be accessed or shared through standard screen-sharing solutions. At the same time, the product was intended for international markets, making usability testing beyond the local context necessary. This created a methodological challenge where remote usability testing became difficult or even impossible to conduct using traditional approaches.

While this constraint stems from how the near-final prototype is embedded in the device, it also reflects the nature of similar systems, which are often physically integrated and difficult to isolate from their hardware environment.

The user interface consists of a 50×50 mm touch screen with limited memory capacity and is intended to be installed directly adjacent to the door as part of a larger door system (see Figure 1.1). Although the prototype was tested as a standalone unit in a laboratory setting, the final product is designed for everyday use, enabling users such as retail staff to control door behavior in real-world environments.

As direct interaction and screen sharing were unavailable, alternative, non-standardized testing configurations were required to enable remote usability testing in this context. However, since such approaches are not well established, their reliability remains uncertain, raising the question of whether they can produce results comparable to traditional usability testing while still supporting meaningful user experiences.

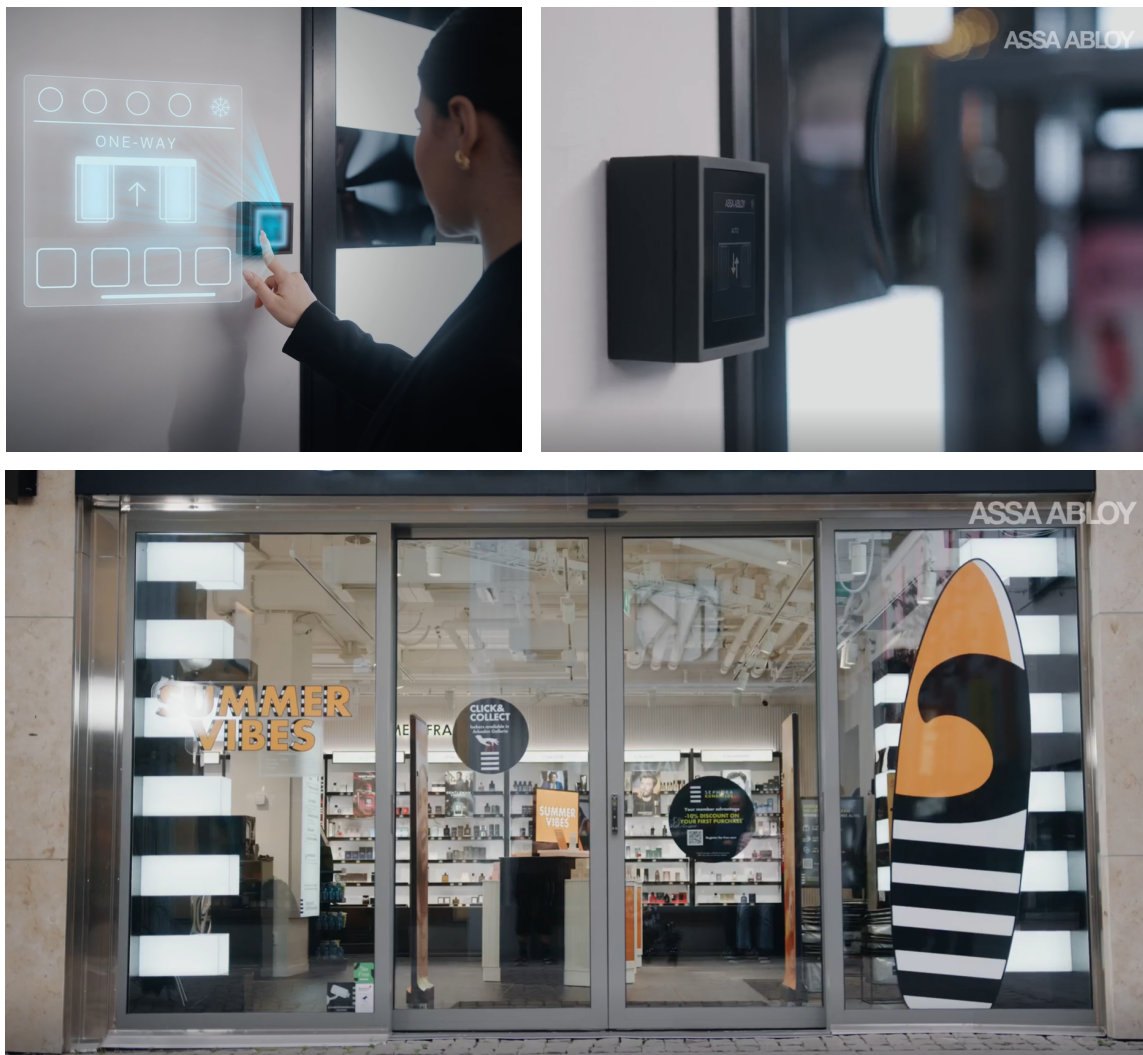


Figure 1.1: Sliding door system with an integrated control panel for configuring operating modes and displaying system feedback (e.g., error messages), illustrating how the final product will appear in its real-world setting. (Pictures from ASSA ABLOY).

1.2 Purpose

Based on this methodological challenge, the aim of this thesis is to investigate alternative approaches to remote usability testing for products that do not allow conventional screen sharing, and to evaluate how these approaches influence usability evaluation outcomes.

To address this, four usability testing methods are compared, varying along two dimensions: physical co-presence (remote vs. in-person) and interaction execution (direct vs. proxy interaction). In this study, proxy interaction refers to a form of interaction in which the participant does not engage with the system directly, but instead provides instructions while the moderator performs actions on their behalf. The moderator is understood as the person facilitating the usability test, responsible for guiding the session and, in proxy interaction methods, executing these actions.

This distinction also gives rise to different forms of verbalisation, enabling comparison between traditional think-aloud protocols and situations where participants articulate instructions rather than actions. Through this comparative approach, the study examines how different configurations affect reliability, interaction behaviour, user experience, and the role of the moderator.

Beyond the methodological comparison, the study aims to support usability testing in industrial contexts involving physical and embedded systems, where user interfaces are tightly coupled to hardware and cannot be easily accessed remotely. By identifying how different testing approaches perform under these constraints, the study provides guidance on how to design and conduct usability tests when direct interaction with the user interface is not feasible.

The control panel serves as a controlled test object, enabling systematic comparison between the methods. While the primary focus is methodological, the study also provides formative insights into the usability of the product in its current development stage.

1.3 Problem Description

Alternative approaches to remote usability testing, such as proxy interaction and mediated setups, introduce fundamentally different forms of interaction compared to traditional usability testing. Rather than interacting directly with the user interface, participants may rely on a moderator to execute actions, resulting in a redistribution of control between participant and moderator.

These changes introduce several methodological uncertainties. When interaction is no longer direct, it becomes unclear how users interpret and communicate their understanding of the system, and how this shapes both interaction behaviour and perceived experience. Similarly, when interaction is mediated or distributed, the identification of usability problems may depend on how instructions are communicated and interpreted, rather than solely on the user interface itself.

Furthermore, these approaches shift the role of the moderator from observer to active participant. This raises questions about influence, consistency, and control, as the moderator's actions and interpretations may shape both the interaction process and the resulting findings.

At the same time, the need for remote testing remains, particularly for products intended for international markets. While such approaches enable access to a broader participant base, they also introduce practical challenges related to communication, technical setup, and reduced access to contextual and nonverbal cues, which may affect the quality and interpretation of the data collected.

The problem addressed in this thesis is therefore methodological: to what extent alternative usability testing methods, particularly those involving proxy interaction and varying degrees of co-presence, can produce reliable and valid usability insights while maintaining meaningful interaction and perceived user experience, and preserving a consistent and controlled role for the moderator.

1.4 Research Questions

To examine this, the study compares usability testing methods varying in co-presence and interaction execution, focusing on the following key aspects: reliability, interaction processes and user experience, and the role of the moderator.

- RQ1. To what extent do different usability testing methods, varying in co-presence and interaction execution, produce comparable results in terms of reliability?
- RQ2. How does proxy interaction affect users' interaction behavior, verbalization patterns, and perceived experience compared to direct interaction?
- RQ3. How does the role and influence of the moderator differ across the tested usability methods?

1.4.1 Concise Analytical Clarification of Each Research Question

RQ1 – Reliability of Testing Methods

RQ1 focuses on methodological reliability. The analysis examines differences in identified usability problems, task performance, and participant feedback across methods that vary in co-presence and interaction execution. Particular attention is given to how differences in verbalisation, specifically the shift from think-aloud to instruction-based communication, influence the identification and interpretation of usability problems. The central concern is the extent to which alternative testing approaches, particularly those involving proxy interaction and remote elements, can provide findings comparable to traditional in-person usability testing.

RQ2 – Proxy Interaction, User Behaviour, & User Experience

RQ2 examines how the redistribution of interaction affects users' interaction behaviour and overall experience. The analysis considers differences between direct interaction and proxy interaction in terms of communication patterns, coordination of interaction, and participants' perceived sense of control and usability. Particular attention is given to how verbalisation shifts from action-based think-aloud to instruction-based communication, and how this reshapes both observable interaction behaviour and the participants' subjective experience of the interaction.

RQ3 – Moderator Role Across Methods

RQ3 investigates how the moderator's responsibilities and influence vary across the different test methods. The analysis focuses on communication patterns, interpretation of instructions, pacing, and the moderator's role as an active part of the interaction system, particularly in mediating user instructions in proxy-based interaction.

1.4.2 Theoretical Framing of the Research Questions

The research questions are examined through three complementary theoretical perspectives: Distributed Cognition, Situated Cognition, and Embodied Interaction. These perspectives are used as analytical lenses to understand how interaction, communication, and interpretation are shaped across the different testing methods.

In proxy interaction, cognitive processes are distributed across the participant, the moderator, and the user interface, forming a shared cognitive system [2, 3]. From this perspective, Distributed Cognition supports the analysis of how mediated interaction may affect users' understanding, decision-making, and identification of usability problems.

From a Situated Cognition perspective, cognition is inseparable from context [4, 5]. In this study, differences in co-presence and moderator involvement shape how participants interpret tasks, feedback, and instructions, and are therefore relevant for understanding variations in reliability and user experience across methods.

From an embodiment perspective, interaction is shaped by users' physical engagement with the user interface [6]. In proxy interaction, this direct engagement is removed, which may affect users' sense of control and understanding. At the same time, the moderator's physical presence, including gestures and timing, becomes more influential in shaping the interaction.

1.4.3 Clarification of "Usability" & "User Experience"

In this thesis, *usability* primarily refers to outcomes related to the usability evaluation, such as task success, identified usability problems, perceived usability, and reliability across testing methods. In contrast, *user experience* refers to how participants experience and engage with the interaction during the test situation.

The term *user experience* is occasionally used in a broader sense to describe how different testing methods shape the overall interaction process. This includes both participants' subjective experiences, such as confidence, effort, control, and naturalness, and observable interaction behaviours, such as verbalization patterns, pacing, hesitation, coordination, and moderator reliance.

However, analytical distinctions are made throughout the thesis between *interaction behaviour*, referring to observable interaction processes, and *perceived experience*, referring to participants' subjective perceptions of the interaction.

1.5 Related Work

This thesis builds on established research in usability evaluation, particularly within think-aloud methodology and remote usability testing. The think-aloud method has long been applied to capture users' cognitive processes during interaction with a system. By encouraging participants to verbalize their thoughts while performing tasks, researchers gain insight into reasoning, misunderstandings, and decision-making processes [7]. Think-aloud testing is widely used in usability evaluation,

particularly in formative design phases, and is often considered an accessible method for identifying usability problems that may not be evident through performance metrics alone [8].

With increasing globalization and distributed product development, remote usability testing has become more common. Several studies have compared remote and in-person usability testing methods, often focusing on differences in problem identification and, in some cases, aspects such as efficiency and participant recruitment [9, 10, 11]. These studies are typically based on synchronous setups that aim to replicate traditional laboratory-based think-aloud testing, relying on video and audio communication combined with remote desktop sharing to allow evaluators to observe the user interface in real time while participants interact with it directly [11].

Such remote synchronous methods have been shown to produce results comparable to in-person usability testing under these conditions [11]. However, these findings rely on the assumption that evaluators have direct visual access to the user interface. This highlights an important limitation in existing research: many remote usability testing approaches assume that screen sharing or equivalent technologies are available. Less attention has been given to situations where such access is not possible, and where interaction must instead be mediated or redistributed.

Research on remote usability testing methods without screen sharing remains limited. In such situations, alternative approaches are required. Previous work has explored assisted or modified think-aloud methods in which participants verbalize instructions while a facilitator performs actions on their behalf [12]. These approaches demonstrate how interaction can be redistributed between participant and facilitator, and suggest that such mediation may influence users' sense of control, coordination of actions, and underlying cognitive processes.

While previous work has explored assisted or modified think-aloud approaches, it has primarily focused on feasibility and adaptation for specific contexts rather than on systematically comparing different interaction structures [12]. In contrast, this study takes inspiration from these approaches but examines proxy interaction as a core methodological condition, where participants consistently instruct a moderator who performs all actions.

This addresses a gap in understanding how mediated interaction affects usability evaluation when direct interaction is not possible, as well as how the moderator's role varies across different test conditions.

To address this, four usability testing methods are systematically compared: in-person moderated usability testing with concurrent think-aloud (On-site think-aloud), in-person assisted think-aloud with proxy interaction (On-site proxy), remote assisted think-aloud with proxy interaction (Remote proxy), and remote moderated usability testing with concurrent think-aloud (Remote think-aloud), as presented in Table 1.1 below. This comparison provides insights into how methodological differences affect reliability, user experience, and moderator involvement.

Table 1.1: The methods and their differences. Horizontal: physical co-presence, vertical: interaction execution.

	On-Site	Remote
Direct (Think-Aloud)	Method 1 (On-site think-aloud)	Method 4 (Remote think-aloud)
Proxy	Method 2 (On-site proxy)	Method 3 (Remote proxy)

1.6 Limitations

Several methodological, practical, and analytical limitations were present during the planning and execution of this study, which may influence the interpretation and generalizability of the findings.

1.6.1 Methodological Limitations

The primary focus of this thesis is the comparison of usability testing methods rather than the evaluation of the final product. Consequently, the control panel serves as a controlled test object to enable systematic comparison between methods. Some aspects of the test procedure were therefore intentionally designed to highlight differences between testing methods, which may have influenced how usability issues were revealed across conditions.

In addition, the tested prototype did not include all planned features of the final product. As a result, some interactions could not be evaluated, potentially limiting the range of usability problems identified.

1.6.2 Practical & Contextual Limitations

The study relied on a simulator to emulate door behaviors such as operating modes and error messages, since the control panel was not connected to an actual door. While this setup was consistent across all test methods and ensured comparability, it may not fully reflect real-world usage conditions. The absence of real system feedback (e.g., visual or auditory responses from the door) could have influenced participants' perception of realism and overall interaction experience.

1.6.3 Participant-Related Limitations

Participants were recruited internally from the company's sales organization due to organizational constraints. As a result, participants possessed greater product familiarity and domain knowledge

than the intended end-user group (e.g., retail staff). This prior knowledge may have reduced the likelihood of encountering usability problems that less experienced users might face or biases.

However, since the study focuses on comparing testing methods, internal participants were considered acceptable for the methodological purpose.

1.6.4 Analytical Limitations

As the study compares different usability testing methods, the observed differences in identified usability problems and user behaviour are shaped not only by the methods themselves, but also by how interaction unfolds within each setup. Factors such as how participants communicate, how instructions are interpreted, and how the moderator enacts interaction are integral to each method and therefore part of what is being examined.

Rather than isolating individual variables, the study captures how these factors jointly influence the outcome of each testing configuration. The results should therefore be interpreted as reflecting how the methods function in practice, rather than as precise measurements of overall usability effectiveness. While this may limit direct generalization to other contexts, it strengthens the study's ability to reveal method-specific characteristics and interaction patterns.

1.6.5 Single-Moderator Considerations

As the study examines how the moderator adapts across methods and shapes the participant-moderator interaction, all sessions were conducted by a single moderator. The findings therefore reflect a specific moderation approach and should be understood within this context. At the same time, using a single moderator ensured consistency across methods, allowing differences in interaction to be more directly attributed to method conditions rather than variation in moderation style.

1.7 Relevance for Sustainable Development

This project relates to sustainable development through its exploration of alternative usability testing methods that may reduce the need for physical co-presence. In particular, remote testing has the potential to lower environmental impact by reducing travel and associated resource use, which relates to global sustainability goals such as Climate Action (Goal 13) and Responsible Consumption and Production (Goal 12) [13].

From a social sustainability perspective, remote methods may also enable broader participation by allowing users from different geographical contexts to take part in usability evaluations. This can support more inclusive design processes, relating to Reduced Inequalities (Goal 10) [13].

At the same time, the project considers potential trade-offs associated with these methods, particularly in how interaction, communication, and user experience may be affected when testing is conducted remotely or through mediated interaction.

1.8 Ethical Considerations

Ethical considerations were taken into account throughout the planning of the study, particularly in relation to participant integrity, informed consent, and data handling. Participants were informed about the purpose of the study and their right to withdraw at any time.

The study also involves methodological ethical considerations. In proxy-based interaction, the moderator performs actions on behalf of the participant, which introduces a risk of influencing both the interaction and the outcomes.

In remote testing contexts, reduced visibility of participants affects how comfort, understanding, and engagement can be assessed during the session. This was addressed in the study design, particularly in how the moderator supports communication and participant well-being.

1.9 Outline of the Thesis

The thesis is structured around three main phases: designing the test plan, conducting the user tests, and analyzing the collected data.

The first phase is presented in Chapter 2. Theory & Literature Study, and Chapter 3. Test Design. It begins with a review of relevant literature on usability testing methods and remote testing. Insights from the literature informed the design of the test plan, ensuring that the selected methods and data collection approaches were aligned with the research questions.

The second phase consisted of conducting the user tests according to the defined test plan. During this phase, all test sessions were documented, and qualitative and quantitative data were collected.

The third phase focuses on the analysis of the collected data and is presented in Chapter 4. Data Collection & Analysis. In this phase, the data were systematically analyzed in relation to the research questions, enabling a comparison of the different testing methods. The outcomes of this analysis are presented in Chapter 5. Results.

Parts of the analysis were conducted alongside the data collection in the second phase for practical reasons, such as managing participant scheduling and coordinating test logistics. However, no adjustments were made to the test setup, and all sessions followed the same predefined procedure to ensure comparability between the test methods.

Chapter 6. Discussion interprets the findings in relation to the research questions and previous work, and reflects on methodological implications and limitations. Chapter 7. Takeaways for Practitioners summarizes practical implications of the findings and provides recommendations for usability testing in contexts involving remote or proxy-based interaction. Finally, the thesis is concluded in Chapter 8. Conclusion, which summarizes the main findings and contributions of the study.

2. Theory & Literature study

This chapter presents the theoretical and methodological foundations of the study. It introduces key perspectives from cognitive science and human–computer interaction that are used to understand how interaction, cognition, and communication take place in usability testing. In addition, it reviews methodological approaches related to usability evaluation, including think-aloud methods, comparison of test methods, and remote testing.

2.1 Theoretical Background

This section presents the theoretical perspectives that inform the analysis of the study. The frameworks are used to understand how cognition and interaction unfold in usability testing situations where interaction with the user interface is shared between participants and moderators. Together, these perspectives provide a lens for examining how thinking, action, and communication are shaped by social, material, and bodily factors.

2.1.1 Distributed Cognition

Distributed cognition describes cognition as something that happens across a system, involving both people and the tools or artefacts in their environment, rather than only inside a single individual [3, 2]. In this view, cognitive processes are shaped by the relationships between different elements in the system, not by where those elements are physically located. This perspective emphasizes how external representations and interactions with artefacts can actively support and shape cognitive processes [14].

In this study, distributed cognition is used to understand usability testing situations where the physical interaction with the user interface is shared between more than one person. The participant makes decisions and gives instructions, while the moderator performs the actions on the user interface. From this perspective, decision-making, instruction, and execution can be seen as parts of the same cognitive process, even though they are carried out by different people. This aligns with prior work showing that cognitive processes can be distributed across individuals and coordinated through interaction [2, 3].

The concepts of intersubjectivity and presence provide additional perspectives on how shared understanding is created in such situations [15, 16]. Intersubjectivity refers to how people develop a shared understanding through interaction, while presence describes the feeling of “being there” in a mediated environment. These concepts support the idea that coordinated thinking and

understanding can still occur even when interaction is distributed across participants and mediated through technology.

2.1.2 Situated Cognition

Situated cognition is a perspective in cognitive science that emphasizes that thinking and learning always take place within a specific context [4, 5]. Rather than viewing cognition as something that happens only inside an individual, this perspective highlights how understanding develops through interaction with other people, tools, and the surrounding environment. Knowledge is therefore seen as something that is constructed through participation in social and practical activities, rather than something abstract and independent of context [5].

From this perspective, understanding and decision-making are closely tied to the situation in which they occur. Factors such as social interaction, available tools, and the physical and social setting influence how people interpret information and act. This means that the same task may be understood and performed differently depending on the context in which it takes place.

In the context of usability testing, this perspective is particularly relevant. The testing environment, the presence of a moderator, and the form of interaction all shape how participants think, communicate, and perform tasks. Changes in co-presence or interaction setup may therefore influence not only user behaviour, but also how usability problems are revealed and interpreted.

2.1.3 Embodied Interaction/Embodiment

Embodied interaction highlights how cognition and interaction with technology are shaped by the human body and its engagement with the physical environment [6]. Rather than viewing understanding as purely mental, this perspective emphasizes the role of physical actions, bodily presence, and direct interaction in how users make sense of a system [6]. It is closely related to distributed and situated perspectives, which similarly stress that cognition emerges through interaction between individuals, artefacts, and their surrounding context.

From this perspective, interaction with a system is closely connected to how users move, touch, and physically engage with user interfaces. The body plays a central role in how people perceive, learn, and make sense of a system, as external actions and interactions with artefacts can support cognitive processes by making them more visible, tangible, and easier to coordinate [14].

In the context of this study, embodiment is particularly relevant in situations where users do not interact directly with the user interface but instead instruct a moderator to perform the actions. The absence of direct physical engagement may influence users' sense of control, involvement, and overall experience. At the same time, the moderator's bodily presence, including gestures, posture, timing, and spatial positioning, can shape how participants communicate and behave during the test. From an embodiment perspective, this highlights how both physical action and presence contribute to how interaction is experienced and interpreted.

2.2 Methodological Background

This section presents methodological concepts and approaches that inform the design of the study. The literature is used to describe how usability testing can be conducted, how different test methods can be compared, and how user data can be collected and interpreted. In addition, the section introduces the concept of proxy interaction, which is central to this thesis, as well as considerations related to remote usability testing. Together, these perspectives provide a foundation for understanding the methodological choices made in the study.

2.2.1 Usability Testing & Usability Problems

Usability testing is a widely used method for identifying problems that users encounter when interacting with a system [17, 18]. During a usability test, participants perform representative tasks while researchers observe their actions, comments, and difficulties. The goal is to reveal issues that may affect task performance, understanding, or overall user experience.

A usability problem can be understood as a situation in which the user struggles to achieve a goal, makes errors, or expresses confusion or dissatisfaction when interacting with the system. According to Nielsen and Landauer, identifying a usability problem depends on two conditions: the user must experience the problem, and the evaluator must recognize that the problem has occurred [19]. This highlights that usability problems are not only about user performance, but also about the interpretation and observation made by the evaluator. This issue is further supported by research on the evaluator effect, which shows that different evaluators may identify different usability problems even when observing the same session [20].

In comparative usability studies, the identification of usability problems is often used as a key measure when evaluating different test methods [11, 21]. However, the number and type of problems found may also be influenced by factors such as the test setup, the interaction format, and the role of the moderator. Therefore, usability problems should be interpreted not only as properties of the user interface, but also as outcomes shaped by the testing method itself.

2.2.2 Comparing Test Methods

Previous research has explored how different usability testing methods can be compared in a systematic way. One common approach is to evaluate methods based on measurable outcomes, such as the number of usability problems identified and the time required for analysis [22]. For example, Bruun et al. [21] compared several testing conditions using a traditional in-person usability test as the baseline, and evaluated other methods in relation to this reference method.

Another comparative study by Van den Haak et al. [23] examined differences between usability testing methods using think-aloud protocols and constructive interaction. Similar comparisons of usability testing approaches have also been explored in other studies of remote usability testing [11].

Their comparison focused on several key aspects:

- the number and types of usability problems detected,
- the relevance of the identified usability problems,
- task performance, and
- participants' experiences during the test.

These criteria provide a useful foundation for comparing usability testing methods, as they capture both objective outcomes, such as problem detection and task performance, and subjective aspects, such as participant experience. At the same time, previous research has shown that such measures may be influenced by the testing method itself, including factors such as interaction format and evaluator interpretation [20]. In the context of this thesis, similar dimensions are considered when comparing the different test methods.

2.2.3 Think-Aloud Methods

The think-aloud method originates from cognitive psychology and belongs to a group of techniques known as verbal report methods [7]. In usability testing, this method is used to capture users' thoughts while they interact with a system, which makes it useful for understanding their reasoning, expectations, and difficulties during task performance [22, 24].

In general, the method consists of two main stages: first, collecting think-aloud protocols in a systematic way, and second, analyzing these protocols to understand the cognitive processes that occur while a person solves a problem [7]. The protocols are gathered by asking participants to perform tasks while continuously verbalizing what they think, see, and intend to do.

Two common variants of the think-aloud method are concurrent and retrospective verbalization. In concurrent think-aloud, participants express their thoughts during the interaction. In retrospective think-aloud, participants instead reflect on their thoughts after completing the task [7]. Concurrent verbalization is often preferred in usability testing because it provides more immediate insight into the user's decision-making process, although it may also influence task performance [24].

The classic concurrent think-aloud approach is based on the framework proposed by Ericsson and Simon [7]. This framework emphasizes the importance of maintaining natural behavior while collecting verbal data. To ensure reliable results, researchers typically use neutral instruction scripts, provide short practice sessions to help participants get used to thinking aloud, and give simple reminders such as "keep talking" if participants become silent. Apart from these prompts, interaction between the evaluator and the participant should be kept to a minimum to avoid influencing the user's behavior [24].

Research has also examined how participants' verbalizations can be categorized. For example, Van den Haak et al. [25] developed classification labels for different types of communicative acts during usability testing. These categories help structure and interpret think-aloud data by distinguishing between types of comments, such as problem identification, reasoning, or expressions of uncertainty (see Figure 2.1).

TABLE I
CLASSIFICATION LABELS FOR THE COMMUNICATIVE ACTS AND EXAMPLES OF CONSTRUCTIVE INTERACTION

	Communicative Act	Description	Examples from CI Sessions
1	Action	performing an action (e.g. reading a task, navigating on the site, clicking on links, writing down the answer to a task)	The actor enters a search term in the appropriate box on the web site [Team 5, Task 2]
2	Description	describing how a particular action is performed	<i>There was something there..., in [the link] "help," I'll give it a try...</i> [Team 1, Task 4]
3	Suggestion	suggesting an action, to be taken in order to complete a task successfully (not in response to a previously formulated problem: cf. 8)	<i>You could maybe try NOT like "language NOT English"</i> [Team 5, Task 2]
4	Evaluation	evaluating the test object, an action or the outcome thereof	<i>We seem to be doing okay...</i> [Team 3, Task 5]
5	Prediction	predicting what the outcome of a particular action will be	<i>I'm sure we'll be able to select results afterwards</i> [Team 4, Task 4]
6	Agreement	agreeing with an action or communicative act performed by the other Team member	<i>Yeah, I suppose that's true...</i> [Team 2, Task 4]
7	Disagreement	disagreeing with an action or communicative act performed by the other Team member	<i>No, we did do that...</i> [Team 4, Task 4]
8	Problem formulation	verbalizing a problem with the test object and/or action performed	<i>I don't know what OR means</i> [Team 5, Task 2]
9	Solution	suggesting a solution, in response to a problem that has been formulated	<i>Well, OR means "of"</i> [Dutch translation] <i>of course</i> [Team 5, Task 2]
10	Question	asking an informative <i>wh</i> -question (what, where, etc.), i.e. one that expects an informative response (but does not constitute a problem with the test object and/or action performed)	<i>So how many results do we have then?</i> [Team 2, Task 1]
11	Answer	providing an informative response to an informative question	<i>I'd say sixteen</i> [Team 2, Task 1]
12	Other comment	uttering a comment that is not related to the task at hand	<i>I know people who study there</i> [the university which developed the online library catalogue that the teams were testing] [Team 5, Task 5]

Figure 2.1: Table describing different alternatives of communicative acts [25].

2.2.4 Proxy Interaction

In this study, proxy interaction refers to situations where the participant does not directly interact with the user interface, but instead instructs another person (here, the test moderator) to perform the actions on their behalf. This concept is related to previous work on proxy users and mediated interaction, but is used here specifically to describe instruction-based interaction during usability testing.

Traditional think-aloud testing has sometimes been criticized for creating an artificial testing situation, where participants may feel uncomfortable verbalizing their thoughts while acting alone. Pair-based approaches, in contrast, introduce a more conversational and collaborative setting [25, 24]. In such contexts, speaking and reasoning aloud may feel more natural, as communication becomes part of a shared activity. This characteristic supports the use of proxy interaction, since the participant both instructs and reflects on their actions in cooperation with the moderator. This aligns with research showing that cognitive processes can be distributed across individuals and supported through interaction [3].

However, involving a facilitator in the interaction also introduces potential risks. As noted by Krahmer and Ummelen, the presence of an expert or facilitator may unintentionally influence the test results, even if verbal guidance is kept to a minimum [26]. The moderator's prior knowledge of the system may affect how instructions are interpreted and executed, which in turn may shape

the participant's experience and the observed usability issues. This can be related to the evaluator effect, where different interpretations and actions may lead to variations in identified usability problems [20].

In this thesis, proxy interaction is interpreted through the perspectives of distributed and situated cognition. Interaction is understood as a shared cognitive system involving the participant, the moderator, and technological artefacts [3, 2], while the specific testing context, such as mediated communication and remote observation, shapes how participants understand tasks and make decisions [4, 5].

2.2.5 Remote Testing

Remote usability testing refers to evaluation situations in which evaluators and users are separated in space and/or time [27]. Two common forms of remote testing are synchronous methods, where the evaluator and participant interact in real time, and asynchronous methods, where feedback is provided at a later time.

Asynchronous remote testing often involves techniques such as diary studies, forum feedback, or written reports submitted after the interaction [21]. Previous studies suggest that these methods can be time-efficient for evaluators, both in terms of conducting the test and analyzing the collected data [21]. However, they may require more time and effort from participants and provide less control over how the test is carried out. This can make recruitment more difficult and increase the risk that participants discontinue the test if they encounter too many obstacles [11].

Synchronous remote testing has most commonly been studied in the context of web user interfaces, where screen sharing allows evaluators to observe user interactions directly [11]. In such setups, the test situation resembles a traditional laboratory-based think-aloud test, with the main difference being the physical separation between evaluator and participant. Research highlights several advantages: it can be more cost-efficient and allows for a more diverse participant sample, since users can participate from different locations [11, 21]. Some studies have also found that remote synchronous tests can identify a similar number of usability problems as in-person tests, and in some cases even more [11].

At the same time, remote testing introduces several challenges. Building rapport and trust with participants may be more difficult at a distance, and technical issues can disrupt the session and affect data quality [11]. In addition, evaluators may miss important nonverbal and paraverbal cues, such as body language, tone of voice, or subtle reactions, which can lead to misinterpretation of user behavior.

Remote testing can make it possible to include participants from different geographical contexts, which has been identified as an advantage in previous studies [11, 21]. However, the lack of physical co-presence may make it more difficult to detect misunderstandings or discomfort during the test.

Previous research also suggests that remote usability testing can be effective under certain conditions, particularly when evaluators are familiar with the system and the testing procedure [11]. Such prior knowledge can help moderators manage the limitations of remote interaction. In the context of this thesis, earlier in-person usability tests of the same product had already been conducted, which supported the feasibility of performing remote test methods in this study.

2.3 Linking Theory to the Research Questions

The theoretical perspectives and methodological literature presented in this chapter did not only inform the analysis, but also shaped the design of the study. The methodological literature provided a foundation for how usability testing is typically conducted and evaluated, particularly in relation to think-aloud protocols, remote testing practices, and assumptions about direct interaction with the user interface. This served as a reference point when designing and adapting the different testing methods.

At the same time, the theoretical perspectives guided how the study approached interaction and what types of data were considered relevant to collect. Rather than being tied to individual research questions, Distributed Cognition, Situated Cognition, and Embodied Interaction offered complementary lenses for framing the study and identifying what aspects of interaction, communication, and context to focus on during data collection.

Together, these perspectives informed both the design of the methods and the analytical focus of the study. Distributed Cognition was particularly relevant for structuring and examining proxy-based interaction, where actions and decisions are shared between participant and moderator. Situated Cognition guided attention towards how context, such as co-presence and testing environment, shapes interpretation and behaviour. Embodied Interaction informed the importance of physical engagement and how the presence or absence of direct interaction influences user experience.

In this way, the literature did not only support the interpretation of results, but also influenced how the study was set up and what data was made visible through the chosen methods.

3. Test Design

This chapter presents the design of the usability tests conducted in this study. It describes the different testing configurations used for comparison and outlines how they were structured and implemented. In this thesis, the term test method refers to the different usability testing configurations used for comparison.

3.1 Research Approach & Methods

The study addresses the research questions by conducting and comparing four usability testing methods. The methods vary along two key dimensions: (1) physical co-presence (on-site vs. remote) and (2) interaction execution (direct interaction vs. proxy interaction), forming a 2×2 comparison design, as illustrated in Table 3.1. The horizontal axis represents co-presence, while the vertical axis represents how interaction is performed.

Table 3.1: The methods and their differences. Horizontal: physical co-presence, vertical: interaction execution.

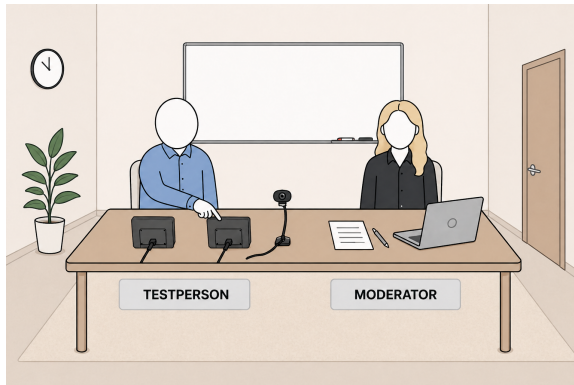
	On-Site	Remote
Direct (Think-Aloud)	Method 1 (On-site think-aloud)	Method 4 (Remote think-aloud)
Proxy	Method 2 (On-site proxy)	Method 3 (Remote proxy)

The four test methods used in the study were:

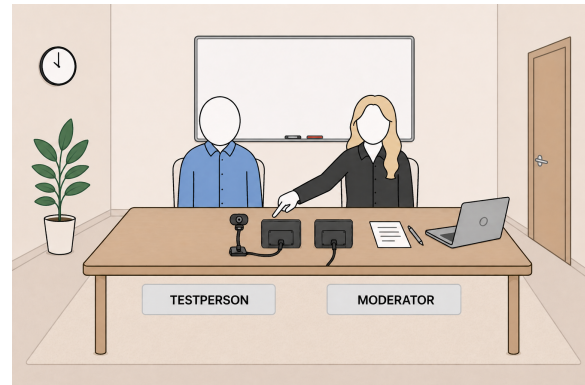
1. In-person moderated usability testing with concurrent think-aloud (On-site think-aloud)
2. In-person assisted think-aloud with proxy interaction (On-site proxy)
3. Remote assisted think-aloud with proxy interaction (Remote proxy)
4. Remote moderated usability testing with concurrent think-aloud (Remote think-aloud)

Methods 1 and 4 involve direct interaction, where participants interact with the user interface themselves while verbalizing their thoughts. In contrast, Methods 2 and 3 involve proxy interaction, where participants instruct the moderator to perform actions on their behalf.

Similarly, Methods 1 and 2 are conducted on-site, with both participant and moderator physically co-located (see Figure 3.1), while Methods 3 and 4 are conducted remotely, with interaction mediated through video communication (see Figure 3.2). Both remote methods are synchronous, meaning that the moderator and participant interact in real time despite being physically separated.

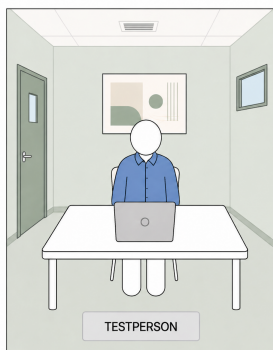


Method 1:
On-site think-aloud.



Method 2:
On-site proxy.

Figure 3.1: Overview of on-site test setups for Method 1 and Method 2.
Illustrations created by ChatGPT.



Method 3:
Remote proxy.



Method 4:
Remote think-aloud.

Figure 3.2: Overview of remote test setups for Method 3 and Method 4.
Illustrations created by ChatGPT.

The order of the test methods corresponds to the order in which they were conducted and was determined by logistical considerations rather than experimental progression.

Method 1 represents a traditional usability testing setup and serves as the baseline for comparison with the other methods. Apart from the differences in co-presence and interaction execution, all other aspects of the usability tests were kept consistent across methods, including the introduction, task set, and debriefing procedure.

Each method was conducted with five participants, resulting in a total of twenty participants, excluding pilot sessions. This sample size is consistent with previous research suggesting that a small number of participants can identify a large proportion of usability problems, a principle widely adopted in usability practice [19, 28].

3.2 User Profile & Participants

A target user profile was defined prior to the study, representing a retail sales associate responsible for opening and closing a store and managing daily interactions with the door control panel. The scenarios and tasks used in the usability tests were designed to reflect this work context and typical responsibilities associated with the role.

Due to organizational constraints, participants were recruited internally from different sales organizations within the company rather than from the intended end-user group. This decision was made as a request by management, as additional usability tests with actual end users were planned after the completion of this study. During the sessions, participants were therefore asked to act according to the defined persona and follow the provided scenarios, even though their professional backgrounds and prior product knowledge might differ from those of the intended users. This may introduce bias, as differences between the participant sample and the target user group can affect the validity and interpretation of usability findings [17, 29]. All participants upon signing up had confirmed that they had not interacted with the product prior to test session.

The study aimed to include five participants per test method, resulting in a total of 20 participants. However, due to a technical malfunction in Method 2, where audio was not recorded for one session, this participant's data was excluded from the observational analysis and only included in the debriefing analysis.

Participants were recruited from Sweden, Denmark, the Netherlands, and Switzerland, with each country assigned to a specific test method. Although cultural and language differences may influence user behavior and communication, participants shared a similar professional background within the company, supporting comparability between groups. All sessions were conducted in English to ensure consistency, although clarifications in Swedish could be provided when necessary.

An overview of participants across the four test methods is presented in Table 3.2.

Table 3.2: Overview of participants across the four test methods. *One participant in Method 2 was excluded from observational analysis due to missing audio data.

Method	Participants	Countries	Age span
Method 1	5	Sweden	25–58
Method 2	5*	Sweden / Denmark	24–35
Method 3	5	The Netherlands	40–51
Method 4	5	Switzerland	40–60

The participant group consisted of 19 men and 1 woman, reflecting the available participant pool. Additional female participants were initially recruited but could not participate due to illness and logistical constraints.

3.3 Test Setup & Equipment

The test setup consisted of a control panel mounted on a flexible prototype placed on a table, allowing participants to interact with the panel in a stable position during the usability tests (see Figure 3.3).

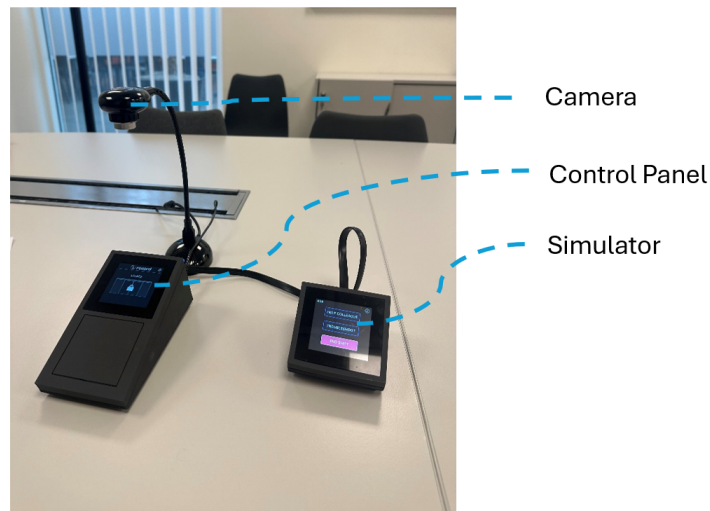


Figure 3.3: The equipment used across all test methods.

The test sessions followed a Wizard-of-Oz approach, where system responses were manually controlled through a simulator rather than generated by a fully implemented system [30]. The simulator had similar screen hardware as the control panel but was mounted on a smaller box to distinguish it from the main user interface. Participants were not informed about the simulator's specific functions, enabling controlled and consistent system behavior while preserving natural interaction.

The simulator was placed adjacent to the control panel to allow efficient triggering of system responses and error states. Both devices were connected via cable and powered through a standard electrical outlet.

A camera was positioned behind the control panel to capture both screen content and interaction input. The video and audio feed were streamed and recorded through a Microsoft Teams meeting, enabling remote observation and subsequent analysis. However, only hand and screen interactions were recorded, meaning that facial expressions and body language were not captured.

The same equipment was used across all test methods, but the physical setup and communication conditions varied depending on the testing configuration. In on-site methods (Methods 1 and 2), sessions were conducted in private meeting rooms in Sweden and Denmark (see Figure 3.4.

In these settings, participant and moderator were physically co-located and could see each other throughout the session, allowing for full access to verbal and nonverbal communication.

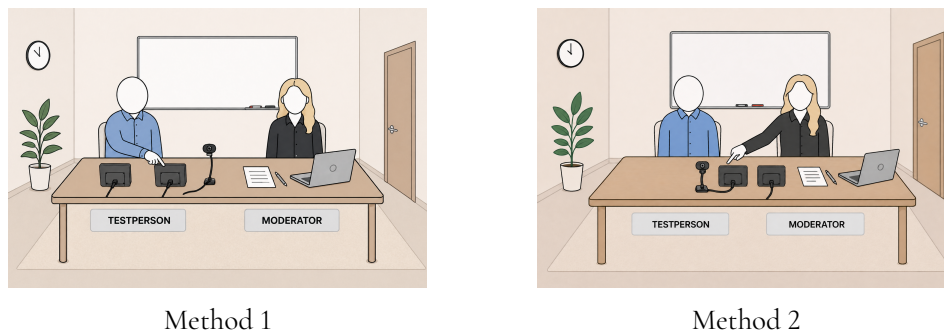


Figure 3.4: Method 1 (On-site think-aloud) and Method 2 (On-site proxy).

In the remote proxy method (Method 3), the moderator remained in Sweden with the full test setup, while participants joined remotely from their homes or offices (see Figure 3.5). Due to technical constraints in Microsoft Teams, it was not possible to display multiple camera feeds simultaneously. As a result, the camera focused on the control panel during task execution, meaning that participants could not see the moderator while performing tasks. Participants were given the option to keep their own camera on or off since the task-execution of the session was recorded and prior to the test they had only been informed of recording their hands and not face; four chose to keep it on, while one turned it off. Cameras were typically switched during the introduction and debriefing, allowing moments of face-to-face interaction.

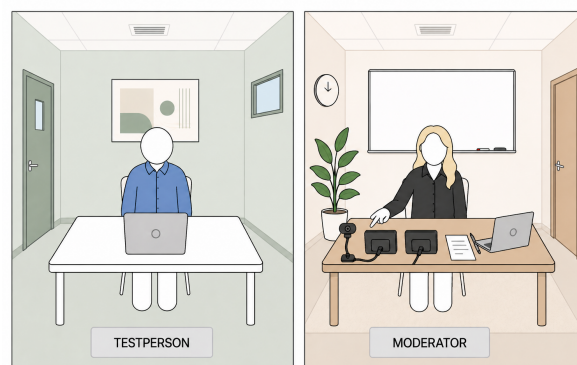


Figure 3.5: Method 3: Remote proxy.

In the remote think-aloud method (Method 4), the setup was distributed, with the moderator located in Sweden and the test equipment placed in Switzerland (see Figure 3.6). Similar technical limitations applied, preventing simultaneous visibility of both participant and the setup during task execution. As a result, the moderator could not see the participants face during the task performance, while the participant could see the moderator. While switching cameras was encouraged between phases of the session, its use varied due to practical constraints.

Across both remote methods, the inability to maintain continuous visual contact reduced access to nonverbal cues such as facial expressions, gaze, and body language. Although moments of

visual contact were present during introductions and debriefings, the visibility conditions differed substantially from the on-site methods, where full co-presence was maintained throughout.

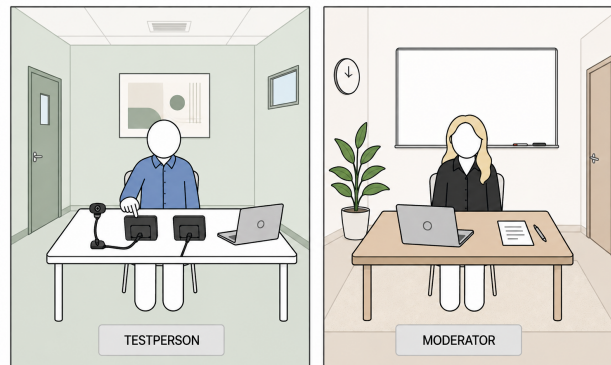


Figure 3.6: Method 4: Remote think-aloud.

Overall, the setup was designed to maintain consistency in equipment and procedure while accommodating the practical constraints of remote and distributed usability testing. However, differences in visual access between participant and moderator represent an important variation between the test methods.

3.4 Pilot Study

Prior to the main study, pilot tests were conducted for each testing method to ensure that the equipment functioned properly, that the tasks were clearly understood, and that the sessions fit within the intended time frame, as recommended in usability testing practice [17].

Feedback from pilot participants indicated that the introduction was too long and felt somewhat unnatural. Participants suggested adopting a more conversational approach to make the sessions feel less scripted. Additionally, the pilot revealed that it was cognitively demanding for the moderator to simultaneously manage proxy interaction, read instructions, and guide the session.

As a result, the test script was revised to create a more seamless and manageable flow. A more conversational introduction was introduced, allowing participants to engage more actively by responding to questions. Additionally, specific sections were added to the script for the moderator to note time stamps and indicate whether tasks were successfully completed. These additions supported the moderator in keeping track of the session in real time and made it easier to resume the script after temporarily acting as a proxy, ensuring a smoother transition back to task facilitation.

The pilot study identified several technical issues related to the testing equipment, which were traced to a hardware problem with the control panel. Replacing the unit with another control panel using the same configuration resolved these issues.

The pilot study also revealed system feedback issues that could not be addressed within the available time. As a result, one task was removed and replaced with two modified tasks.

No data from the pilot sessions was included in the final analysis, as the primary purpose of the pilot study was to refine the testing procedure. The pilot helped identify key procedural and technical issues, ensuring that the main study could be conducted with fewer disruptions and a more reliable setup.

3.5 Test Procedure

Each test session followed a structured three-phase process consisting of:

1. Preparation
2. Execution of the usability test with the participant
3. Post-session reflection by the moderator

This structure ensured consistency across sessions while capturing both participant behavior and moderator insights.

Each session lasted approximately 1.5–2 hours, of which about one hour was dedicated to the usability test itself. Additional time was allocated before and after each session for preparation and reflection.

All sessions were conducted in English to ensure consistency across participants and test methods. However, the moderator's reflections were documented in Swedish (the moderator's native language) to allow for more detailed and nuanced notes. Clarifications in Swedish could be provided when necessary without altering the procedure.

A mixed-method approach was used to collect both qualitative and quantitative data, supporting methodological triangulation [18]. Data were collected through the usability test, participant debriefing, and moderator reflection.

3.5.1 Phase 1: Preparation

Approximately 30 minutes were allocated before each session for preparation. This included setting up the equipment, ensuring that the control panel and simulator were configured according to predefined settings, and connecting all devices to power outlets and the recording system.

Some preparations differed slightly between test methods. For the remote think-aloud method (Method 4), the setup was carried out in collaboration with a test assistant located in Switzerland.

3.5.2 Phase 2: Usability Test Execution

3.5.2.1 Introduction

The session began with a structured introduction lasting approximately 10 minutes. The moderator welcomed the participant, assisted with camera setup if needed, and explained the purpose, format, and expectations of the test.

Particular emphasis was placed on explaining how to perform the think-aloud method. In the proxy interaction conditions (Methods 2 and 3), participants were also instructed on how to provide clear and actionable instructions to the moderator, who executed the interactions on their behalf.

For the remote think-aloud method (Method 4), the initial welcoming and technical setup were primarily handled by a test assistant located in Switzerland. The assistant supported participants in connecting to the Microsoft Teams meeting and ensured that the equipment was properly configured. The assistant then left the meeting room, after which the moderator proceeded with the introduction.

3.5.2.2 Task Performance

The task phase lasted approximately 35 minutes. Participants were first presented with a scenario based on the defined persona, describing a shift supervisor preparing a store for opening. This provided contextual grounding for the tasks.

As an initial step, participants were asked to describe what they could interpret from the control panel's initial screen. This served both as a warm-up for the think-aloud method and as a first impression of the user interface.

In total, eleven tasks were included, each reflecting typical responsibilities associated with the use of the control panel (see Table 3.3). Detailed descriptions of all tasks and scenarios are provided in Appendix A. Task 2.5 was labeled as such following revisions based on pilot study feedback; as the task numbering had already been implemented in the simulator, the existing numbering could not be altered.

Depending on the test method, tasks were completed either through direct interaction or through proxy interaction, where the moderator executed actions based on participant instructions.

Each task had a predefined goal. If a participant did not reach the goal, a hint was provided. If the participant still did not succeed, the correct solution was revealed either immediately or during the debriefing, depending on what best preserved the flow of the test.

After selected tasks, reflective questions were asked to capture participants' reasoning, perceptions, and understanding. This created a semi-structured approach in which tasks were fixed, while reflective questions enabled deeper insight into participants' reasoning, interaction behaviour, and perceived experience.

Table 3.3: Overview of usability test tasks and their success criteria (SCC).

#	Task	SCC
1	Enter system + change mode to auto (no PIN)	Mode set to auto
2	Change time-out (Admin code required)	Time-out set to 20 seconds
2.5	Turn off winter mode	Winter mode turned off
3	Close the door (no PIN)	Mode set to locked
4	Open door with authenticated trigger	Trigger used intentionally
5	Enable user PIN	PIN created with four digits
6	Change mode to auto (PIN) → open (no PIN)	Mode changed to auto and then open
7	Enter wrong PIN repeatedly	User notices PIN does not work
8	Navigate through issues	Error messages located
9	Find contact information	User identifies QR code
10	Understand issues	Stating most severe issue

3.5.2.3 Participant Debriefing

Following the task phase, a participant debriefing was conducted, lasting approximately 15 minutes. The debriefing consisted of six questions, including the System Usability Scale (SUS) and additional statements rated on a similar scale. The use of SUS enabled comparison with previous usability tests of the same product.

The debriefing combined open-ended and scaled questions, and all responses were documented by the participant. The full set of questions is provided in Appendix B.

After the debriefing, the moderator thanked the participant and wrapped up the test session. The participant then left the meeting or testing environment.

3.5.3 Phase 3: Moderator Reflection

After each session, the moderator conducted a structured debriefing lasting approximately 30 minutes. This phase focused on capturing reflections on the test execution, including communication quality, technical challenges, and interaction dynamics. The full set of questions is provided in Appendix C.

The reflections were documented using a predefined debriefing form aligned with the research

questions and theoretical framework of the study. In addition, a diary was maintained to capture more detailed observations, including notable events, interpretations, and reflections on participant behavior and interaction patterns.

Together, the structured form and diary provided a comprehensive account of the moderator's perspective, complementing the participant data and supporting a richer analysis of the test sessions.

3.6 Test Moderator Role

The test moderator played a central role in planning, executing, and documenting all test sessions, as well as in the subsequent data analysis. Before each session, the moderator ensured that the equipment was functioning correctly and configured according to the predefined conditions, either independently or in collaboration with a colleague in Switzerland. During the sessions, the moderator focused on creating a comfortable atmosphere for the participant while ensuring that the procedure followed the planned protocol. After the sessions, the moderator was responsible for organizing and analysing the collected data.

The moderator's practical responsibilities varied slightly between test methods. In the in-person tests, this included preparing the camera setup, managing recordings, and providing debriefing materials. In the remote tests, additional tasks involved managing the meeting setup, guiding participants in adjusting the camera when needed, and distributing digital debriefing forms. Across all methods, the moderator instructed participants to think aloud and, in the proxy interaction methods, also explained how to provide instructions for interaction. The moderator observed participant behavior and verbalizations, and documented task completion and notable comments using a structured protocol.

From a theoretical perspective, the moderator can be seen as part of the testing situation rather than a neutral observer. According to situated cognition, the moderator's presence and actions influence how participants communicate, how the session progresses, and how tasks are understood.

In proxy interaction, this becomes even more important, as the moderator carries out the participant's instructions and becomes part of the interaction. From an embodiment perspective, small cues such as timing and physical presence may also affect how participants behave and communicate, even when verbal guidance is limited.

At the same time, it was important that the moderator followed the predefined test procedure consistently, particularly during the introduction, task instructions, and debriefing. This ensured that any differences between sessions were primarily due to the test methods rather than variations in how the sessions were conducted.

However, the moderator could not remain entirely neutral, as the role required adapting to the specific conditions of each method and responding to participants' needs. The moderator therefore balanced procedural consistency with situational adaptation, ensuring comparable test conditions while maintaining clear communication and mutual understanding during the sessions.

4. Data Collection & Analysis

This chapter presents the data collection and analysis process used in the study. It outlines the different data sources, including observations, participant feedback, and notes gathered during the usability testing sessions. Furthermore, it describes the analytical framework that guided the study and the procedures applied to systematically organize, interpret, and compare findings across the four usability testing methods.

The chapter also details how the collected data were processed and structured to enable identification of patterns, recurring issues, and method-specific insights. Through this process, a set of overarching themes emerged, which form the foundation for the results presented in the following chapter.

4.1 Data Collection

Data were collected from multiple sources to capture both observable behavior and subjective experiences. These included:

- notes taken during and immediately after each test session,
- video and audio recordings of interactions with the control panel,
- participant responses from the debriefing questionnaire,
- the moderator's responses from the debriefing form, as well as entries from a reflective diary.

Together, these sources enabled analysis from both the participant's and the moderator's perspectives. Table 4.1 provides an overview of the collected data and how they were categorized in the analysis.

The dataset consisted of both qualitative and quantitative data. The qualitative data included participants' verbalizations during think-aloud sessions, instructions given during proxy interaction, comments expressed during the test, and open-ended responses from the debriefing. These data were used to identify usability problems, misunderstandings, emotional reactions, and patterns in how participants approached tasks.

The quantitative data consisted of performance-related and preference-related measures. Performance measures included the number of tasks completed without assistance, and the number of identified usability issues. Preference measures included scaled responses from the debriefing questions.

Table 4.1: Overview of collected qualitative and quantitative data used in the analysis.

Data Type	Category	Description
Qualitative Data	Preference-Oriented Data	Emotional responses; subjective opinions shared during the test; comments from debriefing questions
Qualitative Data	Performance-Oriented Data	Observations of task execution; identified usability issues; think-aloud comments and/or instructions reflecting actions or thought processes related to task performance
Quantitative Data	Preference-Oriented Metrics	Scaled responses from debriefing questions
Quantitative Data	Performance-Oriented Metrics	Number of tasks completed without assistance; number of usability issues identified

4.2 Analytical Approach

The analysis in this study was guided by the theoretical perspectives presented earlier, which were used as conceptual lenses rather than strict analytical models. Distributed Cognition, Situated Cognition, and Embodied Interaction informed the interpretation of observed phenomena across the different test methods, particularly in relation to proxy interaction. These perspectives supported the analysis of how cognitive work was distributed between participant and moderator, how the testing context influenced interaction behaviour, and how perceived experience was shaped by mediated interaction.

The collected data were analyzed using complementary approaches. The qualitative data were primarily analyzed using thematic analysis combined with affinity diagramming, enabling the identification of recurring patterns across participants, tasks, and test methods. The quantitative data were analyzed descriptively to support comparisons between methods.

The analysis was conducted in two stages. First, each test method was analyzed individually to identify method-specific patterns. Second, a cross-method comparison was performed as the primary focus of the study, enabling a systematic evaluation of similarities and differences between methods.

The analytical approach was closely aligned with the research questions. Specific metrics and indicators were selected to address each research question. For example, Research Question 2, concerning proxy interaction, interaction behaviour, and perceived experience, was examined through indicators such as differences between instruction-based interaction and direct think-aloud interaction, and indications of how participants interpreted and communicated their understanding of the system during the session.

In addition to the methodological comparison, an overall impression of the product was compiled as a natural outcome of the analysis. This product-focused evaluation was considered a secondary outcome.

4.3 Analysis Procedures

This section describes the step-by-step procedures used to analyze the collected data. The analysis was structured differently for qualitative and quantitative data, as well as for participant and moderator inputs.

The qualitative analysis focused on identifying usability problems and patterns in participant behavior, while the quantitative analysis supported comparisons between methods through performance and preference measures. In addition, participant and moderator debriefings were analyzed separately to capture both user perspectives and reflections on the test execution.

4.3.1 Qualitative Analysis – Usability Problems

The qualitative data formed the primary basis for the analysis and were examined using a combination of thematic analysis and affinity diagramming [31], enabling the identification of recurring themes across participants, tasks, and test methods.

The thematic analysis was initially conducted for Method 1 (On-site think-aloud), following established guidelines for identifying patterns within qualitative data [31]. This analysis resulted in a structured analytical framework, which was subsequently applied to Methods 2–4 to support systematic comparison across all test methods.

This approach introduced a potential risk of bias, as analyzing the methods in a fixed sequence may have influenced the identification of usability problems. To mitigate this risk, additional data sources, including task success measures and debriefing responses, were used to support and complement the findings.

In total, 19 video and audio recordings were transcribed and reviewed to identify usability issues, notable comments, and patterns in participant behavior. The recordings ranged from 19 to 42 minutes and focused on the task performance phase of the test sessions.

For Method 1, which involved a more thorough analysis, each test recording required approximately six hours for transcript editing, annotation, review, and documentation of comments in a structured spreadsheet. Methods 2–4 followed a more time-efficient approach, requiring approximately three to four hours per test, depending on the session and the richness of participant comments. Among these, Method 3 (remote proxy) required slightly more time due to a higher volume of reflective comments from participants, whereas Method 4 (remote think-aloud) was generally at the lower end in terms of time required.

The remote methods (3 and 4) also required less time for transcription, as the audio recordings captured two input sources, reducing the need for manual clarification during the transcription process.

Based on the initial set of comments, approximately 130 merged comments were generated and transferred to Miro as individual post-its. Each post-it represented a summary of related observations, including a brief description and an indication of frequency, and was organized according to the 23 groups. Through this process, broader patterns were identified and subsequently clustered into themes.

In total, 27 themes were generated. Of these, four related to the testing procedure and the moderator's role, while the remaining 23 concerned the product itself (see Table 4.2). These themes later served as the foundation for the analysis of Methods 2–4.

Table 4.2: Overview of themes generated from Method 1 (on-site think-aloud).

General	PIN/Login
• Users click on icons that are not buttons • Feedback behavior creates confusion about possible actions • Poor affordance • Confusing door modes and when to use them • Inconsistent design • Mental model – authentication trigger • Mental model – reset • Mental model – terminology • Mental model – winter mode • Natural path for exploring error codes • Unclear icons and terminology • Lack of error prevention • Clear signifiers, feedback, and visual cues • Recommendations • Intuitive settings • Difficult to press buttons • Clear icons and terminology	• Difficulty creating a user PIN, leading to issues in usage • Unclear when the user is logged in and how to log in • Unclear when a PIN is required • Unclear what happens after “Too many wrong attempts” • Unclear what “User PIN” means • Unclear which PIN should be used for what purpose
	Methodological Observations • Behavior patterns and think-aloud • Lack of communication • Results may be influenced by the test design • Influence of the test leader

The post-its were color-coded based on the nature of the findings: negative (red), positive (green), mixed (orange), or neutral (blue). This supported further refinement and interpretation of the themes.

Based on the 23 product-related themes, the post-its were regrouped into specific usability problems. See Figure 4.2 for an overview of the grouped problems in Miro.

Method 1



Figure 4.2: Overview of the Miro board with regrouped post-its for Method 1 (on-site think-aloud). Each post-it represents a comment or observation, including information about the number of participants, associated task, and frequency of occurrence. The post-its were color-coded based on the nature of the findings: negative (red), positive (green), mixed (orange), or neutral (blue).

The usability problems were documented in a structured table, where each issue was assigned a severity rating and a brief motivation based on predefined criteria (see Table 4.3). The ratings were informed by established usability evaluation principles [32]. The color-coding from the previous step supported this assessment by providing an overview of the relative impact of each issue.

The complete list of usability problems are presented in Appendix D, while a summary is presented in Chapter 5. Results.

Table 4.3: Severity rating scale for usability problems.

Level	Severity	Description
4	Critical	User cannot complete the task; blocks main flow; no reasonable workaround
3	Major	Task completed with difficulty; causes confusion; requires extra steps or guesswork
2	Minor	Annoying but manageable; affects efficiency; easy to work around
1	Cosmetic	No impact on task completion; minor visual or textual issues

4.3.1.2 Methods 2–4

For Methods 2–4, the previously identified 27 themes (see Table 4.2) served as the analytical foundation, reducing both time consumption and repetitive work.

The methods were analyzed separately and in sequence. Once the identification of usability problems and the assignment of severity had been completed for one method, the process proceeded to the next. The same procedure was applied consistently across all methods, beginning with transcription and followed by the subsequent analysis steps.

For each method, the initial transcription primarily served to familiarize the researcher with the data. The video recordings were then analyzed, and observations were documented directly in a spreadsheet structured around the 27 predefined themes. In contrast to Method 1, where all relevant observations were recorded, this stage focused specifically on the predefined themes, enabling a more efficient and targeted analysis.

Within the spreadsheet, comments were documented for each participant, specifying relevant actions or verbal expressions associated with each theme, as well as the corresponding task.

After completing the analysis for each participant in a method, the data were transferred to Miro. The post-its were color-coded based on participants and then broken down into smaller comments, which were grouped into specific usability problems, following the same approach as in Method 1.

However, the color-coding strategy differed between the methods. In this case, post-its were assigned colors based on participants, whereas in Method 1 they were color-coded according to severity. This structural difference reflects how the methods were organized: Method 1 was less tied to predefined themes initially, and color-coding by severity helped highlight where the most critical issues occurred.

In contrast, Methods 2–4 were structured around specific themes from the outset. To avoid re-examining each round to track occurrences, comments were instead organized by topics within each theme, and frequency was determined by counting the number of post-its. See Figure 4.3 for overview of Method 3. (Overview of Method 2 and 4 are included in the Appendix E: Figures E.1 and E.2).

The identified usability problems were compared against the existing list. If a problem had already been identified, additional observations were incorporated into the corresponding entry, along with the brief motivation next to the severity score. If the problem was new, it was added as a separate entry. In this way, the list of usability problems was iteratively expanded throughout the analysis of each method.

Severity ratings were assigned based on the data from each individual method. As a result, a single usability problem could receive different severity ratings across methods (e.g., critical in one method and minor in another), depending on the observed data (see Results for the number of usability problems identified per method).

Method 3



Figure 4.3: Overview of the Miro board with regrouped post-its for Method 3 (remote proxy). Each post-it represents a comment or observation from a specific participant and task. The post-its are color-coded based on participant identity.

4.3.2 Qualitative Analysis – Moderator Debriefing

The moderator debriefing responses were analyzed using a structured approach inspired by thematic analysis and established practices in UX research [31, 33].

In total, the moderator debriefing responses comprised approximately 35 pages of full-text data. In addition, the reflective diary consisted of 29 pages. The diary primarily functioned as a complementary source, supporting the debriefing responses and helping to address potential gaps.

The responses were reviewed question by question, and answers from participants within the same group were synthesized into group-level descriptions. This process aimed to capture recurring patterns and overall impressions within each group, resulting in approximately 25 pages of condensed descriptive summaries.

The group-level summaries were subsequently compared across groups to identify similarities and differences in user experiences and perceptions. Each group was treated as a separate case, and the comparison can therefore be understood as a form of cross-case analysis [29].

These steps were supported by generative AI to increase efficiency in organizing and condensing the data. All AI-assisted outputs were reviewed and verified to ensure consistency with the original material.

Step 1: Within-method analysis

The analysis began by reviewing the responses question by question within each method (eight questions in total). For each question, the answers from all five participants within the method were read and compared. Some questions were only answered in certain methods, depending on the test configuration and whether the question was applicable.

Based on this, concise summaries were developed to capture recurring patterns and common perspectives. This procedure was repeated for all questions within each method. The same process was then applied across all methods, resulting in a set of detailed summaries for each method.

Step 2: Cross-method comparison

Once all methods had been analyzed individually, the summaries were reviewed again and compared across methods. For each question, similarities and differences between methods were identified. Key themes were extracted and organized into a table, and higher-level summaries were formulated for each method.

The analysis resulted in six overarching themes:

- Instructional strategies across methods
- How proxy interaction affects rhythm, cognition, and behavior
- Misunderstandings and coordination problems across methods
- Situations in which participants adapted more to the moderator than to the system
- Moderator influence on participant behavior and problem-solving
- Perceived moderator and environmental influence across groups

Step 3: Theme refinement

In this step, the identified themes were iteratively refined to achieve a more abstract and comparable level of description (see Appendix E Table E.1 - E.5) for first iterations). The focus was on narrowing down patterns and expressing them as concise, descriptive subthemes that could be systematically compared across methods. Table 4.4 presents the finalized set of subthemes.

Step 4: Structuring Results

Finally, tables were created with the subthemes, providing concise descriptions of how each method related to the respective subtheme. These tables were iteratively refined to clarify the core definition of each subtheme across methods, enabling a structured comparison that highlights similarities and differences. The resulting tables are presented in Chapter 5. Results.

Table 4.4: Overview of identified themes and corresponding subthemes from the analysis.

Theme	Subthemes
Interaction characteristics	Verbal labels Spatial references Pointing/gestures Instruction richness
Participant Experience and Cognitive Processing	Rhythm and pace Cognitive load Articulation challenges Emotional/behavioral effects
Moderator-oriented adaptation	Adaption style Social alignment Triggers for adaptation Impact on exploration Dependence on moderator
Coordination, Misunderstanding, and Context	Main challenge Coordination/spatial challenges Role dynamics Primary source of influence Environmental impact Role of moderator Observed effect
Moderator influence	Triggers for influence Type of influence Participant response Impact on problem-solving

4.3.3 Quantitative Analysis

Quantitative data, including task performance measures and response ratings, were analyzed and presented using tables and figures to support comparisons between the four test methods. Both participant responses and moderator assessments were compiled at the method level and presented side by side.

The System Usability Scale (SUS) consists of 10 items rated on a five-point Likert scale ranging from “strongly disagree” to “strongly agree” and is used to assess perceived usability [34]. This scale was retained to ensure comparability with prior research.

The scoring procedure was conducted in three steps. First, each of the ten item responses (rated 1–5) was converted to a contribution score. The SUS items alternate between positively and negatively worded statements. For positively worded items (1, 3, 5, 7, 9), the contribution was calculated as (*response* – 1), whereas for negatively worded items (2, 4, 6, 8, 10), it was calculated

as $(5 - \text{response})$. This transformation ensures that higher values consistently indicate better usability across all items.

Second, the contribution scores for all ten items were summed, resulting in a total score ranging from 0 to 40. Finally, the summed score was multiplied by 2.5 to obtain the overall SUS score, which ranges from 0 to 100 [34]. The final SUS score for each method was calculated by averaging the scores across participants.

Additional usability questions from the participant debriefings were analyzed using a similar scoring procedure to ensure consistency. Items were adjusted based on whether they were positively or negatively worded, following the same principle as for the SUS. The adjusted scores were then summed, resulting in a total score ranging from 0 to 16. To enable comparison with the SUS, this score was transformed to a 0–100 scale by multiplying by 6.25.

5. Results

This chapter presents the results from the four usability testing methods. The results are structured around key aspects relevant to the research questions: usability problems, task performance, interaction patterns, user experience, and the role of the moderator.

5.1 Summary of Method Differences

Figure 5.1 revisits the test environments introduced in Chapter 3 (Figures 3.1 and 3.2). The figure serves as a reminder for the results section, highlighting whether the participant or the moderator has access to the setup, as well as who is responsible for interacting with it.

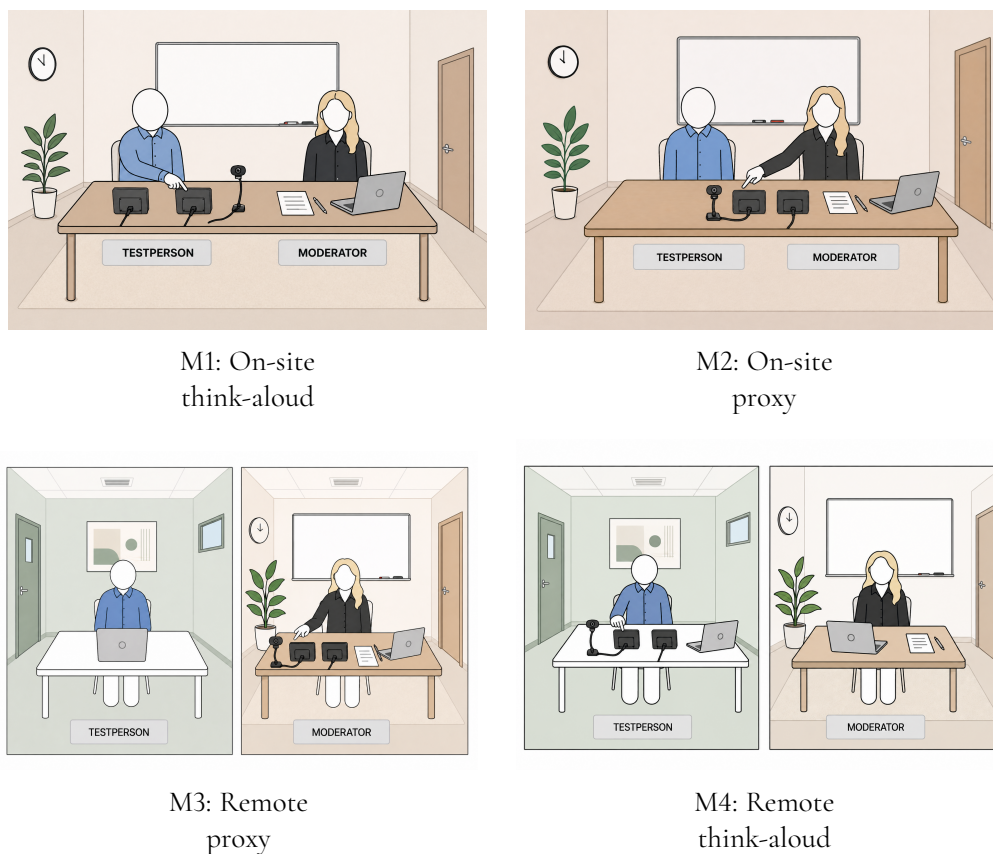


Figure 5.1: Overview of Method conditions. Illustrations created by ChatGPT.

Table 5.1 provides an overview of the key characteristics of the four test methods, including physical co-presence, interaction type, participant age range, and location.

Table 5.1: Summary of key characteristics of the four test methods. *Two of the tests were conducted in Sweden, Three in Denmark.

Method	Presence	Interaction	Ages	Location
M1	On-site	Direct	25–58	Sweden
M2	On-site	Proxy	24–35	Sweden/Denmark*
M3	Remote	Proxy	40–51	The Netherlands
M4	Remote	Direct	40–60	Switzerland

5.2 Usability Problems & Severity

Table 5.2 presents the number of usability problems identified for each method. In total, 52 usability problems were identified across the four methods.

Table 5.2: Overview of usability problems identified per test method. In total, 52 distinct usability problems were identified. The “Identified by all methods” column represents problems detected by all four methods (i.e., the overlap). The “Unique problems” row represents problems identified exclusively by a single method; the final column (13) indicates the total number of such problems.

	M1	M2	M3	M4	<i>Identified by all methods</i>
Number of problems identified	35	30	34	35	16
Percentage of problems identified (%)	67	58	63	67	31
Unique problems	2	3	1	7	13

The identified usability problems were assigned severity ratings based on the criteria presented in Table 4.3. In some cases, the same usability problem was assigned different severity levels depending on the method. The distribution of usability problems across severity levels is presented in Table 5.3.

The column “Max severity count” represents the number of usability problems categorized according to their highest observed severity across methods. For example, if a problem received severity ratings of 3, 3, 4, and 1 across the four methods, it is classified as severity 4 in this column.

A complete list of identified usability problems and their corresponding severity ratings for each method is provided in Appendix D.

While several issues were observed across multiple methods, each method also revealed distinct types of usability problems.

Table 5.3: Overview of the severities of usability problems identified per test method. "Max severity count" indicates the number of problems categorized by their highest observed severity across all methods.

	M1	M2	M3	M4	<i>Max severity count</i>
Critical problems	10	2	2	5	12
Major problems	9	10	12	10	18
Minor problems	11	10	18	12	14
Cosmetic problems	5	8	2	8	8

Method 1 (on-site think-aloud) identified a broad range of usability problems, including several critical issues. In particular, problems related to PIN-related interactions and login state were prominent. Participants frequently struggled with creating and using PIN codes, and additional issues were observed in interactions involving login and system state. This method also revealed problems related to system feedback and users' understanding of system status.

In Method 2 (on-site proxy), usability problems related to terminology and interpretation of user interface elements were prominent. Participants showed uncertainty in understanding icons, button functionality, and system concepts.

Similarly, Method 3 (remote proxy) revealed problems related to terminology and user interface interpretation. Compared to Method 2, fewer immediate usability breakdowns were observed, but confusion regarding symbols, labels, and system behavior remained evident.

Method 4 (remote think-aloud) revealed usability problems related to touch interaction, including difficulties in pressing buttons and issues with touch sensitivity. While some of these problems were also observed in Method 1, they were largely absent in the proxy-based methods. In addition, Method 4 identified issues related to feedback and system responsiveness.

Across methods, certain usability problems were consistently observed, particularly those related to PIN handling, system feedback, and mismatches between users' expectations and system behavior. However, the severity of these problems varied depending on the method. The methods also differed in the types of problems identified: direct interaction methods (Methods 1 and 4) more frequently revealed interaction-related issues, while proxy-based methods (Methods 2 and 3) highlighted issues related to terminology and interpretation.

The on-site methods more frequently revealed issues related to exploratory interaction and direct manipulation, while the proxy-based methods more often highlighted problems emerging through more structured and instruction-based interaction.

5.3 Task Performance

Table 5.4 presents task performance and time-related matter across the four methods. The task success rate ranges from 76% (M1) to 82% (M2).

Time-related measures vary more between methods. The average test time ranges from 42 minutes (M2) to 58 minutes (M3), and the average task execution time ranges from 25 minutes (M1) to 35 minutes (M3).

Table 5.4: Comparison of task performance across the four test methods, including task success rate, number of completed tasks, and time-related measures (11 tasks in total).

	M1	M2	M3	M4
Task success rate (%)	76	82	80	78
Tasks completed	8.4	9	8.8	8.6
Average test time (minutes)	43	42	58	56
Average task execution time (minutes)	25	26	35	30

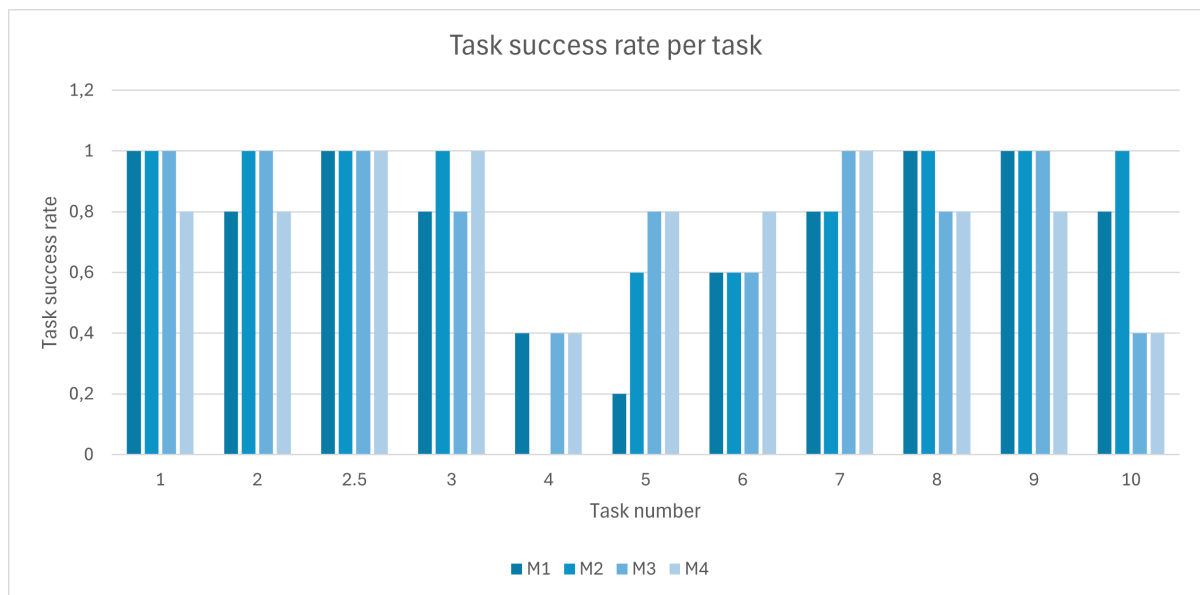


Figure 5.2: Task success rate among each task. 1 indicates 100%.

Figure 5.2 presents task success rates across all methods for each task (see Table 5.5 for list of Tasks, initially presented in Test Design). A task was assigned a value of 1 if completed successfully and 0 if unsuccessful.

Several tasks (e.g., Tasks 1, 2.5, and 9) show consistently high success rates across all methods. With Task 2.5 being the only task where all participants succeeded.

Greater variation between methods is observed for Tasks 4, 5, and 10. Task 4 shows consistently low success rates across all methods.

The lowest success rate is observed for M2 on Task 4, and M1 on Task 5, while higher success rates for Task 5 are observed for M3 and M4.

Table 5.5: Overview of usability test tasks and their success criteria (SCC).

#	Task	SCC
1	Enter system + change mode to auto (no PIN)	Mode set to auto
2	Change time-out (Admin code required)	Time-out set to 20 seconds
2.5	Turn off winter mode	Winter mode turned off
3	Close the door (no PIN)	Mode set to locked
4	Open door with authenticated trigger	Trigger used intentionally
5	Enable user PIN	PIN created with four digits
6	Change mode to auto (PIN) → open (no PIN)	Mode changed to auto and then open
7	Enter wrong PIN repeatedly	User notices PIN does not work
8	Navigate through issues	Error messages located
9	Find contact information	User identifies QR code
10	Understand issues	Stating most severe issue

5.4 Perceived Usability & Experience

This section presents participants' subjective experiences across the four methods, based on SUS scores and additional debriefing questions related to confidence, clarity, and navigation.

Table 5.6 presents perceived usability across the four methods. Mean SUS scores ranged from 74.5 (M3) to 78.5 (M4). The lowest individual score was 57.5 and the highest 92.5, both observed in Method 3. Mean ratings for the additional questions ranged from 73.75 (M2) to 86.25 (M1).

Table 5.6: Comparison of perceived usability across the four test methods, including mean SUS scores, score ranges, and aggregated results from additional participant perception questions (see Figure 5.3).

	M1	M2	M3	M4
SUS score (mean)	77	76	74.5	78.5
Lowest SUS score	65	65	57.5	70
Highest SUS score	90	85	92.5	90
Additional questions (mean)	86.25	73.75	76.25	81.25

Figure 5.3 presents variation between methods across additional questions. For the positively worded items, M1 shows the highest ratings, particularly for confidence, while M2–M4 are slightly lower and relatively similar.

For the negatively worded items, M2 shows the highest ratings, followed by M3, whereas M1 and M4 show lower values. The lowest rating of the two is observed for M4 on the statement regarding confusing symbols.

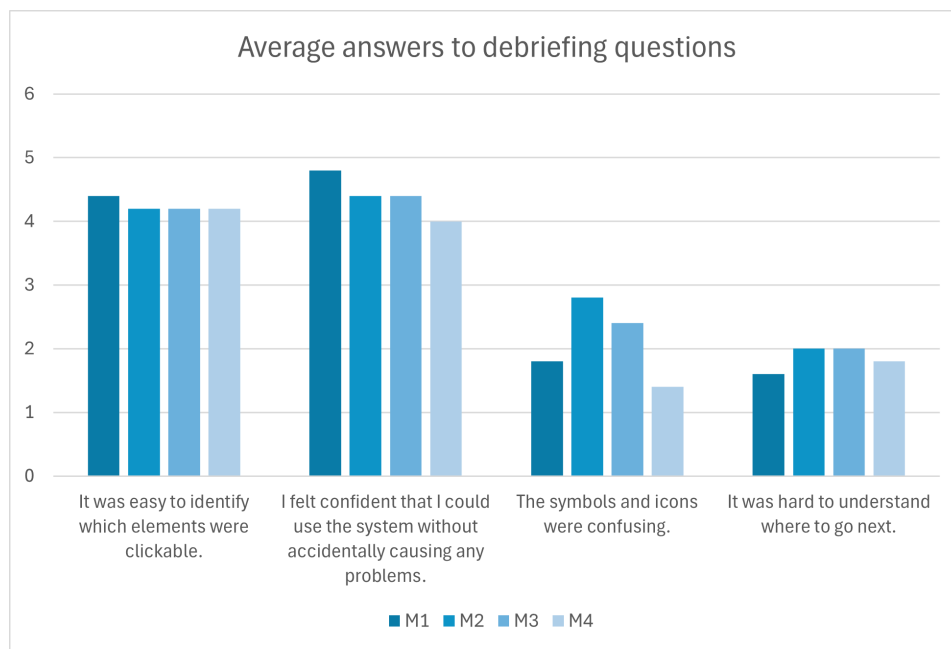


Figure 5.3: Average responses to debriefing questions across the methods, highlighting participants' perceptions of clarity, confidence, and navigation. Scores range from 1 ("strongly disagree") to 5 ("strongly agree").

5.5 Interaction Characteristics

This section presents observed interaction patterns across the four methods, focusing on how participants expressed actions, maintained control, and engaged with the user interface. Participants expressed actions either through concurrent think-aloud verbalizations or instruction-based communication, resulting in differences in communication style, level of detail, and interaction flow across methods.

Table 5.7 summarizes how participants expressed their actions across the methods. The qualitative ratings (very low–high) reflect the relative frequency and richness of observed behaviors, including verbal labeling, spatial referencing, and gesture use (see Appendix E Table E.1 for detailed analysis).

Proxy-based methods (M2, M3) show higher levels across all categories than direct interaction methods. The highest levels are observed in the remote proxy condition (M3), except for gesture use, which is most prominent in the on-site proxy condition (M2). The lowest levels are found in the remote think-aloud condition (M4), followed by the on-site think-aloud (M1).

Table 5.7: Comparison of instructional strategies across methods. Ratings (very low–high) reflect relative frequency and richness of behaviors observed across sessions.

	M1	M2	M3	M4
Verbal labels	Low	Medium	High	Very low
Spatial references	Low	Medium	High	Low
Pointing / gestures	Low	High	Medium	Low
Instruction richness	-	Medium	High	-

Table 5.8 presents a condensed overview of recurring interaction patterns in proxy methods (see Appendix E Table E.2 for detailed analysis).

Table 5.8: Comparison of how proxy interaction affects rhythm, cognition, and behavior. Descriptions summarize recurring interaction patterns observed across sessions.

	M2 (On-site proxy)	M3 (Remote proxy)
Rhythm and pace	Slower, disrupted flow	Slower, more hesitant and deliberate
Cognitive load	Increased (wording effort)	Increased (reflective processing)
Articulation challenges	Frequent difficulty describing actions	Moderate difficulty; some adaptation
Emotional / behavioral effects	Reduced confidence; cautious behavior	Increased caution; occasional hesitation
Overall pattern	Disrupted spontaneity	Reflective, slowed interaction

Both methods are associated with slower interaction compared to direct interaction. M2 shows a more disrupted interaction flow, while M3 is characterized by a more hesitant and deliberate pace.

In both methods, more cautious behavior is observed, with reduced confidence and increased hesitation reported during interaction.

5.6 Moderator-Oriented Adaptation

This section presents patterns in moderator behavior and how participant–moderator interaction varied across methods. Table 5.9 shows differences in how participant behavior related to the moderator in each method (see Appendix E Table E.3 for detailed analysis).

Building on this, Table 5.10 presents a quantified interpretation of these patterns. The ratings (very

low–very high) reflect the relative extent of adaptation, dependence, and impact on exploration observed across methods.

Table 5.9: Summary of how participant behavior relates to the moderator across methods. Descriptions summarize recurring patterns of social alignment, dependence, and changes in agency observed across sessions.

	M1	M2	M3	M4
Adaptation style	Adjusting pace	Waits for cues	Polite instructions	Adjusts to clarity
Social alignment	Seeks validation	Leans back and disengages	Holds back and wait for prompts	Struggles to align
Triggers for adaptation	Unclear expectations	Unclear permission to explore	Unclear who leads	Misunderstood goals
Exploration	Slightly reduced	Strongly reduced	Hesitant	Structured by moderator
Overall pattern	Performance-oriented	Depends on cues and permission	Cue-dependent and socially hesitant	Participant–moderator misunderstandings

Table 5.10: Degree of participant adaptation to the moderator across methods. Ratings (very low–very high) reflect the relative extent of adaptation, dependence, and impact on exploration observed across sessions.

	M1	M2	M3	M4
Adaptation to moderator	Medium	Very high	High	Medium
Social alignment	Medium	High	High	Medium
Impact on exploration	Low	Very high	Medium	High
Dependence on moderator	Low	Very high	High	Medium

Taken together, Tables 5.9 and 5.10 show differences in how participants adapted their behavior in relation to the moderator across methods.

Across methods, clear differences are observed in how participants relied on the moderator and how this affected exploration. M2 stands out with the highest levels of dependence, where participants consistently waited for confirmation and showed reduced independent exploration. In contrast, M1 shows lower dependence, with participants maintaining greater control while still seeking validation.

Compared to M2, M3 shows a more hesitant and polite interaction style. Although social alignment remains high in both methods, M3 is characterized by lower levels of dependence and a less

pronounced impact on exploration, with participants more often waiting for prompts rather than actively seeking confirmation.

M4 differs from both M1 and the proxy-based methods by showing a pattern driven by misunderstandings of task goals. Participants adapted to clarity and pacing and relied on moderator guidance in these situations, while also struggling to establish shared understanding.

Overall, M2 represents the strongest shift toward moderator-dependent interaction, while M1 reflects more independent behavior. M3 and M4 fall in between, but differ in that M3 is characterized by hesitation, whereas M4 is shaped by misunderstandings and alignment difficulties.

5.7 Coordination, Misunderstandings, & Context

This section presents coordination patterns and misunderstandings observed in proxy interaction, focusing on how tasks were interpreted and performed. Table 5.11 shows differences in coordination challenges between M2 and M3, (see Appendix E Table E.4 for detailed analysis).

Table 5.11: Summary of coordination challenges in proxy interaction.

	M2 (On-site proxy)	M3 (Remote proxy)
Main challenge	Inconsistent instruction clarity	Spatial and role ambiguity
Coordination / spatial challenges	Unclear instructions and spatial confusion	No shared spatial reference
Role dynamics	Participant reaches for control	Moderator steps in to guide
Overall	Interaction-driven issues	Structurally driven issues

Coordination challenges are observed in both proxy methods, including shared disruptions related to system timing and feedback (e.g., system timeouts), as well as difficulties in conveying directional gestures such as swiping (e.g., ambiguity in what constitutes left and right).

In M2, misunderstandings are primarily linked to variation in instruction clarity, with participants showing inconsistencies in how actions are described. Role distribution is generally clear but occurs in an unfamiliar interaction setting, sometimes leading participants to attempt to take over interaction.

In contrast, M3 is characterized by spatial ambiguity and role uncertainty. Participants lack a shared reference frame and show uncertainty regarding who is responsible for acting and instructing, resulting in unclear shifts between participant-led and moderator-led actions.

Table 5.12 shows differences in the primary sources of influence and corresponding observed participant responses across methods, based on moderator reports and observed interaction patterns (see Appendix E Table E.5 for detailed analysis).

Table 5.12: Comparison of primary sources of influence, moderator roles, and corresponding participant responses across methods.

	M1	M2	M3	M4
Primary source of influence	Seeks confirmation from moderator	Subtle interpersonal cues (pacing, hesitation)	Environmental limitations (lack of feedback)	Communication barriers (audio, language)
Environmental impact	Prototype constraints	Sensitivity to moderator cues	Missing system feedback	Technical constraints
Role of moderator	Clarifications become leading	Indirect influence through social signals	Prompts interpreted as hints	Increased intervention and support
Observed effect	Confusion and repeated clarification	Reduced exploration	Interpreting moderator intent and increased reliance	Increased cognitive effort, hesitation, and stress
Overall pattern	Confirmation-driven interaction	Social cue-driven interaction	Environment-driven interaction	Communication-driven interaction

Across methods, participant behavior is shaped by different sources of influence, with clear differences in how these are expressed.

In M1 and M2, influence is tied to the interaction itself. In M1, this appears as participants seeking confirmation from the moderator, resulting in confusion and repeated clarification. In contrast, M2 is shaped by subtle interpersonal cues, where tone and pacing influence behavior, leading to reduced exploration and sensitivity to the moderator.

In M3 and M4, influence is instead linked to contextual constraints. In M3, environmental limitations, particularly the absence of system feedback, lead participants to interpret prompts as hints, resulting in interpreting moderator intent and increased reliance. In M4, communication barriers such as audio quality and language are associated with increased cognitive effort, hesitation, and reliance on moderator support.

Overall, the methods differ in whether participant behavior is primarily shaped by interaction dynamics (M1–M2) or by environmental and communication constraints (M3–M4).

Figure 5.4 presents moderator ratings of test setup and structural conditions across methods. Ratings for having the time to prepare are high across all methods, as well as ratings for equipment functioning which are consistently high across all methods, with slightly lower values observed for M4.

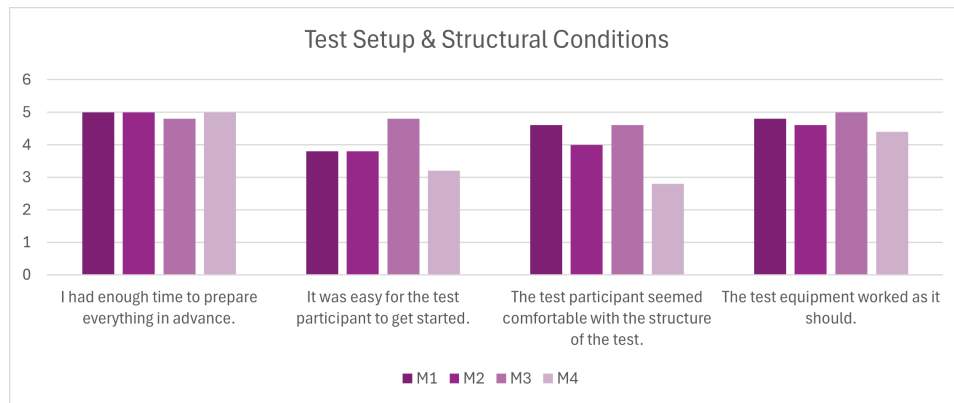


Figure 5.4: Average response for the test moderator debriefing form regarding the test setup and structural conditions. Scores range from 1 (“strongly disagree”) to 5 (“strongly agree”).

Differences are observed in how easily participants were able to get started, with higher ratings for M3 and lower ratings for M4. Participant comfort with the test structure is rated highest in M1 and M3, and lower in M2 and M4.

Figure 5.5 presents moderator ratings of environmental conditions across methods. Ratings for a calm and distraction-free environment are high across all methods, with M4 showing the highest value.

Clarity of audio is rated highest for M1 and M3, and slightly lower for M2 and M4. And visibility of participant interaction (applicable to M1 and M4) is rated slightly higher in M1 than in M4.

As the item regarding the effect of the test environment is negatively worded, higher values indicate a greater perceived negative impact. Ratings are generally lower, with slightly higher values observed for M3. Some level of negative impact is reported across all methods.

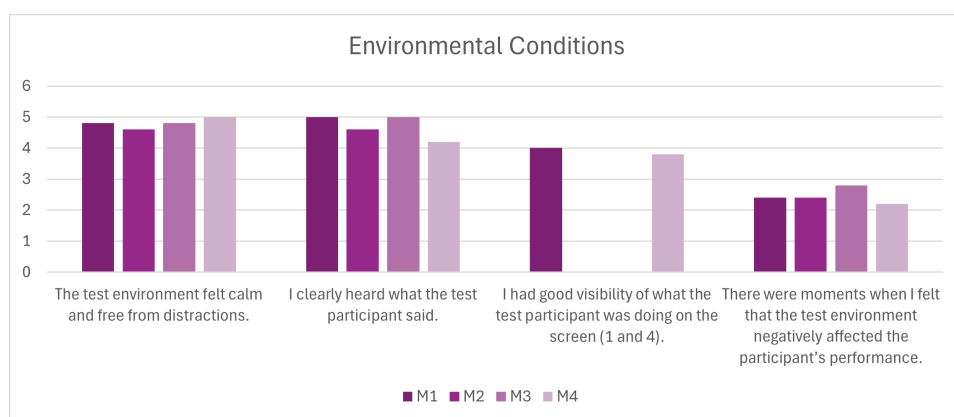


Figure 5.5: Average response for the test moderator debriefing form regarding environmental conditions. Scores range from 1 (“strongly disagree”) to 5 (“strongly agree”). The third question was only applicable for method 1 and 4.

5.8 Moderator Influence

Moderator influence was observed across all methods, with differences in how it was expressed and how participants responded.

Table 5.13 presents how moderator influence is expressed across methods, based on qualitative analysis of moderators' free-text self-assessments, (see Appendix E Table E.6 for detailed analysis).

Table 5.13: Comparison of how moderator influence shapes participant behavior and problem-solving across methods.

	M1	M2	M3	M4
Triggers for influence	System confusion	Need for confirmation	Uncertainty	Unclear expectations
Type of influence	Subtle cues (tone, reactions)	Clarifications and timing	Guides through questions	Waits for moderator input
Participant response	Self-correction	Caution and hesitation	Adjusted reasoning	Reliance on moderator for help
Impact on problem-solving	Mild unintended guidance	Reduced confidence	Reduced independence	Moderator-led actions

Moderator influence varies across methods in both expression and impact. M1 shows the lowest level of influence, characterized by subtle cues and participant self-correction, with limited impact on problem-solving. In contrast, M2 shows more explicit influence, where instruction timing and clarification shape participant behavior, resulting in increased hesitation and reduced confidence.

Compared to M2, M3 shows a shift toward influence through guided reasoning. Participants adjust their thinking and often wait when tasks are unclear, corresponding to reduced independence in problem-solving, although to a lesser extent than in M2.

M4 stands out with the strongest moderator influence. Participants rely on moderator input to proceed, and problem-solving is frequently moderator-led, particularly in situations involving unclear expectations.

Figure 5.6 complements these findings by presenting moderators' quantitative self-assessments across the same methods.

The perceived role of the moderator's presence in participants' thinking-aloud is strongest in M2 and weakest in M4, with M1 and M3 showing similar intermediate levels.

For perceived neutrality, M4 shows the highest rating, while M1 and M2 report similar levels, and M3 shows a lower value. The frequency of unintentional cues is reported as lowest for M2 and M4, and higher for M1 and M3, with M3 showing the highest value.

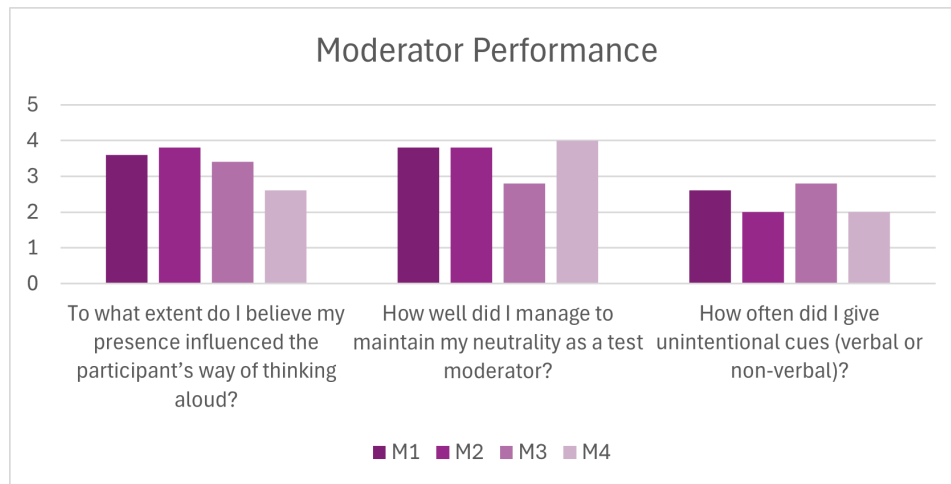


Figure 5.6: Average response for the test moderator debriefing form regarding influence, neutrality, and unintentional cues. Scores range from 1 (“Not at all”) to 5 (“To a great extent”).

5.9 Participant Connection

This section presents moderator-reported assessments of social interaction and participant connection across methods, with observed differences related to communication flow and social alignment.

Figures 5.7 and 5.8 present moderator assessments of participant connection and test experience across methods.

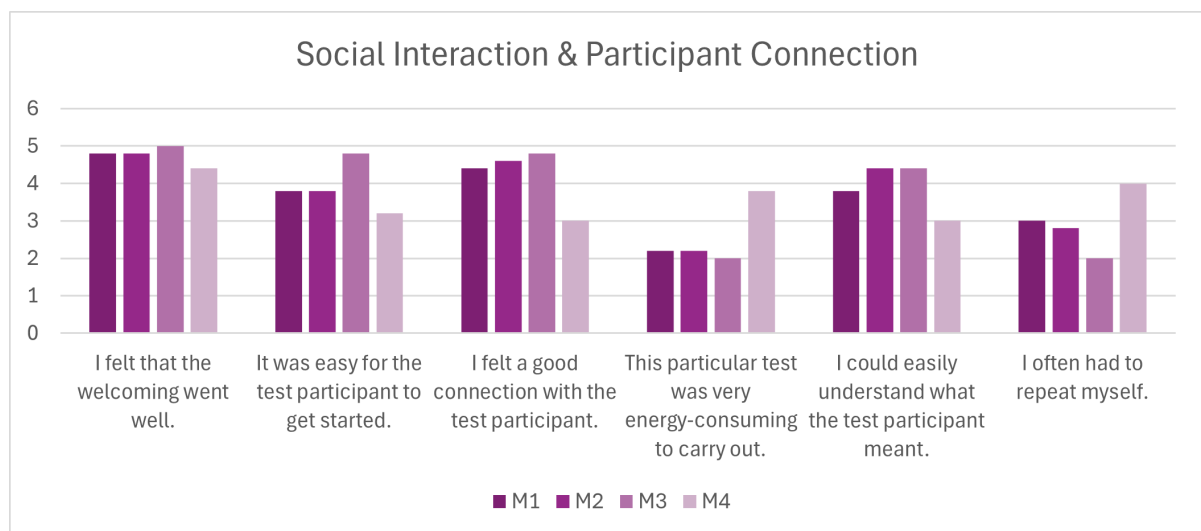


Figure 5.7: Average response for the test moderator debriefing form regarding social interaction and participant connection. Scores range from 1 (“strongly disagree”) to 5 (“strongly agree”).

Ratings related to social interaction and participant connection vary across methods. M3 shows consistently high ratings across items related to welcoming, ease of getting started, and perceived connection, indicating strong reported interaction quality.

Compared to M3, M2 shows slightly lower but still high ratings, particularly for welcoming and understanding the participant.

M1 shows similar ratings to M2 for welcoming and ease of getting started, but slightly lower for connection and communication clarity.

M4 differs from the other methods, with lower ratings for ease of getting started and perceived connection. Communication-related challenges are more pronounced, with higher ratings for repetition.

For the negatively worded item regarding energy consumption, higher values indicate greater perceived effort. M4 shows the highest rating, while M1–M3 show lower values.

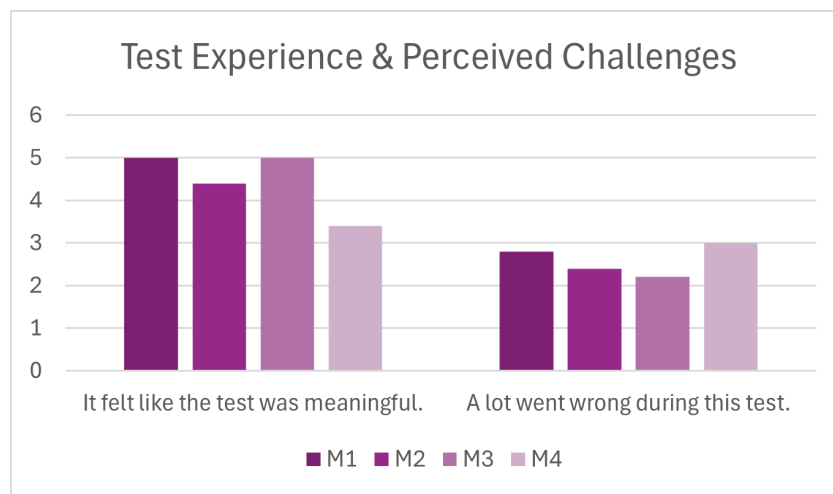


Figure 5.8: Average response for the test moderator debriefing form regarding test experience and perceived challenges. Scores range from 1 (“strongly disagree”) to 5 (“strongly agree”).

In terms of overall test experience (Figure 5.8), the test is rated as meaningful across all methods, with lower values observed for M4.

For the negatively worded item regarding issues during the test, ratings are moderate across methods. Slightly higher values are observed for M4, while M3 shows the lowest rating.

5.10 Debriefing & Post-Task Behavior

Participants completed a debriefing form after the tasks. In on-site settings, the form was provided on the moderator’s computer, while in remote settings, participants received a link and completed it during the session. In M3 and M4, participants occasionally asked whether the form could be completed at a later time.

Across methods, participants were generally focused and engaged. Engagement was consistent in M2 and M3, either through quiet completion (M2) or verbal reflection (M3). In M1, engagement varied more, with some distraction and need for clarification. In M4, engagement was quieter, with participant activity less visible to the moderator.

After submitting the debriefing form, participants engaged in casual conversation, marking a transition out of the test session. This phase was more relaxed, with participants appearing calmer and less focused on performance. In sessions that ran longer, less time naturally remained for casual conversation.

In the on-site methods (M1 and M2), some participants used this time to explore the control panel, while others engaged in small talk. In M3, conversations were more polite and oriented toward getting to know each other. In M4, the interaction typically ended quickly with minimal conversation.

These observations were considered separate from the test session but reflect patterns of exploration and social interaction.

5.11 Summary of Key Results

To summarize, the key results are presented based on the main differences observed across methods. Table 5.14 provides an overview of the methods.

Table 5.14: The methods and their differences. Horizontal: physical co-presence, vertical: interaction execution.

	On-Site	Remote
Direct (Think-Aloud)	Method 1 (On-site think-aloud)	Method 4 (Remote think-aloud)
Proxy	Method 2 (On-site proxy)	Method 3 (Remote proxy)

Usability Findings

Direct interaction methods (M1, M4) primarily revealed interaction-related issues, including touch interaction and system feedback. In contrast, proxy-based methods (M2, M3) more often identified problems related to terminology and interpretation of user interface elements. Some issues, such as PIN handling and system feedback, were observed across all methods, although their severity varied.

Task Performance

Task success rates were high across methods for straightforward tasks (e.g., entering the system (Task 1), turning off winter mode (Task 2.5), and finding contact information (Task 9)), with all participants completing Task 2.5 successfully. In contrast, tasks requiring more complex interaction or system understanding showed greater variation between methods. Opening the door using an authenticated trigger (Task 4) showed consistently low success rates, with none completing it in M2, while creating a user PIN (Task 5) showed lower success rates in M1 and higher in M3 and M4. Overall, no method consistently outperformed the others, and performance differences were primarily task-dependent.

Perceived Usability

Perceived usability ratings were relatively similar across methods, with slightly higher SUS scores in M4 and lower in M3, with M3 also showing a wider range of scores. In contrast, M1 showed the highest ratings across the additional questions, particularly for confidence, while M2 and M3 showed higher values for negatively worded items, indicating more perceived difficulty.

Instructional & Communicative Strategies

Proxy-based methods (M2, M3) showed higher levels of verbal labels, spatial references, and instruction richness compared to direct interaction methods, with the highest levels observed in M3. In contrast, M2 relied more on gestures, reflecting co-located interaction, while M1 and M4 showed lower levels across these measures, with M4 showing the lowest use overall.

Interaction Patterns & Behaviour

Proxy methods (M2, M3) were associated with slower and more cautious interaction, reduced interaction confidence, and greater reliance on the moderator. M2 showed the highest level of dependence, with frequent confirmation-seeking, while M3 was characterized by hesitation and a more deliberate pace. In contrast, M1 showed more independent interaction, while M4 was marked by misunderstandings and difficulty establishing shared understanding.

Coordination & Communication in Proxy Methods

Both proxy methods showed coordination challenges, including difficulties with spatial communication (e.g., left/right) and disruptions related to system feedback. M2 was characterized by inconsistencies in instruction clarity, while M3 showed ambiguity in roles and lack of a shared reference frame, resulting in unclear shifts in interaction control.

Sources of Influence

Participant behavior was shaped by different dominant factors across methods. M1 was characterized by confirmation-seeking behavior, M2 by sensitivity to social cues, M3 by environmental constraints, and M4 by communication barriers. Overall, M1–M2 were primarily interaction-driven, while M3–M4 were shaped by contextual constraints.

Moderator Influence

Moderator influence varied across methods and affected both interaction and outcomes. Influence was lowest in M1, where participants retained control but frequently sought confirmation, and highest in M4, where interaction was often moderator-led. M2 showed explicit influence through timing and clarification, while M3 reflected guided reasoning, resulting in more hesitant but structured interaction.

Experience & Engagement

Participant connection varied across methods. M3 showed the highest ratings for interaction quality and connection, while M4 showed lower ratings, including higher effort and more communication-related challenges. Engagement during debriefing was consistent in M2 and M3, more variable in M1, and quieter in M4.

6. Discussion

This chapter discusses the findings of the study in relation to the research questions, previous work, and theoretical perspectives. Particular attention is given to how different usability testing methods influenced reliability, interaction patterns, perceived experience, and the role of the moderator. In addition, methodological choices, limitations, and implications of the findings are considered in order to provide context for how the results should be interpreted.

6.1 Process & Methods

Several methodological choices influenced how the results were produced and should be interpreted. While the study was designed to enable systematic comparison across testing setups, the execution introduced trade-offs and constraints that shaped both data collection and analysis. The following sections reflect on these aspects, focusing on how they may have influenced the findings.

6.1.1 Preparations

The preparation phase of the study required substantial time and effort, particularly in relation to participant recruitment and logistical coordination. Conducting several test methods with a relatively large participant sample introduced a level of complexity beyond typical small-scale usability studies.

Although participants were recruited internally, the process depended on coordination across organizational roles and locations, highlighting how recruitment is shaped by availability and organizational structures.

This reflects how usability testing in practice often requires more planning and coordination than initially anticipated. At the same time, the chosen approach made it possible to include a broader participant sample across different countries, which was particularly relevant given the study's focus on remote testing and international applicability.

In addition, the preparation phase revealed important methodological trade-offs. While efforts were made to standardize the setup across test methods, differences in physical location, technical configuration, and participant context introduced variability that could not be fully controlled.

Whether remote usability testing reduces time was not assessed in this study. The complexity of coordinating multiple setups made it difficult to isolate preparation time, so the analysis focused on session duration.

6.1.2 Participants

The study ultimately included 20 participants, of whom 19 were men and 1 was a woman, resulting in a heavily male-dominated sample. While participants were recruited from a niche group requested by the company, the gender distribution within these groups was not specified. Participation was voluntary and mediated through managers, meaning that the final sample composition was not fully controlled by the author.

Since participants from different countries were assigned to different test methods, contextual factors such as cultural differences, work environments, and communication styles may have influenced the results.

Although all participants had English as a second language, noticeable differences were observed between locations. This was particularly evident in Method 4, where the participants had lower confidence and proficiency in English compared to other groups. As a result, communication became more challenging, especially in the remote setting.

In addition, the age distribution varied across methods. M1 included a wider range (25–58), M2 consisted of younger participants (24–35), while M3 and M4 included older participants (40–51 and 40–60, respectively). This variation may have influenced how participants interpreted symbols and system functions, as well as their comfort with language and technology, although all participants had professional experience working with technology.

6.1.3 Order Effects & Data Analysis

The methods were run one after the other due to practical constraints, rather than being randomized. This likely introduced some order effects, as the moderator became more familiar with the procedure, equipment, and product over time. As a result, later sessions may have benefited from improved pacing and smoother handling of the setup, which could have influenced the consistency between methods.

The analysis also followed a sequential structure, where Method 1 was analyzed in depth first and subsequent methods were interpreted in relation to emerging patterns. This facilitated comparison across methods but also shaped the analytical lens through which later data were interpreted.

As a result, the identification of patterns, such as differences in interaction, verbalisation, and moderator influence, was not entirely independent across methods, but developed iteratively over time. This reflects the evaluator effect, where findings are influenced by the researcher's perspective and analytical framing [20]. Rather than being identified in isolation, some patterns were reinforced through repeated observation across methods.

Generative AI was used primarily to analyze moderator debriefings, as the volume and complexity of the data exceeded expectations. While the structured format supported comparison across research questions, many findings were interrelated, making manual analysis difficult. AI was therefore introduced to improve efficiency, using an iterative process where outputs were continuously reviewed and validated against the original data to mitigate misinterpretation and data mixing across methods, which was a recurring challenge.

In contrast, the thematic analysis of usability problems was conducted manually without AI assistance, allowing for clearer traceability between observations and findings, but requiring significantly more time. Hence the decision to adopt a more efficiency-driven approach for the analysis of Methods 2–4.

Despite differences in analytical approach, recurring patterns were observed consistently across methods, supporting the overall credibility of the findings while also highlighting different aspects of the interaction. Thematic analysis offered more systematic and traceable insights, whereas the moderator-focused analysis, shaped by time constraints and supported by AI, was more efficiency-driven. This also reflects an initial underestimation of the moderator's role, which emerged as a central factor in the study.

6.1.4 Single-Moderator Limitation

During the data analysis process, it became increasingly evident that the moderator influenced both the test procedure and its outcomes. While it was still possible to identify patterns across the different methods, these patterns should be understood in relation to the fact that all sessions were conducted by the same moderator.

Although the relatively large number of participants contributed to a broad range of data and insights, all interactions were shaped by the same moderator guiding the sessions. This introduces a limitation when interpreting the results, as the observed patterns may partly reflect moderator-specific effects rather than solely differences between methods.

Including multiple moderators could have strengthened the generalizability of the findings by reducing the influence of individual moderation styles. However, such an approach would require a larger study with a different scope and was therefore beyond the aims of the present work.

6.2 Results

This section discusses the findings from the study, focusing on key aspects such as task performance and perceived usability, identified usability problems, interaction patterns and verbalisation, and the role of the moderator across testing conditions.

6.2.1 Task Performance, Perceived Usability & Experience

At first glance, the results suggest that the choice of method does not matter that much. Participants were generally able to complete the tasks, and perceived usability scores were similar across methods. However, this similarity starts to break down when looking at how the interaction actually unfolded.

What stands out is not whether participants succeeded, but how they got there. In some methods, tasks were completed through quick trial-and-error, while in others, participants spent more time reasoning, interpreting, and trying to understand the system before acting.

Time also reflects this shift. Remote and proxy-based sessions took longer, not necessarily because tasks were more difficult, but because participants had to explain, interpret, or wait rather than act directly. In this sense, the interaction becomes less about doing and more about thinking through the system.

This makes the results somewhat misleading if taken at face value. Similar task success rates and SUS scores suggest comparable usability, but they mask important differences in how effortful, intuitive, or natural the interaction actually was. This reflects a known limitation in usability evaluation, where outcome-based measures may overlook important aspects of user experience and interaction processes [22]. What appears as equivalent performance may therefore represent fundamentally different ways of interacting with and interpreting the system. Evaluating usability based only on outcomes therefore risks overlooking how the interaction is actually experienced in practice.

6.2.2 Usability Problems

Differences in how interaction unfolds also affect which usability problems become visible. Direct methods (M1, M4) primarily revealed interaction-related issues, such as touch sensitivity and system feedback, whereas proxy methods (M2, M3) more often exposed problems related to terminology and interpretation. This also influenced perceived severity across methods. These findings support prior research showing that usability problems are shaped not only by the user interface, but also by the testing method and interaction conditions [11, 21].

A key difference is the absence of physical interaction in proxy methods. From an embodied interaction perspective, physical engagement is central to how users perceive and understand a system [6]. In the direct methods, participants compared the control panel to familiar devices such as smartphones and reacted to its touch response. When input was not registered, they repeated actions, leading to persistent frustration that shaped their understanding of the system. These issues were largely absent in proxy methods, where no physical interaction occurred.

Instead, focus shifted toward visible user interface elements. Without direct interaction, participants attended more to labels, icons, and system logic, resulting in a more analytical reading of the user interface. However, this does not necessarily reflect real-world use. From a situated cognition perspective, this illustrates how the testing context shapes user interpretation and engagement [4, 5]. Interaction became more clue-driven, with participants reasoning through the user interface and often scanning the entire screen before acting. This more deliberate approach sometimes led them to engage with non-interactive elements, such as pressing icons not intended as buttons.

In contrast, participants in the direct methods acted more instinctively, exploring through action rather than reflection. They pressed buttons without fully reading the user interface, allowing understanding to emerge through use. As a result, perceptions of what was clickable developed more spontaneously.

Some issues, such as PIN handling and system feedback, appeared across all methods, indicating core design problems. However, they manifested differently. In proxy methods, the slower pace made details like PIN length more noticeable. In direct methods, these were often missed, affecting

subsequent tasks. Participants eventually resolved login through trial and error, often forming incorrect assumptions about system logic and the purpose of the PIN. In remote methods, where interaction was less exploratory, participants instead struggled to understand the system despite having the correct PIN.

The method influences not only which problems are found, but how they are discovered. Direct interaction exposes breakdowns as they occur, while proxy interaction enables reflection and interpretation. This can be understood through distributed cognition, where interaction is distributed across participant, moderator, and artefacts, shaping how problems are identified and articulated [3, 2]. Together, the methods support two distinct modes of engagement: direct interaction foregrounds physical usability and immediate feedback, while proxy interaction highlights reflective interpretation of labels, icons, and system structure.

The relatively low overlap between methods further suggests that no single approach captures the full range of usability issues; rather, they complement each other by revealing different aspects of the user experience.

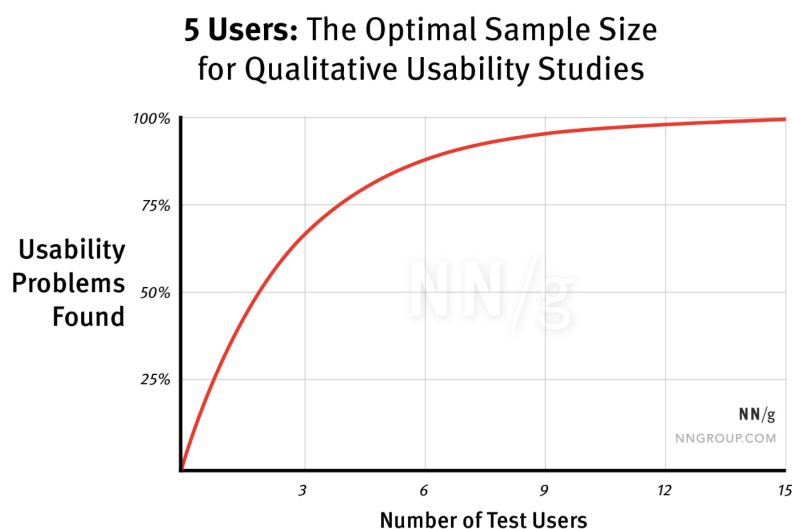


Figure 6.1: Usability problems found based on number of test users according to Nielsen Norman Group [28].

This pattern is also reflected in the cumulative number of identified problems. In total, 52 usability problems were identified across 19 participants. Method 1 identified 35 (67%) with 5 participants. After Method 2, this increased to 41 (79%) with 9 participants, and after Method 3 to 45 (87%) with 14 participants.

Compared to Figure 6.1, the curve would suggest that close to 100% of problems should be identified by this point. However, only 87% were identified, corresponding roughly to the number of problems identified by six users according to the curve

This indicates that problem discovery was not solely a function of total sample size. Instead, different methods contributed distinct and complementary insights, resulting in a more comprehensive understanding of the evaluated product.

6.2.3 Think-Aloud vs Instruction-Based Interaction

A central question in this study was whether proxy-based interaction provides richer insight into participants' understanding of a system than traditional think-aloud. At first glance, it appears to do so: in the proxy conditions (M2, M3), participants continuously described intended actions, leading to more structured and explicit verbalizations.

As shown in Table 5.7, this was reflected in higher levels of verbal labeling, spatial references, and overall instruction richness, particularly in M3 (remote proxy). Participants frequently used phrases such as “press the button on the right,” “go to the top menu,” or “select the lock icon,” making their interpretation of the user interface more visible than in the think-aloud conditions. Rather than merely reacting to the user interface, they articulated what they believed the system afforded. This aligns with research on communicative acts in usability testing, which suggests that instruction-based interaction promotes more explicit articulation of reasoning and intent [25].

This contrasts with the traditional think-aloud approach described by Ericsson and Simon [7], where verbalizations are intended to accompany action without interfering with it. In practice, much of the reasoning in M1, and especially M4, remained implicit. Participants often acted before speaking or produced fragmented utterances, leaving interpretation to the researcher rather than making their reasoning fully explicit.

In the proxy methods, verbalization became the interaction itself. Participants had to explain actions in advance, requiring them to structure their understanding. In this sense, they were not only expressing their thoughts but actively shaping them to make them communicable.

However, this richer verbal access came at a cost. As shown in Table 5.8, participants in M2 often struggled to describe actions precisely, leading to pauses, reformulations, and vague instructions. In M3, they adapted over time and produced more complete instructions, but the interaction became slower and more hesitant. This suggests that participants were not only solving the task but also managing the effort of translating their understanding into language. From a think-aloud perspective, this reflects how verbalization can impose additional cognitive load and influence task performance [7, 24].

This raises the question of whether richer verbal data is always beneficial. While the need to formulate instructions sometimes prompted deeper reflection, making misunderstandings or incorrect assumptions easier to identify, it also introduced new challenges. Ambiguous instructions, such as unclear spatial references or vague descriptions, occasionally led to moderator misinterpretations, affecting task execution. In this sense, the method not only reveals understanding, but can also shape and occasionally distort it.

In contrast, think-aloud methods offered less explicit but more natural insight into participants' interaction behaviour. Verbalizations were typically brief reactions or immediate interpretations, requiring greater researcher interpretation but more closely reflecting natural interaction with the system.

Taken together, proxy interaction yields richer verbal data, but this insight is constructed through the interaction itself. The added cognitive load and slower pace are part of how this data is produced, while misunderstandings and coordination challenges show that more verbalization does not necessarily lead to clearer insight.

Seen in this light, the choice between think-aloud and instruction-based approaches is less about data richness than about the kind of understanding sought. Proxy methods make users' interpretations explicit at the cost of natural interaction, while think-aloud preserves more realistic behavior but requires the researcher to infer users' reasoning. Neither approach is inherently better, but they reveal different aspects of interaction, reasoning, and perceived experience.

6.2.4 Remote vs On-Site: Control & Connection

On-site and remote methods differed in how connection and control were intertwined. In on-site settings, the moderator could directly shape both the environment and the interaction: welcoming the participant, guiding the setup, and observing subtle behaviors helped establish shared understanding early in the session. From a situated cognition perspective, this reflects how cognition and interaction are shaped by physical and social context [4, 5]. This appeared to reduce hesitation and enabled more independent engagement.

Remote methods differed in how control was distributed and when it came into play. In Method 4 (remote think-aloud), the loss of control was most evident during setup, where an assistant prepared the environment and welcomed the participant before the session began. Although the assistant left once the connection was established, this initial handover meant the moderator lacked full control over how the test was introduced and configured.

As a result, the moderator entered at a later stage, adapting to pre-established conditions and limiting early influence. From a distributed cognition perspective, parts of the system were already in place before the moderator became fully involved, shaping subsequent coordination [3, 2]. The moderator thus shifted from guiding to reacting, for example when participants moved the control panel, misunderstood instructions, or deviated from the task flow. In some cases, the simulator itself became a source of confusion when placed out of sight, requiring small interruptions.

Reduced visibility in M4 also made it harder to judge when to intervene. Without access to body language or subtle cues, the moderator had less basis for interpreting participant uncertainty, sometimes leading to delayed guidance or increased participant reliance. This aligns with research on remote usability testing, where the absence of nonverbal cues complicates interpretation and timely support [11].

In contrast, Method 3 (remote proxy) was less affected by setup but introduced limitations during interaction. The moderator remained directly involved, yet the lack of shared physical context made instructions such as “press left” or “go there” ambiguous. Rather than being lost, control had to be continuously negotiated, as meaning and intent were not always aligned. This reflects how proxy-based interaction becomes a shared, distributed activity, requiring ongoing coordination of understanding [3].

These differences also shaped participants' relation to the moderator. In on-site methods, they appeared more independent, with a clearer separation of roles. In remote settings, this boundary weakened: M3 often became more collaborative, reflected in higher connection ratings, whereas M4 felt more distant and effortful, with greater reliance on verbal clarification. From a situated perspective, this suggests that connection is shaped by how interaction is structured and supported within the testing context.

Taken together, remote testing not only reduces control over setup but also reshapes how connection is established and maintained. As control decreases, interaction becomes more dependent on communication and coordination, shaping participant comfort and session flow. In this sense, connection is not only a social factor but also tied to the moderator's ability to manage and stabilize the testing situation.

6.2.5 The Moderator's Role in Balancing Support & Neutrality

Across all methods, a central tension concerned the moderator's role in balancing two responsibilities: ensuring session progress and participant comfort while maintaining neutrality. This reflects a well-known challenge in usability testing, where moderators are expected to support the session while minimizing their influence on participant behavior [7, 24].

In practice, this balance proved difficult. Participants frequently showed signs of uncertainty, hesitation, or stress, requiring the moderator to decide whether to intervene or remain passive. While intervention could reduce frustration and help participants progress, it also risked influencing their behavior and the study outcomes.

This tension was particularly visible in the proxy methods (M2, M3), where the moderator was directly involved in executing actions. Here, the role became dual: both facilitator and the participant's "hands." When instructions were unclear, such as "press left" or "go there," the moderator had to interpret intent. From a distributed cognition perspective, interaction was thus shared between participant and moderator, with decision-making and execution forming a joint system [3, 2]. This meant that the moderator's judgments directly shaped the interaction: clarifying too much could lead participants, while too little could cause confusion or stall progress.

In M2, this often led participants to wait for confirmation before acting, suggesting that the moderator's presence influenced not only actions but also confidence. In M3, responsibility shifted slightly, with participants expressing themselves more cautiously and the moderator guiding through subtle cues rather than explicit clarification. In both cases, the moderator remained central to maintaining progress, though at the cost of reduced participant independence.

In contrast, M1 placed greater emphasis on moderator neutrality, in line with traditional think-aloud principles [7]. The moderator intervened less, allowing more independent exploration, but this also left participants stuck or uncertain for longer periods. In M4, communication barriers and reduced visibility made it harder to judge when to intervene, leading to delayed support or increased participant reliance.

These differences show that the moderator's role is shaped by the method rather than fixed. In more controlled settings, neutrality can be maintained more strictly, whereas proxy and remote contexts often require a more active role to sustain the interaction. This raises a key question: when does intervention support a fair and meaningful test, and when does it begin to shape the results?

Rather than a clear boundary, the findings point to a continuous negotiation, where the moderator constantly adjusts between stepping in and holding back to reduce participant stress, maintain progress, and preserve test integrity. In this sense, neutrality is not a fixed state but something actively managed throughout the interaction.

The moderator thus not only facilitates the test but also shapes both the participant experience and the outcomes. The challenge is not to eliminate this influence, but to understand and manage it.

6.3 Theoretical Interpretation

The results both align with and challenge the theoretical perspectives presented in the theory chapter. Rather than fitting neatly into each framework, the findings reveal tensions between what the theories suggest and how interaction unfolded in practice. Across methods, usability testing does not merely capture interaction but actively shapes how cognition, communication, and experience emerge.

From a Distributed Cognition perspective, the proxy conditions functioned as distributed systems, where decision-making, instruction, and execution were shared between participant and moderator [2, 3]. While this supported task completion, it also introduced coordination as part of the cognitive work, reflected in increased task time, hesitation, and misunderstandings. At the same time, participants often sought direct control, suggesting that distributed interaction may be cognitively viable but not always intuitively accepted.

From a Situated Cognition perspective, the results confirm that interaction is shaped by context, with remote and on-site settings affecting communication, coordination, and perceived connection [4, 5, 35]. However, the findings also suggest that interaction structure plays a key role: remote proxy interaction could support strong connection (sometimes even stronger than on-site proxy) despite lack of co-presence, indicating that connection depends not only on physical context but on how interaction is organized and coordinated.

From an embodied interaction perspective, differences in control, confidence, and engagement can be linked to the absence of direct physical interaction in the proxy methods. Participants showed lower confidence, increased hesitation, and attempts to regain control, highlighting the importance of bodily engagement with artefacts [6, 14]. At the same time, embodiment did not disappear but shifted form, as interaction was mediated through the moderator's gestures, timing, and presence. This suggests that embodied aspects can be shared and redistributed across individuals, partially compensating for the lack of direct control. However, this compensation remained incomplete, as participants still expressed a need for direct interaction, particularly in tasks requiring precision or immediate feedback.

The three perspectives highlight different aspects of how the methods function. In proxy interaction, action is shared between participant and moderator, which enables task completion but requires continuous coordination. In remote settings, the lack of shared context makes interaction more dependent on communication and alignment. In direct interaction, participants act more intuitively, but their reasoning becomes less explicit. Rather than contradicting the theories, these findings show that each perspective captures only part of the interaction, while the methods themselves shape how it unfolds in practice.

6.4 Relation to Research Questions

Looking across the findings, the three research questions are closely connected. Rather than pointing to clear-cut answers, the results highlight a number of trade-offs between methods, particularly in terms of what is captured and how the interaction unfolds.

6.4.1 RQ1 – Reliability of Testing Methods

To what extent do different usability testing methods, varying in co-presence and interaction execution, produce comparable results in terms of reliability?

At first glance, the methods seem fairly comparable. Participants were generally able to complete the tasks, and the overall usability scores did not differ that much. In that sense, it would be easy to conclude that the choice of method does not matter too much.

However, looking more closely, this picture becomes more nuanced. While the outcomes were similar, the paths leading to them were quite different. The methods revealed different types of usability problems, and the same issues were sometimes experienced and interpreted in different ways. Direct interaction tended to expose more immediate, physical interaction issues, while proxy methods brought out problems related to interpretation and understanding of the user interface.

This makes reliability a bit more complicated than just comparing results. It is not only about whether participants succeed, but about what kinds of problems become visible and how they emerge. From that perspective, the methods are not fully interchangeable. Instead, they seem to complement each other, each highlighting different aspects of usability.

6.4.2 RQ2 – Proxy Interaction, User Behaviour, & User Experience

How does proxy interaction affect users' interaction behavior, verbalization patterns, and perceived experience compared to direct interaction?

Proxy interaction clearly changed how participants engaged with the system. Even though task success remained relatively stable, the experience of getting there was quite different.

One of the most noticeable differences was how participants expressed themselves. In the proxy methods, they had to describe what should be done, which led to more structured and explicit verbalizations. This made participants' reasoning and interpretation of the system more visible, since they had to explain not just what they did, but what they thought was possible.

At the same time, this also made the interaction slower and more effortful. Participants had to translate their intentions into words, which often led to pauses, hesitation, and reformulation. The interaction process became less about direct action and more about reasoning, coordination, and explanation.

What becomes clear across the results is that proxy interaction does not just change how users interact, it changes what interaction is. When action is replaced by instruction, the whole experience

shifts. It is not necessarily worse in terms of outcomes, but it appears different: slower, more effortful, and less direct. Users still appeared able to form an understanding of the system, but through explanation rather than exploration, which affected both confidence and interaction flow.

In contrast, direct interaction felt more immediate and intuitive. Participants explored the user interface more freely, and their understanding seemed to develop through action rather than explanation. This made the interaction feel more natural, but also meant that parts of their reasoning remained implicit.

So while proxy interaction does provide richer verbal insight, it also changes how the interaction is experienced and carried out during the test. It does not only reveal how users think and interact with the system, but also shapes how these processes unfold during the test.

6.4.3 RQ3 – Moderator Role Across Methods

How does the role and influence of the moderator differ across the tested usability methods?

The role of the moderator turned out to be more central than expected, especially in the proxy and remote methods.

In theory, the moderator is supposed to stay neutral, but in practice this was not always possible. Across all methods, situations came up where participants hesitated, got stuck, or seemed unsure. In those moments, the moderator had to decide whether to step in or hold back. Helping could keep the session moving, but also risk influencing the outcome.

This tension became especially clear in the proxy methods, where the moderator was directly involved in executing actions. In these cases, the moderator was not just observing, but actively shaping the interaction. Interpreting unclear instructions or deciding how to act introduced an additional layer where the moderator's decisions directly affected how tasks were carried out.

At the same time, this makes it difficult to treat the moderator as neutral. As soon as interaction is mediated, the moderator becomes part of how thinking, acting, and understanding unfold. What stood out was not only that the moderator influenced the interaction, but how differently that influence played out depending on the setting. In some situations, it created structure and support; in others, it introduced uncertainty or subtle steering.

In the remote methods, the situation was slightly different but just as challenging. Limited visibility and communication made it harder to judge what participants needed, which sometimes led to delayed support or increased dependence on the moderator.

Overall, the moderator's role was not fixed, but constantly adjusted depending on the situation. Rather than being a neutral observer, the moderator became part of the interaction itself, influencing both how the session unfolded and what results were produced.

Taken together, these findings suggest that methodological differences are not only shaped by technical setup or interaction format, but also by how the moderator participates in and influences the interaction process.

6.4.4 Overall Reflection

Across the results, there is also a clear tension. More integrated interaction, as in proxy settings, can create stronger alignment and even a sense of collaboration. At the same time, that same integration increases the moderator's influence, making it harder to separate user behavior from the testing setup itself. Less interaction, as in remote think-aloud, reduces this influence but instead leads to a more fragmented and less engaging interaction.

Taken together, this suggests that usability testing is less about observing a user interacting with a system, and more about observing a situation where user, moderator, and method are tightly intertwined. Proxy interaction makes this especially visible, but it is present across all methods.

This raises a broader question: if interaction changes this much depending on how it is mediated, to what extent are we actually evaluating the system, and to what extent are we evaluating the conditions under which it is used?

6.5 Relevance for Sustainable Development

The results point to remote usability testing as a potentially more sustainable alternative to traditional in-person testing, particularly from an environmental and resource-related perspective. By reducing the need for travel, these methods can lower both emissions and logistical costs.

From a social sustainability perspective, remote testing also opens up for broader participation across geographical contexts, which can support more inclusive design processes, especially for products targeting international user groups. These aspects were part of the motivation for including remote methods in the study.

At the same time, the findings show that these methods are not neutral substitutes. Remote and proxy-based setups changed how interaction unfolded, affecting communication, coordination, and the role of the moderator. While overall outcomes such as task success remained relatively stable, the experience of interaction and the types of insights generated differed across methods.

This points to a clear trade-off. More sustainable testing approaches can increase accessibility and reduce environmental impact, but they also influence what is observed and how it is interpreted. In that sense, choosing a more sustainable method is not only a practical decision, but also a methodological one, where different approaches highlight different aspects of usability.

6.6 Ethical Considerations

Ethical considerations such as informed consent, participant integrity, and data handling were central throughout the study, but the results also highlight how these aspects were shaped by the chosen methods. This became particularly visible in the remote conditions, where reduced visibility made it harder to assess participants' comfort and understanding, increasing the moderator's responsibility during the session.

The use of proxy interaction introduced additional ethical challenges in practice. Since the moderator performs actions on behalf of the participant, there is an inherent risk of influencing both the interaction and the outcomes. The findings suggest that this influence is not only methodological, but also affects how participants experience and engage with the interaction, for example through reduced confidence or increased dependence on the moderator.

This raises a broader ethical implication. If the method shapes how participants act and what becomes visible in the data, it also influences which aspects of usability and interaction are ultimately evaluated. Ethical responsibility in usability testing therefore extends beyond participant protection to include transparency about how methodological choices influence results.

Remote testing also introduces challenges related to participant authenticity. Compared to on-site testing, it is more difficult to verify that the interaction reflects independent user behavior, for example if participants receive external assistance. While this was not a central focus of the study, it highlights an additional consideration when interpreting results from remote sessions.

Finally, the significant gender imbalance among participants limits the inclusivity of the findings and highlights the importance of diversity as an ethical aspect of research design.

6.7 Further research

The findings point to several directions for future research. To begin with, future studies should include more diverse and representative participant samples, particularly in terms of gender, to improve the generalizability of findings.

One direction concerns further investigation of the moderator's role. The findings indicate that the moderator may influence both the interaction and the resulting data. Future research could therefore explore this influence in more detail, for example by examining how different ways of guiding sessions affect participant behavior and the identification of usability problems. This could include studies with multiple moderators, as well as closer analysis of how and when moderators intervene across different testing contexts.

Second, proxy-based interaction warrants further exploration, including how moderator influence can be minimized, how these methods perform across different user groups and task types, and how they function in more realistic settings using fully implemented products rather than prototypes. The use of Wizard-of-Oz techniques in remote proxy settings could also be explored, as they may enable flexible simulation of system behavior without requiring a fully implemented solution.

An additional direction concerns the potential use of proxy interaction for training purposes. Comments during the tests suggest that instruction-based interaction may support initial understanding of a system, raising questions about its effectiveness as a learning approach.

7. Takeaways for Practitioners

Based on this study, there is no single best usability testing method for *all* situations. Different methods are useful for different goals, and the choice of method directly affects the insights you gain. The key is therefore to choose the method based on what you want to learn. From the findings and discussion, several practical recommendations can be drawn.

If the goal is to understand *why* users struggle, proxy interaction can be valuable. Since the moderator performs the interaction at the participant's request, users often explain their intentions and reasoning more clearly. This can be especially useful when evaluating complex flows or unclear user behavior.

If the goal is to observe how users naturally behave or evaluate physical interaction, direct interaction is more suitable. Letting users interact independently makes it easier to identify issues related to intuitiveness, navigation, and real-world use.

Remote testing can broaden participation, but it requires careful test design. Direct remote testing requires clear structure and instructions so participants can proceed independently. Proxy-based testing can be easier to manage remotely since the moderator can more easily move the session forward, while also placing fewer demands on participants. This can simplify both recruitment and organization, and may be particularly useful when users need more support or when the goal is to better understand user reasoning rather than observe natural interaction. However, the method must match the purpose of the evaluation, otherwise, there is a risk of generating misleading insights.

Another important takeaway is that the moderator always influences the session. The key challenge is therefore not to remove that influence, but to manage it consciously.

Finally, participants in the remote proxy setup also experienced the session as a way to become familiar with the product. This suggests that proxy interaction may have value not only for evaluation, but also for onboarding and training purposes.

8. Conclusion

This study shows that usability testing methods do not simply reveal usability, but actively shape it. While overall outcomes such as task success and perceived usability were relatively similar across methods, important differences emerged in how interaction unfolded, how users reasoned, and what kinds of usability problems became visible.

Direct interaction primarily revealed functional and physical issues, such as touch sensitivity and system feedback, whereas proxy-based interaction highlighted problems related to terminology, user interface interpretation, and system logic. Although 67% of usability problems were identified in the first method, additional issues continued to emerge across subsequent methods, indicating that each method revealed different aspects of usability and contributed complementary insights.

Proxy interaction generated more explicit and structured verbalizations, making users' understanding of the system more visible. At the same time, it increased cognitive load, slowed down interaction, and introduced risks of misinterpretation between participant and moderator. In contrast, think-aloud methods supported more natural and direct interaction, but left more of the reasoning implicit and required greater interpretation in the analysis.

The role of the moderator proved central across all methods. In proxy-based setups, the moderator became an active part of the interaction, directly influencing how tasks were carried out. In remote settings, reduced visibility and control made it more difficult to assess participant needs and establish a shared understanding, particularly in remote think-aloud, while this was less pronounced in the remote proxy condition where interaction was structured through instruction.

Taken together, these findings show that usability testing methods are not interchangeable. Rather than producing the same type of results, they generate different insights, shape interaction in distinct ways, and influence the relationship between participant and moderator. Usability findings cannot therefore be separated from the method used to produce them, as each method highlights certain aspects while making others less visible.

This has important implications for practice. Choosing a method is not only about how to conduct a test, but about what kind of knowledge can be generated. Each method comes with trade-offs that shape both user behavior and the resulting data, and these should be considered as part of the study design rather than as limitations identified afterwards.

Finally, the findings highlight that the moderator cannot be treated as a neutral observer. Instead, moderation involves a continuous balancing between supporting participants and maintaining reliable results. Rather than striving for complete neutrality, the moderator's influence should be acknowledged as an integral part of how usability testing unfolds, rather than something that can be fully controlled or eliminated. Taken together, this challenges the idea of usability testing as a neutral evaluation of a system.

References

- [1] ASSA ABLOY;. Accessed 28 Apr 2026. Available from: <https://www.assaabloy.com>.
- [2] Hutchins E. *Cognition in the Wild*. Cambridge, MA: MIT Press; 1995. Available from: doi.org/10.7551/mitpress/1881.001.0001.
- [3] Hollan J, Hutchins E, Kirsh D. Distributed cognition: Toward a new foundation for human–computer interaction research. *ACM Trans Comput Hum Interact*. 2000;7(2):174-96. Accessed 26 Apr 2026. Available from: doi.org/10.1145/353485.353487.
- [4] Brown JS, Collins A, Duguid P. *Situated Cognition and the Culture of Learning*. *Educ Res*. 1989;18(1):32-42. Accessed 26 Apr 2026. Available from: doi.org/10.3102/0013189X018001032.
- [5] Lave J, Wenger E. *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press; 1991. Accessed 26 Apr 2026.
- [6] Dourish P. *Where the Action Is: The Foundations of Embodied Interaction*. Cambridge, MA: MIT Press; 2001. Accessed 26 Apr 2026.
- [7] Ericsson KA, Simon HA. *Protocol Analysis: Verbal Reports as Data*. Revised ed. Cambridge, MA: MIT Press; 1993. Available from: doi.org/10.7551/mitpress/5657.001.0001.
- [8] Nielsen J. *Thinking Aloud: The #1 Usability Tool*; 2012. Accessed 8 Jan 2026. [Blog post, Nielsen Norman Group]. Available from: <https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/>.
- [9] Tullis T, Fleischman S, McNulty M, Cianchette C, Bergel M. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. In: *Proceedings of the Usability Professionals Association Conference*; 2002. .
- [10] Thompson KE, Rozanski EP, Haake AR. Here, There, Anywhere: Remote Usability Testing That Works. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM; 2004. p. 1325-8. doi:10.1145/985921.986094.
- [11] Andreasen MS, Nielsen HV, Schröder SO, Stage J. What happened to remote usability testing?: an empirical study of three methods. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. New York, NY, USA: Association for Computing Machinery; 2007. p. 1405-14. Accessed 26 Apr 2026. Available from: doi.org/10.1145/1240624.1240838.

-
- [12] Wood R, Dixon E, Elsayed-Ali S, Shokeen E, Lazar A, Lazar J. Investigating best practices for remote summative usability testing with people with mild to moderate dementia. *ACM Trans Access Comput.* 2021;14(3). doi:10.1145/3460942.
- [13] United Nations. Sustainable Development Goals; 2015. Available from: https://sdgs.un.org/#goal_section.
- [14] Kirsh D. Thinking with External Representations. *AI & Society.* 2010;25:441-54. Available from: doi.org/10.1007/s00146-010-0272-8.
- [15] Chiappe D, Vervaeke J. Distributed Cognition and the Experience of Presence in the Mars Exploration Rover Mission. *Frontiers in Psychology.* 2021;12:689932. doi:10.3389/fpsyg.2021.689932.
- [16] Ligorio MB, Cesareni D, Schwartz N. Collaborative Virtual Environments as Means to Increase the Level of Intersubjectivity in a Distributed Cognition System. *Journal of Research on Technology in Education.* 2008;40(3):339-57.
- [17] Rubin J, Chisnell D. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* 2nd ed. Indianapolis, IN, USA: Wiley Publishing, Inc.; 2008.
- [18] Denscombe M. *The Good Research Guide: For Small-Scale Social Research Projects.* 4th ed. Maidenhead, England: Open University Press / McGraw-Hill Education; 2010.
- [19] Nielsen J, Landauer TK. A Mathematical Model of the Finding of Usability Problems. In: *Proceedings of the 11th ACM SIGCHI Conference on Human Factors in Computing Systems.* Amsterdam, The Netherlands: ACM; 1993. p. 206-13. Available from: doi.org/10.1145/169059.169166.
- [20] Jacobsen NE, Hertzum M, John BE. The evaluator effect in usability studies: problem detection and severity judgments. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.* 1998;42(19):1336-40. doi:10.1177/154193129804201902.
- [21] Bruun A, Gull P, Hofmeister L, Stage J. Let Your Users Do the Testing: A Comparison of Three Remote Asynchronous Usability Testing Methods. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Boston, MA, USA: Association for Computing Machinery; 2009. p. 1619-28. Available from: doi.org/10.1145/1518701.1518948.
- [22] Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. *International Journal of Medical Informatics.* 2009;78(5):340-53. doi:10.1016/j.ijmedinf.2008.10.002. Available from: doi.org/10.1016/j.ijmedinf.2008.10.002.
- [23] Van den Haak MDT, Maaikje J De Jong, Schellens PJ. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with Computers.* 2004;16(6):1153-70. doi:10.1016/j.intcom.2004.07.007.
- [24] McDonald S, Edwards HM, Zhao T. Exploring Think-Alouds in Usability Testing: An International Survey. *IEEE Transactions on Professional Communication.* 2012;55(1):2-19. doi:10.1109/TPC.2011.2182569.
-

- [25] Van den Haak MJ, De Jong MDT, Schellens PJ. Constructive Interaction: An Analysis of Verbal Interaction in a Usability Setting. *Behaviour & Information Technology*. 2006;25(4):339-51. Available from: <https://ris.utwente.nl/ws/files/249702798/VandenHaak2006constructive.pdf>.
- [26] Krahmer EJ, Ummelen N. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*. 2004;47(2):105-17. doi:10.1109/TPC.2004.828205.
- [27] Castillo JC, Hartson HR, Hix D. Remote usability evaluation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '98)*. New York, NY, USA: Association for Computing Machinery; 1998. p. 253-4. Available from: doi.org/10.1145/286498.286736.
- [28] Nielsen J. Why You Only Need to Test with 5 Users; 2000. Accessed 8 Jan 2026. [Blog post, Nielsen Norman Group]. Available from: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>.
- [29] Yin RK. *Case Study Research: Design and Methods*. 5th ed. SAGE Publications; 2014.
- [30] Kelley JF. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*. 1984;2(1):26-41. doi:10.1145/357417.357420.
- [31] Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 2006;3(2):77-101. doi:10.1191/1478088706qp063oa.
- [32] Nielsen J. Severity Ratings for Usability Problems. Nielsen Norman Group. 1994. Available from: <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>.
- [33] Nielsen Norman Group. How to Analyze Qualitative Data from UX Research: Thematic Analysis; 2022. Accessed 9 April 2026. Available from: <https://www.nngroup.com/articles/thematic-analysis/>.
- [34] Brooke J. SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. *Usability Evaluation in Industry*. Taylor & Francis; 1996. p. 189-94.
- [35] Rogers Y. *HCI Theory: Classical, Modern, and Contemporary*. Morgan & Claypool Publishers; 2012.

A. Script for the Usability Tests, incl. Tasks

See following pages.

Usability Test Spring 2026 – Script

Checklist (control panel) before the test:

- Time-out: 30 sec
- Brightness: 30% (Adjusted based on camera picture)
- User PIN: disabled
- Winter mode: activated
- Simulator: press END SHIFT:
 - Mode: Locked
 - Clear issues

Checklist (setup) before the test:

- Connect the camera to the computer – Change the setting in the computer to rotate the camera 180 degrees.
- Start Teams meeting
- Choose the HUE camera for the Teams meeting – perhaps unable the mirror setting
- Choose the correct microphone for the Teams meeting

Test person:

Test group:

Klockslag:

I'd like to start by saying thank you for taking the time to help us today.

My name is Frida, and I'm doing my master's thesis with the sliding door team here at Assa Abloy. Can you tell me a little bit what you're doing at Assa Abloy?

A big part of my work is to run user studies like this one, so I'm very glad we could organize this session today.

Before we begin, I want to clarify the purpose of the test:

We are testing the design — not you.

If something feels unclear, complicated, or unexpected, that's exactly what we want to know. There are no right or wrong answers, and all feedback is valuable.

Have you participated in a usability test before?

The whole session for this test will take about an hour, including this introduction, the tasks, and a few questions at the end. If something unrelated comes to mind, that's fine — I'll make a note and we can come back to it later.

For usability reasons, we will record audio and video of **your hands only**. Your face or other identifying features will not be recorded. You have already consented to this in the form, but I want to double-check that you're still comfortable with it.

If you need a break or want to stop at any time, just let me know — no explanation needed.

Do you have any questions so far?

Let me briefly show you what we'll be working with:

In front of us is a control panel that operates sliding doors, similar to those used at building entrances. Do you know what these control panels normally do?

If no: they simply control how the door behaves — for example open, closed, or automatic. The panel also includes a few additional functions, which you'll get to explore during the test.

This one will be instead of the older door mode selectors, like key-turn versions or physical button panels, and will be part of the next generation of sliding doors currently under development.

And next to it is a simulator which is just meant for this test, and we'll only use it when a task specifically tells us to.

Does everything feel clear so far?

When we start, I'll read one task at a time. After each one, just do what you think the task is asking for, and let me know when you feel that you're done.

Method 1 & 4:

During the test, I'd like you to think aloud. Are you familiar with this technique?

It is to say what you're thinking and why you're doing what you're doing. For example: "I'm choosing this light bulb icon because it looks like a clue," or "I'm not sure what to do here, so I'll try the menu."

And feel free to share any impressions along the way, such as if something looks strange or well designed, if anything is confusing, or if the system behaves differently than expected. There are no right or wrong thoughts — this just helps us understand your reasoning.

Do you have any questions before we begin?

Then I will start the recording.

Method 2 & 3 (proxy):

*During the test, **I want you to tell me where to press on the control panel.***

*You will guide the interaction, and **I'll be the one who actually presses the buttons.***

If I misunderstand your instructions, just correct me — that's completely fine.

I'd also like you to think aloud. Are you familiar with this technique?

It means saying what you think and why you take certain actions. For example: "I'm choosing this light bulb icon because it looks like a clue," or "I'm not sure what to do, so I'll try the menu." You can also mention anything that seems strange, well designed, confusing, or unexpected. There are no right or wrong thoughts — it just helps us understand your reasoning.

Do you have any questions before we begin?

Then I will start the recording.

START THE RECORDING

I will now read a short background story. It gives a bit of context for why you're using the control panel. You don't need to remember the details—it's just to help you get a feel for the situation.

*“You are working as a **shift supervisor in a clothing store** located on a busy street. You care a lot about customer service, and it’s important to you that the store feels welcoming and that customers have a pleasant experience.*

*Your shifts are usually **hectic** — you assist customers, handle deliveries, and are often the main person responsible in the store. When problems appear, you want to deal with them as smoothly as possible so you can stay focused on the customers.*

Today is a typical spring morning, with brighter and warmer days. You arrive at the store before opening. You unlock the door, step inside, turn off the alarm, and lock the door again while you prepare the store for the day.

*You keep your **phone close**, since delivery staff or colleagues may need to reach you at any time during your shift.”*

Before we start with the tasks: What can you understand from the initial screen, without interacting with it? Tell me what information you can understand or what you think the screen allows you to do.

I will now begin to read short tasks for you, one at a time. When I have finished reading the task, I want you to solve the task while thinking aloud. Let me know when you feel like you are done with the task.

Klockslag:

Tasks:

Task 1:

☐

“You are preparing the store for opening and customers will soon arrive. To ensure a smooth start to the day, adjust the door settings so that it opens when customers pass through and closes afterward.”

Task 2:

☐

“You have noticed that the control panel stays active for longer than you would like before returning to the initial screen. Adjust the control panel so it remains logged in for 20 seconds.”

Give out Admin PIN

Task 2.5:

☐

“While standing by the entrance door you notice the spring air has arrived, and that means winter mode is no longer required. You therefore turn off the winter mode”.

Task 3:

☐

“The beginning of your shift has been busy, and it’s now time for your lunch break. Since you’ll be working alone for the next few hours, the routine is to close the door for a 30-minute lunch in the staff room. Please make the necessary changes so the door stays closed during your break.”

Task 4:

☐

“During your lunch break, a delivery driver arrives with a trolley of clothes for tomorrow’s sale. You need to open the door once to let him in while keeping the door in locked mode. Show how to perform a one-time door opening without changing the mode.”

Follow-up question: What would you expect to happen after pressing this button?

Task 5:

☐

“The store has introduced a new rule: only shift supervisors are allowed to handle the entrance door. Because of this, a user PIN must now be added to the control panel. Please show how you would create a user PIN, using any number combination of your choice.”

Task 6:

- ☐ *“Your lunch break has ended, so you switch the door mode back to ‘Auto’. Right after doing so, you change your mind since the weather has warmed up, and instead set the door to the mode that keeps the door leaves continuously open.”*

Two actions must’ve been made!

I have a few follow-up questions I’d like you to reflect on:

1. When do you need to enter a PIN code?
2. What do you have to do in order to stay logged in? What do you have to do if you *don’t* want to stay logged in anymore?
3. What PIN code can you enter to make the changes you need?

If ‘Open’ not found: Unfortunately, you selected the incorrect mode. I want to you to see if you can find a mode called ‘Open’.

Task 7:

- ☐ *“A colleague calls and asks you to test her old user PIN now that the PIN setting is enabled. She wants you to try the combination ‘1996’ and not give up. Switch to another mode to see if the PIN works and then switch back again.”*

Task 8:

- ☐ *“A few hours later, a colleague joins your shift. While you’re unpacking the earlier delivery in the storage room, she comes in and says the door is acting strangely and needs your help. In the simulator, press ‘Help colleague’ to begin.”*

If logged in: “Before starting the next task, please go back to the initial screen.”

Task 9:

- ☐ *“You would like to understand if there’s anything you can do to resolve this issue. Since you don’t want to troubleshoot on your own, you decide to contact a service technician for guidance. Begin by looking for their contact information so you can call them.”*

Task 10: *Open – Or other*

“You successfully called a service technician, who instructed you to start looking into the matter by pressing ‘Troubleshoot’ on the simulator.”



“After doing so, they tell you to change the door mode to ‘Locked’ and then to ‘Auto’ and asks if the error remains. Please try as suggested and report back.”

“Your shift has come to an end, and the next shift supervisor arrives and thanks you for being such a helpful colleague throughout the day.”

Stop the recording.

Klockslag:

Thank you — you’ve now completed all the tasks in the session and I have stopped the recording.

You can now switch back the camera if you would like.

Before we finish, I’d like you to answer a few final questions. You’ll find them in the questionnaire, and you can fill it in yourself. If you have any questions at any point, feel free to ask.

Dubbelkolla att debriefingen är inskickad

Then I just want to say thank you so much! I really appreciate that you took the time to participate, complete all the tasks, and answer all my questions. It’s been great running this test with you!

If you have any questions later, feel free to message me on Teams or email. I may also reach out with a follow-up question — Is that okay with you?

Thank you!

Klockslag:

B. Debriefing Questions - Participant

See following pages.

2. Please fill in the option that suits the most for each statement. (System Usability Scale)

*

	Strongly disagree	Disagree	Not sure	Agree	Strongly
I think that I would like to use this system frequently.	☐	☐	☐	☐	☐
I found the system unnecessarily complex.	☐	☐	☐	☐	☐
I thought the system was easy to use.	☐	☐	☐	☐	☐
I think that I would need the support of a technical person to be able to use this system.	☐	☐	☐	☐	☐
I found the various functions in this system were well integrated.	☐	☐	☐	☐	☐
I thought there was too much inconsistency in this system.	☐	☐	☐	☐	☐
I would imagine that most people would learn to use this system very quickly.	☐	☐	☐	☐	☐
I found the system very cumbersome to use.	☐	☐	☐	☐	☐
I felt very confident using the system.	☐	☐	☐	☐	☐
I needed to learn a lot of things before I could get going with this system.	☐	☐	☐	☐	☐

3. Please rate the following additional usability statements specific to this product. *

	Strongly disagree	Disagree	Not sure	Agree	Strongly
It was easy to identify which elements were clickable.	☐	☐	☐	☐	☐
The symbols and icons were confusing.	☐	☐	☐	☐	☐
It was hard to understand where to go next.	☐	☐	☐	☐	☐
I felt confident that I could use the system without accidentally causing any problems.	☐	☐	☐	☐	☐

4. Is there anything in the design that you particularly liked? If so, please describe why.

5. Is there anything that felt unclear, confusing, or in need of improvement? Please explain.

6. If you had to teach someone else how to use this system, what would you tell them first? *

7. Based on your experience with customers and similar products, did the system behave in a way you would expect? Please explain. *

8. Is there anything else you'd like to add?

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

 Microsoft Forms

C. Debriefing Questions - Moderator

See following pages.

Distributed cognition

Fokus: kognition distribuerad mellan personer, artefakter, miljöer och system. Kommunikation, koordinering och informationsflöde.

3. Likert-frågor *

	Stämmer inte alls	Stämmer knappt	Neutral	Stämmer delvis	Stämmer helt	N
RQ1: Jag hörde tydligt vad testpersonen sa	☐	☐	☐	☐	☐	(
RQ1: Jag kunde enkelt förstå vad testpersonen menade	☐	☐	☐	☐	☐	(
RQ1: Testutrustningen fungerade som den skulle.	☐	☐	☐	☐	☐	(
RQ1: Mycket gick fel under detta test.	☐	☐	☐	☐	☐	(
RQ1: Jag behövde ofta upprepa mig	☐	☐	☐	☐	☐	(
RQ1: Jag hade bra sikt över vad testpersonen gjorde på skärmen (1 och 4)	☐	☐	☐	☐	☐	(

4. RQ2: Uppstod missförstånd eller koordinationsproblem när deltagaren instruerade mig? *

5. RQ2: Använde deltagaren ord, pekningar, metaforer eller spatiala referenser när de instruerade mig? *

Använde den de när den tänkte högt?

6. RQ2: Uppstod situationer där deltagaren försökte ta över skärmen spontant, även om uppgiften var att instruera mig? När? *

(2 och 3)

Situated Cognition

Fokus: hur handlingar påverkas av den specifika situationen, miljön och den sociala kontexten.

7. Likert-frågor *

	Stämmer inte alls	Stämmer knappt	Neutral	Stämmer delvis	Stämme
RQ3: Jag hann förbereda allt i god tid.	☐	☐	☐	☐	☐
RQ3: Jag upplevde att välkommandet fungerade bra.	☐	☐	☐	☐	☐
RQ1, RQ3: Det var enkelt för testpersonen att komma igång.	☐	☐	☐	☐	☐
RQ3: Testpersonen verkade trygg med testets upplägg.	☐	☐	☐	☐	☐
RQ3: Jag kände en bra kontakt med testpersonen	☐	☐	☐	☐	☐
RQ1, RQ3: Testmiljön kändes lugn och utan störningar.	☐	☐	☐	☐	☐
RQ1: Det fanns tillfällen då jag kände att testmiljön påverkade deltagarens prestation negativt.	☐	☐	☐	☐	☐
RQ1, RQ2: Det kändes som att testet var givande	☐	☐	☐	☐	☐
RQ3: Det tog mycket energi att utföra just detta test.	☐	☐	☐	☐	☐

8. RQ1, RQ3: Var det något i miljön, mitt kroppsspråk eller sättet jag gav instruktioner som jag tror formade hur deltagaren utförde uppgifterna?

9. RQ1: Hur upplevde jag deltagarens närvaro medan hen fyllde i sitt debriefingformulär (t.ex. fokus, engagemang, tystnader, osäkerhet, distraktioner)? *

Embodiment

Fokus: kroppslighet, motorik, fysisk närvaro, gestik, sensoriska villkor, interaktion via proxy.

10. Skolor (1-5) *

1 = Inte alls, 5 = I mycket hög grad

	1	2	3	4	5
RQ3: I vilken grad tror jag att min närvaro påverkade deltagarens sätt att resonera högt (think-aloud)?	☐	☐	☐	☐	☐
RQ3: Hur väl lyckades jag hålla fast vid min neutralitet som testledare?	☐	☐	☐	☐	☐
RQ2, RQ3: Hur ofta gav jag oavsiktliga ledtrådar (verbalt eller icke-verbalt)?	☐	☐	☐	☐	☐

11. RQ1, RQ3: Beskriv ett tillfälle där jag tror att min närvaro påverkade deltagarens beteende eller problemlösning. *

12. RQ2, RQ3: I vilka situationer verkade deltagaren anpassa sig mer till mig än till systemet? *

13. RQ2: Hur påverkades deltagarens rytm och flöde när de behövde "översätta" sina motoriska handlingar till verbala instruktioner? *

(2 och 3)

D. Usability Problems with Severity Scores

See following pages for complete list of usability problems.

Severity score was based on following guidelines:

Kritisk/Critical (Severity 4) Användaren kan inte slutföra sin uppgift.

- Blockerar huvudflödet
- Leder ofta till avbrott eller felaktiga beslut
- Inga rimliga workarounds

Allvarlig/Major (Severity 3) Användaren kan slutföra uppgiften, men med stor svårighet.

- Skapar förvirring eller leder till fel
- Kräver gissning, extra steg eller instruktioner
- Påverkar många användare

Mindre/Minor (Severity 2) Problemet är irriterande men hanterbart.

- Påverkar effektivitet eller användarupplevelse
- Lätt att ta sig runt
- Drabbar inte alla

Kosmetisk/Cosmetic (Severity 1) Påverkar inte användbarheten i någon större utsträckning.

- Visuella eller språkliga brister
- Ingen påverkan på uppgiftens genomförande

Usability Problems and Severity Score

	Tema	UP	Kommentar		M1	M2	M3	M4
					Classic think-aloud	Proxy on site	Remote proxy	Remote think-aloud
1	Auth.trig.	"Authenticated Trigger" saknar lämplig feedback.	1 / 4 - T.ex. Pop-up eller animation: Dörrar som öppnas, hänglåset låses upp. 4 - 2st tror det är fel knapp, 2st verkar säkra ändå. 3 - Knappen blir grön när man trycker på den. Den ger en pop.up.		4 - Användare förstår inte vad knappen gör.	4 - Användare förstår inte vad knappen gör.	4 - Användare förstår inte vad knappen gör.	3 - Skapar förvirring och osäkerhet.
2	Auth.trig.	Authenticated trigger är svår att hitta.	1/2 - Användare letar bland modes. 2 - Föreslår större ikon eller att den har funktion på illustrationen av dörrarna (visas genom symbol bredvid). Eller annan design på pilarna. 3 - Key impulse, Last man button, handen: borde visa en nyckel. 4 - Utseendet av knappen ser ut som en meny-knapp.		4 - Leder till att användare byter dörrläge istället.	4 - Leder till att användare byter dörrläge istället.	4 - Användare har olika associationer för hur knappen ska se ut: Hittar den inte.	2 - Samtliga trycker där men är förvirrade pga ingen feedback.
3	Inloggad	Användare saknar tydlig feedback kring om de är inloggade eller inte.	1 - Dörrlägena har inget hänglås så som i settings. 2 - Tror man är utloggad när man byter till Locked. 3 - Noterar inte hänglåsen i settings. Glömmer snabbt vad som krävs.		4 - Blockerar huvudflödet, och leder till många misstag bland majoriteten av användarna.	3 - Påverkar effektiviteten samt skapar förvirring gällande inlogg.	3 - Skapar förvirring.	4 - Blockerar huvudflödet, och leder till många misstag bland majoriteten av användarna.
4	Reset	"Reset" fungerar inte och saknar feedback.	3 - Användare trycker på knappen flera gånger trots att ignet händer. Tror att knappen fungerar men att annat gör att den inte funkar. Borde ge en pop-up om "Reset lyckades eller failed".		4 - Saknar implementation.	3 - Behöver åtgärdas. Men testarna rör den inte ofta och nämner inte.	3 - Användare blir förvirrade men tror att knappen fungerar ändå.	3 - Användare blir förvirrade men tror att knappen fungerar ändå.
5	Inloggad	Saknar tydlighet om hur man loggar in och ut, samt när/Konceptet inloggad är abstrakt.	1/2/3/4 - Användare går in till Settings. 4 - Uttrycket matchar inte användarnas förväntan.		4 - Skapar stor förvirring och stämmer inte med den mentala modellen.	2 - Skapar förvirring, men drabbar inte alla.	2 - Skapar förvirring, men drabbar inte alla.	4 - Skapar stor förvirring och stämmer inte med den mentala modellen.
6	PIN	Otydligt vad "User PIN" innebär.	2 / 4 - Tror admin pin har ändrats vid skapande av User pin. Tror man kan skapa flera. 3 - Verkar tror att man har ändrat koden och tror man kan slå in user PIN i settings. 4 - Tror att user pin och att pinskärmen dyker upp är olika funktioner.		4 - Otydligheten skapar en mental modell som inte överensstämmer med systemet.	2 - Leder till förvirring.	2 - Leder till förvirring.	3 - Leder till förvirring och kan inte slutföra uppgift.
7	PIN	Otydligt hur många siffror "User PIN" ska vara.	1 - Knappsatsen visas touch-event. 3 / 4 - Tror initieellt 6 siffror, men ser att det bara registreras 4. 4 - En misslyckas, men I R skapar han med 4 siffror.		4 - 4/5 missförstår vad deras pinkod är och får problem med adminkoden.	3 - Skapar problem vid bytande av dörrlägen.	2 - Användare kan lösa uppgiften.	2 - Användare kan lösa uppgiften.

Usability Problems and Severity Score

8	PIN	"Disable User PIN" har ingen error prevention.			3 - Kräver extra steg och risk för att återskapa "fel" pinkod.		3 - Kräver extra steg och risk för att återskapa "fel" pinkod.	4 - Användare förstår inte vad som har hänt.
9	PIN	Saknar instruktioner och feedback i processen av att skapa en user PIN.	1 - Avsaknad av hänglåset på "Enable User PIN". Otydligt om User PIN har sparats eller avbrutits när man går tillbaka. (när man skapar). 3 - Får inget meddelande om vad som har hänt. Skriver in vald User PIN istället för Admin PIN. Osäker när pinkoden har sparats. 4 - 4/5 Skriver in vald User PIN istället för Admin PIN. En behöver ledtråd pga han förstår inte varför det inte går.		4 - Användaren vet inte vad som har hänt.		3 - Användare misslyckas med uppgiften.	3 - Skapar förvirring och extra steg. 1 användare misslyckas med uppgiften.
10	Touch	Bekräftelseknappen i pinskärmen är svår att trycka på.	2 - Testpersonerna anmärker inte på det. 4 - En säger att touch sensitiviten är dålig och inte lika effektiv som en smartphone.		4 - 2/5 tror att deras pinkod är fel och kan inte logga in.	1 - Upptäcks endast en gång av testledare.		4 - Återupprepanande tillfällen, 1/5 tar sig inte vidare/tror koden är fel.
11	PIN	Otydligt vilken kod som behövs för pinskärmen för att ändra modes.	1 - 4-6 siffror kan registreras. 2 - Testar först user PIN, men byter till admin om den inte funkar. Tror User PIN bara funkar för vissa mdoes.		4 - Leder ofta till avbrott eller felbeslut.	2 - Drabbar inte alla, och lätt att sig runt (med admin).		2 - Drabbar inte alla, och lätt att sig runt (med admin).
12	Reset	Det finns ingen möjlighet att reseta systemet om det inte visas någon felkod.	Om det är något fel på kontrolpanelen och meddelanden förviner och kommer snabbt, hur resetar vi?					4 - Inga rimliga workarounds.
13	PIN	Ointuitivt när pinskärmen dyker upp och inte.	3 - 4 - Användare testar för att se om de lyckades spara sin User PIN.		3 - Användare förväntar sig annat beteende. Vissa vänjer sig, vissa fortsätter vara förvirrade.	2 - Förvirrar användare men de flesta kan lösa det.	3 - Skapar förvirring och användare får problem med att avklara uppgiften.	3 - Skapar förvirring och användare får problem med att avklara uppgiften.
14	Modes	Otydligt vad One-Way Locked och Auto Locked gör.	1 - Pga namn och hänglås i ikon, Användare tror att One-Way Locked och Auto Locked stänger dörren. 3 - Tror Auto Locked stänger dörren. 4 - Tror det är "always locked".		3 - Användare ska inte behöva gissa om dörren är låst eller inte.	3 - Skapar förvirring och fel beslut.	3 - Drabbar få, men leder till fel beslut och potentiell säkerhetsrisk.	2 - Skapar mindre förvirring, drabbar få.
15	PIN	Otydliga namn och funktioner i User settings-menyn.	Mental modell om User PIN. 1 - Ordet "Enable" är inte helt självklart för "Enable User PIN". 3 / 4 - Otydligt vad "disable user pin" innebär och förmedlar. Väljer Change admin PIN istället för att skapa User PIN. 4 - De förstår inte skillnaden av admin pin och user pin. Svåra att komma ihåg under testets gång.		2 - Användare löste uppgiften ändå.	3 - Användare löste uppgiften endast för att admin pin ej fungerar.	3 - Användare löste uppgiften endast för att admin pin ej fungerar. Tar längre tid att slutföra uppgifter.	3 - Förvirrad över Disable-knappen.

Usability Problems and Severity Score

16	Winter mode	Winter mode har fel pop-up meddelande.			3 - Skapar förvirring.	3 - skapar förvirring.	2 - Drabbar endast en, som tror att det är fel inställning.	2 - Drabbar endast en, som tror att det är fel inställning.
17	Modes	Förväntar sig ytterligare funktion av modesikonerna i hemskärmen.	1 - Trycker på modes-ikonerna för att expandera till en meny. 3 - Händer ofta, och de verkar tro att dörren både öppnas och stängs, och att något speciellt sker när man trycker på knapparna. 4 - Verkar trycka där för att återgå dörrläget eller öppna dörren via.		2 - Lätt förvirring och snabbt åtgärdat.	2 - Lätt förvirring och snabbt åtgärdat.	3 - Påverkar många användare, och kräver extra steg.	2 - Drabbar få och är lätt åtgärdat.
18	Change Admin pin	"Change Admin PIN" fungerar inte och saknar feedback.	3 - Den ger dock touchevent.		3 - Kräver gissning för att ta sig vidare.	3 - Kräver gissning för att ta sig vidare.	3 - Kräver gissning för att ta sig vidare.	
19	Auth.trig.	Authenticated Trigger trycks på i fel sammanhang.	2 - De har olika idéer om vad knappen ska göra. 3 - Väljer istället för "Open mode". 4 - Tror att knappen ska leda till en meny pga knappens utseende (refererar till smartphones där pilarna indikerar meny).		2 - Drabbar inte alla.	3 - Påverkar de flesta och skapar förvirring.	2 - Som följd av annat UP.	2 - Skapar förvirring och osäkerhet.
20	Error	Otydligt vilket errormeddelande som är mest akut.	3 - 3 st beskriver vad meddelanden gör och refererar inte till färg/form. Tror att varningstriangeln innebär att systemet "gör checkar".		2 - Drabbar inte alla.		3 - De flesta användarna svarar fel.	3 - De flesta användarna svarar fel.
21	Modes / scroll	Scrollfunktionen saknar tydlig affordance.	Användare missar att man kan scrolla. 2 - Tycker att signifier är ett måste, t.ex. pilar. 3 - Hittar slider pga den blå baren. Saknar dock tydliga signifiers. 4 - en behöver ledtråd, de andra 4 hittar lätt.		2 - Påverkar huvudfunktionen. Kräver ledtråd för 2/5.	3 - Påverkar många användare. Svårigheter med att klara av uppgiften.	2 - 2/5 ser direkt. 1/5 klarar inte uppgiften.	1 - Påverkar knappt användare.
22	Locked mode	Otydligt vad man kan och inte kan göra i dörrläget Locked.	2 - tror inte man kan öppna med auth.trig. i Locked. Tror inloggning samt pinkodsnivå är kopplat till just det Locked. 3 - Osäker på om man behöver låsa med nycklar manuellt, eller om läget är stängd. Verkar tro att Locked gäller kontrollpanelen och inte dörrläget. Vad står på skärmen när man låser upp dörren manuellt med nycklar?			3 - Skapar förvirring och potentiell säkerhetsaspekt.	3 - Skapar förvirring och potentiell säkerhetsaspekt.	
23	PIN	Tror att bara vissa specifika dörrlägen kräver pinkod.					2 - Skapar förvirring.	3 - Skapar stor förvirring.

Usability Problems and Severity Score

24	Reset	Användare har olika tankar om vad som ska ske när man trycker på "Reset".	3 - Eftersom knappen finns på varje meddelande så tror användare att knappen gäller specifikt för det meddelandet.		3 - Otydligt vad knappen gör och saknar info.		2 - Skapar förvirring över vad knappen ska göra.	
25	PIN	Oklart när pinförsöken nollställs.	3 - Användare tror att det bara är 2 ggr pga att han gick till låsskärmen efter en gång.		3 - 2/3 får "Too many attempts" när det ser ut som att de försökt endast en gång.		2 - Användare missförstår funktion.	
26	PIN	Inkonsekvent skapande av User PIN som en konsekvens av att det är otydligt när man är inloggad.			3 - Kräver extra steg och skapar stor förvirring.	2 - Påverkar effektivitet, men lätt att ta sig runt.		
27	Modes	Mental modell över vad dörlägen har för funktion stämmer inte överens med hur det är.					3 - Huvudfunktion av kontrollpanelen. Leder till fel beslut.	
28	PIN	Otydligt om pinkod är aktiverat och hur den stängs av (som följd av terminologi)	4 - Tillsammans med otydlig pinförfarande så saknar användare bekräftelse på om pinkodsfunktionen är av eller på.					3 - Skapar förvirring och användare vet inte vad som har hänt.
29	Touch	Raderaknappen i pinskärmen är svår att trycka på.						3 - Påverkar effektivitet och upplevelse. Drabbar 4/5.
30	Auth.trig.	Authenticated Trigger har ingen pinskärm.			3 - Det är en som har nämnt detta. Ska det vara så?			
31	PIN	Otydligt och inkonsekvent rubrik för pinskärmen	Flera användare slår in 20 när de ska ändra timeout. Flera användare vet inte vilken kod som ska skrivas in.		2 - Skapar förvirring och extra steg.	2 - Lätt åtgärdat.	2 - Skapar förvirring och extra steg.	2 - Skapar förvirring och extra steg.
32	Modes	Förväntar sig funktion av bilderna i hemskärmen.	1- Trycker på hänglåset på Locked mode i hemskärmen för att låsa upp dörren. 2 - Försöker öppna dörren genom att hålla inne någonstans på dörrbilden, samt trycka på varningssymbolen på dörrarna.		2 - Lätt förvirring och snabbt åtgärdat.	2 - Lätt förvirring och snabbt åtgärdat.	2 - Lätt förvirring och drabbar endast få.	
33	Error	Errormeddelanden har otydliga beskrivningar.	Kan bli viktigt att se över i framtiden. 3 - Läser noggrannt.		2 - Drabbar inte alla.		2 - Användare läser noggrannt.	2 - Användare läser noggrannt.
34	Modes	Ikonen för manual är inte helt tydlig (open).	2 / 3 / 4 - användare trycker på den när de känner sig osäkra i tillfällen då de ska öppna dörren (open, auth.trig.)			2 - Drabbar inte alla, men gör att de fastnar i uppgiften.	2 - Skapar förvirring och leder till felbeslut, men drabbar få.	1 - Drabbar en, och åtgärdar det.
35	Låsskärm	Användare trycker på symboler i låsskärmen som inte leder någonstans.	1/2/3 - Framförallt hänglåset i "Locked". 2/3 - snöflingan. 4 - Hänglåset 2 ggr, samt errorsymbolen och snöflingan en gång.		1 - Påverkar inte uppgiften.	1 - Påverkar inte uppgiften.	2 - Drabbar många men åtgärdas lätt.	1 - Påverkar inte uppgiften.

Usability Problems and Severity Score

36	PIN	Förväntar sig hårdare säkerhet när man slår in fel pinkod.	2 - pga mental modell från iPhones. 3 - Förväntar sig att Admin måste åtgärda det, men kommer inget sådant meddelande. 4 - om admin-koden inte fungerar då ska ingenting fungera. Dvs, blir du utlåst med adminkoden så borde du inte kunna logga in med user PIN.			1 - Skapar ett försiktigare beteende.	2 - Användare missar korrekt funktion.	1 - Skapar mindre förvirring, drabbar få.
37	Info	Saknar info kring inställningar/Otydligt hur man får hjälp med kontrollpanelen.	3 - Inte helt säker på vad som finns i infoknappen.		2 - Sällan förekommet och kom lätt runt det.		2 - Användare får gissa sig fram för att lösa uppgiften.	
38	Inloggad	Oklart om "tycka på skärmen" också håller pinkoden inloggad.	Som följd av otydlighet kring om man är inloggad eller ej.		2 - Bipproblem av ett större problem.	2 - Användare har olika tankar. Men påverkar inte uppgiften.		
39	Error	Ej intuitivt att gå in på "More info".	3 - Tar en stund innan de går in på knappen, felsöker först. 4 - 3/5 har inga problem. 2/5 går hela vägen i nästa uppgift.				2 - Skapar lätt förvirring. De löser problemen oavsett.	1 - Tar ibland en stund men hittar dit tillslut.
40	Settings/Terminologi	Settings stämmer inte med funktioner på hemskärmen.	3 - Reduced width är på 100% när winter mode är på. Är ordet "Reduced" korrekt här?				2 - Skapar förvirring kring funktionen.	1 - Mindre funktionell förvirring.
41	Touch	Siffrorna i pinskärmen är svåra att trycka på.						2 - Påverkar effektivitet och upplevelse.
42	PIN	Oklart hur länge en pinkod är aktiv efter att man slagit in den.	4 - Användare tror att timeout styr hur länge den är aktiv.					2 - Drabbar få, skapar förvirring.
43	Modes	Saknar information om vilket dörrläge som aktiveras när man skriver in pinkod.						2 - Drabbar få, leder till osäkerhet.
44	Touch	Modes-knapparna är svåra att trycka på.			2 - Drabbar endast en, och löser sig snabbt.			
45	Winter mode	Snöflingan i låsskärmen är inte helt tydlig.	4 - Användare förstår direkt innebörd och funktion.		1 - Tydligt när de hör funktionen "Winter mode".	1 - Tydligt när de hör funktionen "Winter mode".	1 - Vissa kopplar direkt med dess funktion.	
46	Winter mode	Winter mode knappen har inte helt tydlig utseende (hemskärm).	3 - Snöflingan i låsskärm och hemskärm är inte samma (föredrar låsskärm). Grön ser ut som en blomma. 4 - Användare tycker att det är tydligt att grön är på och grå är av (trots fel pop-up).		1 - En nämner att grönt inte associeras med vinter och snö.		1 - Inkonsekvent design: skapar förvirring.	
47	Mental model	Användarna referar inkonsekvent vad som är "main screen" och menu.	4 - En kallar låsskärmen för "main display".		1 - Skapar förvirring kring terminologi, och visar att det inte är helt intuitivt.			1 - Språklig mental modell.
48	Information/hjälp	Finns ingen tillgång till info om dörrlägen i kontrollpanelen.	2 - Vill ha info om hur man öppnar dörren.		1 - Endast en som nämnt.	1 - Påverkar inte uppgiften.		

Usability Problems and Severity Score

49	Error	Otydligt vad errorsymbolen i låsskärmen indikerar	4 - En användare tror att errorsymbolen i låsskärmen betyder att kontrollpanelen inte funkar.					1 - Drabbar få, förstår sen.
50	Error	Inget sätt att nå gamla errormeddelanden.				1 - Påverkar inte uppgiften.		
51	Settings	Oväntat beteende när man sparar i Settings.	2 - Förväntas att man går tillbaka automatiskt.			1 - Drabbar endast en. Lätt att gå runt det.		
52	Settings	Saknar dörrlägesspecifika inställningar.	2 - Förväntar sig att settings-alternativ ändras beroende på vilket dörrläge man är i.			1 - "Nice to have", kommer mest från förvirringar skapat av annat.		

E. Data Analysis – Supporting Figures & Tables

M2 - Proxy on site

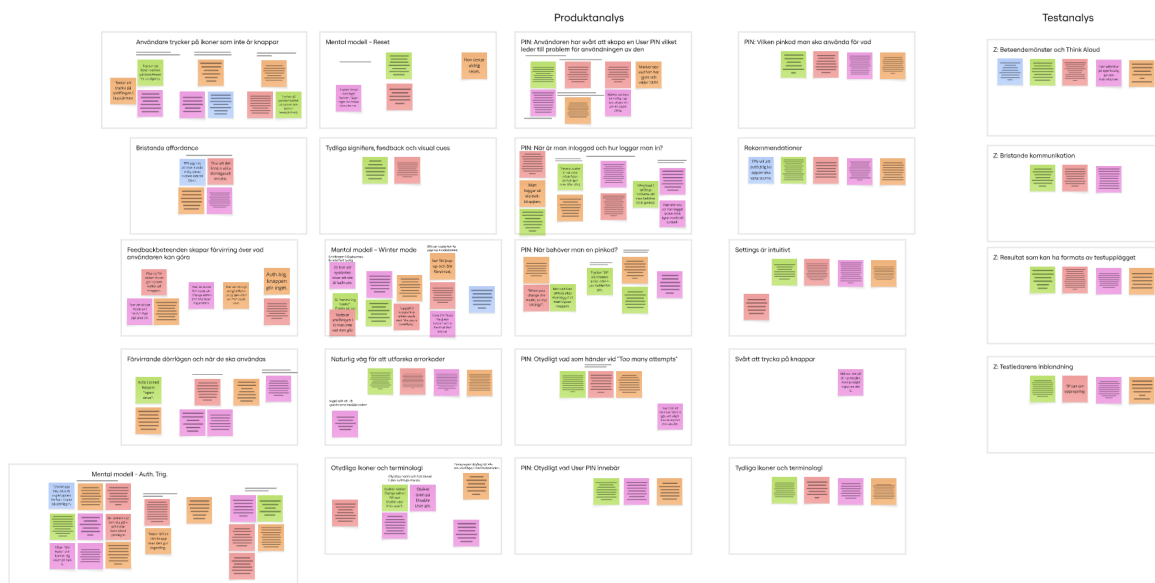


Figure E.1: Overview of the Miro board with regrouped post-its for Method 2. One test failed to record audio and was not included for this step, however, a few comments (blue notes) could be extracted from the diary.

M4 - Remote think aloud

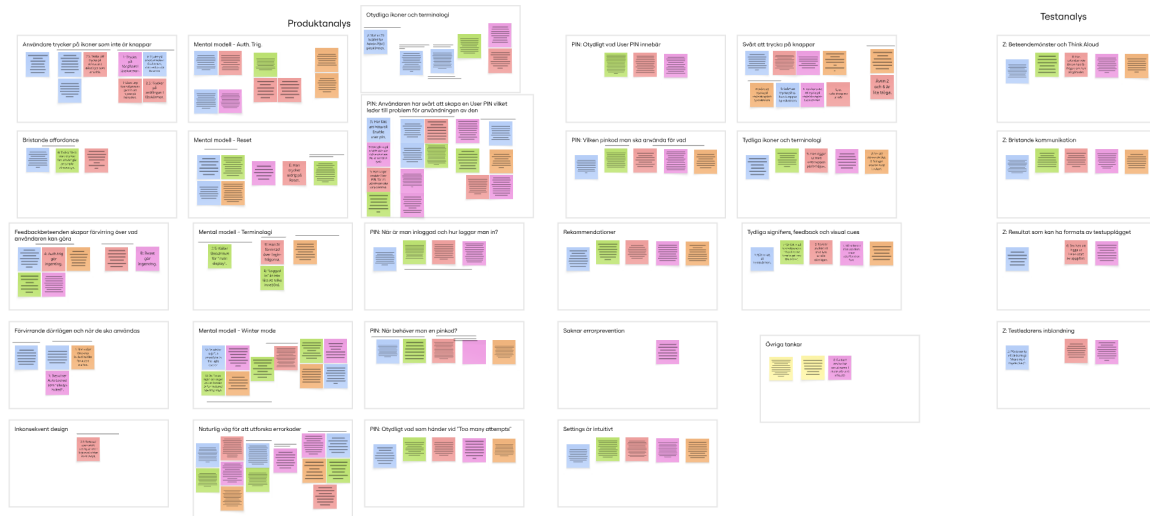


Figure E.2: Overview of the Miro board with regrouped post-its for Method 4.

Table E.1: Comparison of instructional strategies across methods (RQ2).

Dimension	M1	M2	M3	M4
Use of verbal labels	Literal icon naming; minimal elaboration	Frequent icon naming with varying precision	Frequent icon naming with richer visual descriptors	Very limited verbal labelling
Use of spatial references	Rare; mostly simple statements ("here")	Common; corner- and position-based cues	Common; more precise spatial cues ("second from the left")	Sparse; occasional basic positional cues
Use of pointing/gestures	Minimal pointing	Pointing used despite proxy format	Pointing observed even in remote view	Very rare gestures
Overall instruction richness	Low; simple literal descriptions	Medium; consistent mix of labels + spatial cues	High; labels + spatial cues + occasional metaphors	Very low; sparse communication overall

Table E.2: Comparison of how proxy interaction affects rhythm, cognition, and behavior (RQ2).

Dimension	M2	M3
Effect on rhythm and pace	Verbalization slowed participants down, creating pauses and disruptions in natural flow	Interaction became slower and more deliberate, with increased hesitation when phrasing actions
Cognitive load	Increased effort in word choice; precision demanded more mental resources	Participants became reflective, focusing more cognitively on each step than in direct interaction
Challenges in articulation	Difficulty describing icons or locations; trial-and-error in wording and occasional apologies	Some struggled briefly with terminology; others adapted but still paused to find correct phrasing
Emotional or behavioral effects	Caution replaced spontaneity; reduced confidence during uncertain descriptions	More careful and considered actions; flow disrupted by uncertainty or technical issues
Overall emphasis	Verbalization disrupted spontaneity and induced a more cautious, deliberate rhythm	Verbalization introduced a reflective, slower pace with moments of hesitation or friction

Table E.3: Comparison of situations in which participants adapted more to the moderator than to the system (RQ2/RQ3).

Dimension	M1	M2	M3	M4
Forms of adaptation to moderator	Adjusting pace, repeating steps, orienting device for moderator visibility	Waiting for cues, permission, or confirmation before acting	Softened instructions, polite phrasing, deferring exploration	Adjusting pace and communication based on moderator clarity and speed
Indicators of social alignment	Seeking validation; framing actions for the moderator rather than system	Leaning back or giving physical space to emphasize moderator's role	Holding back when unsure; relying on moderator's questions to proceed	"Performing" thinking aloud for the moderator's sake
Situations triggering adaptation	Confusion about purpose, desire to "do things right" for moderator	Uncertainty about allowed exploration; moderator disruptions	Unclear role boundaries or moderator-managed system steps	Misunderstanding of task goals; unclear or fast instructions
Impact on exploration and agency	Some reduced independence; others ignored system logic	Lowered exploration; strong moderator-directed behavior	Hesitation and reduced agency in unclear guidance situations	Reliance on moderator for structure, slower pace, increased hesitancy
Overall emphasis	Performance orientation toward moderator	Dependence on moderator's cues and permission	Social hesitation and cue-driven reasoning	Social alignment shaped by clarity, interruptions, and role expectations

Table E.4: Comparison of misunderstandings and coordination problems across methods (RQ2).

Dimension	M1	M2	M3	M4
Type of problems	–	Subtle coordination issues caused by inconsistent instruction clarity	Subtle misunderstandings related to vague spatial descriptions and role ambiguity	–
Unique characteristics	–	Spatial confusion despite being co-located; system timeouts created additional disruptions	Problems caused by distance, mismatched direction cues, and unclear boundaries on when the moderator should act independently	–
Level of misunderstanding	–	Low–Medium	Low–Medium	–

Table E.5: Comparison of perceived moderator and environmental influence across groups (RQ1/RQ3).

Dimension	M1	M2	M3	M4
Primary source of influence	Instruction ambiguity and challenges maintaining moderator neutrality	Subtle interpersonal cues such as tone, pacing, and hesitation	Environmental limitations, especially missing door feedback	Communication barriers including language, audio issues, and moderator fatigue
Role of moderator behaviour	Strong influence due to repeated clarifications that became leading	Moderate influence through small social signals interpreted by participants	Moderator prompts interpreted as hints due to unclear system feedback	Moderator adaptations (slower speech, extra help) increased participant dependence
Impact of environment	Artificial setup and unclear task phrasing shaped participant decisions	Uncertainty heightened when moderator appeared unsure	Lack of system feedback forced reliance on moderator timing and prompts	Technical issues and unclear audio contributed to hesitation and stress
Participant response patterns	Confusion and need for task repetition; reliance on moderator guidance	Reduced exploration when sensing moderator uncertainty	Focus shifted from system understanding to interpreting moderator intent	Increased stress, hesitation, and reliance due to communication difficulties
Overall emphasis	Clarity of tasks and neutrality	Interpersonal subtleties	Environmental constraints	Communication challenges

Table E.6: Comparison of moderator influence on participant behavior and problem-solving (RQ1/RQ3).

Dimension	M1	M2	M3	M4
Type of moderator influence	Encouragement, clarifications, tone; occasional unintended guidance	Instruction precision and timing shaping carefulness and caution	Leading questions, pacing, and rephrasing steering reasoning	Clarifications and motivational cues interpreted as guidance
Participant sensitivity to moderator	Some participants corrected themselves after noticing the moderator	Participants became cautious or hesitant due to moderator timing	Participants adjusted reasoning to align with moderator cues	Participants relied on moderator for reassurance and direction
Role of task or instruction clarity	Confusion created moments where moderator influence increased	Clarification questions like “here?” acted as subtle prompts	Unclear task framing led participants to wait for direction	Uncertainty about moderator involvement created hesitation
Impact on problem-solving	Moderator reactions shaped participant actions more than intended	Moderator speed altered participant discovery and confidence	Moderator pacing reduced opportunities for independent reasoning	Participants followed moderator cues rather than system behavior
Overall emphasis	Tone and reactions influencing behavior	Timing, precision, and interpersonal cues	Leading guidance and framing effects	Dependence on moderator availability and clarity