USABILITY TEST PLAN

# SageMaker Unified Studio Release Center

## Moderated Usability Study

*Prepared by Ofir Levy, Senior UX Designer · April 2025*

| STUDY TYPE | PARTICIPANTS | DURATION | PROTOTYPE |
|---|---|---|---|
| **Moderated Remote** | **12 total (4 per persona)** | **60 min per session** | **High-Fidelity Figma** |

## 1. Study Overview

This usability test plan covers the moderated remote study for the Amazon SageMaker Unified Studio Release Center. The study is designed to validate the core interaction model, identify friction in the release initiation and approval workflows, and assess whether the information architecture supports the three key personas who interact with the feature.

### What We Are Testing

- Release initiation flow from project workspace to submitted state
- Approval and rejection workflow, including the side panel interaction model
- Release history navigation, filtering, and search
- Rollback initiation and confirmation dialog
- Audit log export discoverability and execution
- Status badge system legibility and mental model alignment

### What We Are Not Testing

- Backend performance, latency, or reliability
- Onboarding or first-run experience for new SMUS users
- Mobile or tablet layout adaptation
- Notification and email delivery workflow
- Advanced configuration options (approval chain setup, policy editing)
- Cross-project or domain-level release aggregation views

## 2. Research Questions

The following primary and secondary research questions guide task design, observation focus, and analysis priorities for this study.

### Primary Research Questions

- Can ML engineers initiate a release with the correct fields and reach the submitted state without facilitator assistance?
- Do project leads understand the approval side panel interaction model, and can they make an approval decision with confidence using only the information surfaced inline?
- Is the rollback action discoverable, and does the confirmation dialog provide sufficient context for approvers to act without external investigation?
- Can compliance stakeholders locate, filter, and export the audit log without navigating away from the Release Center?

### Secondary Research Questions

- Does the status badge system (Draft, Pending Approval, Approved, Rejected, Rolled Back) map to participants' mental models of a release lifecycle?
- Is the Release Center navigation entry point (under Governance in the left rail) discoverable without prompting?
- Does the diff view provide sufficient context for approvers who are reviewing a release from a different team or workstream?
- Do participants understand the distinction between approving a release and deploying it to production?

## 3. Participant Criteria

Participants are recruited across three personas that represent the primary users of the Release Center. All sessions are conducted remotely via video conference with screen sharing. Participants must have access to a desktop or laptop computer during the session.

PERSONA A

**ML Engineer / Data Scientist**

*n = 4*

- 2+ years working with ML pipelines or model training
- Has promoted or deployed a model to a production environment

PERSONA B

**Project Lead / Tech Lead**

*n = 4*

- Has reviewed and approved technical or ML-related changes in a professional context
- Manages or coordinates work across at least 2

PERSONA C

**Compliance / Platform Stakeholder**

*n = 4*

- Works in a compliance, security, audit, or platform engineering role
- Regularly reviews deployment logs or

- Familiarity with at least one ML platform (SageMaker, Databricks, W&B, or MLflow)
- Not required to have prior SMUS experience
- Mix of data scientists and ML engineers

engineers or data scientists

- Experience with code review tools (GitHub, GitLab, Gerrit) or release approval processes
- Not required to have ML domain expertise, but must understand release risk concepts

change records for governance purposes

- Experience with admin dashboards, audit tools, or compliance review portals
- Does not need to be an ML practitioner — governance context is sufficient

### 3.1 Screener Highlights

**Disqualifying Criteria (applies to all personas)**

- Currently employed by AWS, Amazon, or a direct AWS competitor working on an ML platform product
- Has participated in an SMUS usability study in the past 6 months
- UX researcher, UX designer, or product manager by primary role (may bias observation and think-aloud quality)
- Unable to join via desktop or laptop with screen sharing capability

## 4. Methodology

### 4.1 Study Format

All sessions are conducted as moderated, task-based usability tests using a clickable Figma prototype. The facilitator will guide participants through a structured set of tasks using a think-aloud protocol, asking participants to verbalize their reasoning as they navigate the prototype. The observer will take notes using a structured observation guide shared in real time via a collaborative document.

**Session Structure**

- 0:00 — Welcome and consent (5 min)
- 0:05 — Warm-up questions (8 min)

**Facilitation Approach**

- Think-aloud protocol throughout all tasks
- Non-directive probing only: 'What would you do

**Recording & Consent**

- Sessions recorded (audio, video, screen) with explicit written consent
- Recordings stored in secure internal folder,

| | | |
|---|---|---|
| • 0:13 — Prototype tasks (35 min)<br>• 0:48 — Debrief and open questions (8 min)<br>• 0:56 — SUS questionnaire (4 min)<br>• 1:00 — Session close | next?' / 'What are you looking for?'<br>• No assistance unless participant is stuck for more than 2 minutes with visible distress<br>• Post-task probes after each task before moving to the next<br>• Open debrief using laddering technique to uncover mental models | accessible to study team only<br>• Participants may withdraw consent at any time without consequence<br>• Clips shared internally only for synthesis purposes — not in external materials<br>• Participant names replaced with codes (P1A, P2A, P1B, etc.) in all outputs |

## 4.2 Prototype Scope

- The Figma prototype covers the following flows at high fidelity: Release List, Release Detail, New Release modal, Approval side panel, Rollback confirmation dialog, and Audit Log Export.
- Prototype is device-responsive to desktop viewport (1440px) only. Participants will be directed to view it in full screen.
- Data is pre-populated with realistic fictional content (project: 'fraud-detection-v3', assets, release notes, and thread comments) to support naturalistic task completion.
- Links that lead outside the tested scope display a standard 'This area is not part of the prototype' placeholder screen — facilitator will redirect from these.

## 4.3 Metrics

| Metric | Definition | Collection Method |
|---|---|---|
| **Task Completion Rate** | Percentage of participants who complete each task without facilitator assistance, within the allotted time | Facilitator observation log (pass / fail / assist) |
| **Time on Task** | Elapsed time from task prompt delivery to task completion or abandonment | Screen recording timestamp analysis |
| **Error Rate** | Number of incorrect paths, dead ends, or incorrect submissions made per task before correct completion | Facilitator observation log |
| **Ease of Use Rating** | Post-task single-question rating: 'How easy or difficult was that task?' (1–7 Likert) | Post-task verbal rating captured in observation doc |
| **System Usability Scale (SUS)** | Standardized 10-question post-session questionnaire scored 0–100 | Qualtrics survey delivered post-session |

| | | |
|---|---|---|
| **Think-Aloud Observations** | Qualitative comments, hesitations, confusions, and positive reactions captured verbatim | Live note-taking in shared observation guide |

# 5. Usability Tasks

Tasks are presented below in the order they will be administered. Each task includes the participant prompt (read aloud by the facilitator verbatim), the steps the facilitator will observe, success criteria, and the metrics to capture. Tasks are ordered to follow the natural flow of the release lifecycle.

### TASK 1 Navigate to the Release Center

*All Personas · 4 min*

**Participant Prompt:** *"You've just joined the fraud-detection-v3 project in SageMaker Unified Studio. Your team has asked you to check on the status of recent releases. Where would you go to find that information?"*

**Steps to Observe**

1. Participant lands on the SMUS project overview screen (pre-loaded prototype starting state)
2. Observe where participant looks first for navigation options
3. Note whether participant locates 'Release Center' in the left rail under 'Governance' independently
4. Note any hesitation, incorrect navigation attempts, or verbalizations about navigation structure
5. Task is complete when participant reaches the Release List view

**Success Criteria**

- ✓ Participant locates Release Center without facilitator hint
- ✓ Participant navigates correctly within 90 seconds
- ✓ No more than one incorrect navigation path before reaching the correct destination

**Metrics to Capture**

- Task completion rate (assisted vs. unassisted)
- Time to navigate from project overview to Release List
- Number of incorrect clicks before reaching Release Center
- Verbal comments on navigation label or location

### TASK 2 Submit a New Release

*Persona A — ML Engineer · 10 min*

**Participant Prompt:** *"Your model retraining run finished this morning and the results look good. You need to submit it for approval so it can be promoted to production. The model is called 'fraud-scoring-xgb', version 2.4.1, and the target environment is prod-us-east-1. Go ahead and submit a new release."*

**Steps to Observe**

6. Observe how participant locates the release initiation entry point (New button in toolbar)
7. Note whether participant completes all required fields (name, target environment, linked assets, release notes) without prompting
8. Observe how participant interacts with the asset selection step — does the asset manifest make sense?
9. Note whether participant opens or ignores the 'Advanced options' section
10. Observe submit confirmation behavior — does participant expect a confirmation screen?
11. Task is complete when participant reaches the submitted/pending approval state

**Success Criteria**

**Metrics to Capture**

- Task completion rate and time on task

- ✓ Participant completes all required fields without facilitator guidance
- ✓ Participant submits the release and reaches the Pending Approval status without error
- ✓ Participant does not express confusion about what the 'linked assets' field expects
- ✓ Task completed within 8 minutes

- Number of form fields that cause hesitation or require re-reading
- Whether participant opens Advanced options and why
- Post-task ease rating (1–7)
- Verbatim comments on form structure, field labels, and asset selection UX

## TASK 3 Review and Approve a Pending Release

*Persona B — Project Lead · 10 min*

**Participant Prompt:** *"You've received a notification that a new release has been submitted for your approval. The release is for the fraud-scoring model. Take a look at what's being proposed and decide whether you're comfortable approving it."*

### Steps to Observe

12. Observe whether participant navigates to the pending release from the Release List or expects a direct link
13. Note how participant uses the release detail page — do they read the diff? The release notes? The approval thread?
14. Observe whether participant opens the side panel or expects to navigate to a separate approval page
15. Note participant's confidence level before approving — do they feel they have enough information?
16. Observe any confusion about the distinction between approving the release vs. deploying it
17. Task is complete when participant clicks Approve and sees the status update to Approved

### Success Criteria

- ✓ Participant locates the pending release and opens the detail view without assistance
- ✓ Participant references the diff or release notes before making an approval decision
- ✓ Participant completes the approval action without confusion about what 'Approve' means
- ✓ Participant does not attempt to navigate away from the side panel to complete the approval
- ✓ Task completed within 8 minutes

### Metrics to Capture

- Task completion rate and time on task
- Which sections of the release detail participant reads before approving
- Whether participant expresses uncertainty about what 'Approve' commits them to
- Post-task ease rating (1–7)
- Verbatim comments on information sufficiency for an approval decision

## TASK 4 Initiate a Rollback

*Persona B — Project Lead · 8 min*

**Participant Prompt:** *"The release that went to production this morning is causing an unexpected spike in false positives. You need to roll back to the previous stable version as quickly as possible. How would you do that?"*

### Steps to Observe

18. Observe whether participant locates the Rollback button without facilitator guidance — note whether they look in the action bar, a menu, or elsewhere
19. Note participant's reaction to the rollback confirmation dialog — do they read it? Does it feel reassuring or alarming?
20. Observe whether participant understands what 'v2.3.8' means in the rollback target description
21. Note whether participant expects additional confirmation or notification after initiating rollback
22. Task is complete when participant confirms the rollback dialog and sees the status update to 'Rolled Back'

### Success Criteria

- ✓ Participant locates the Rollback button without facilitator assistance
- ✓ Participant reads the confirmation dialog before confirming — not a reflexive click-through
- ✓ Participant expresses confidence (not uncertainty) in what the rollback will do
- ✓ Task completed within 6 minutes

### Metrics to Capture

- Task completion rate and time to locate Rollback button
- Whether participant reads the confirmation dialog content
- Verbal confidence level before confirming rollback
- Post-task ease rating (1–7)
- Any requests for additional information not present in the confirmation dialog

## TASK 5 Filter and Export the Audit Log

*Persona C — Compliance Stakeholder · 8 min*

**Participant Prompt:** *"Your team needs to provide a record of all production releases in the fraud-detection project for the past 30 days for an upcoming compliance review. Please find the release history and export it."*

### Steps to Observe

23. Observe how participant navigates to the Release List from their starting point
24. Note whether participant uses the filter controls independently and selects the correct date range and status filter
25. Observe how participant locates the Export Audit Log action — is it visible in the toolbar?
26. Note any confusion about what the export will contain (format, fields, date range applied)
27. Task is complete when participant initiates the export download

### Success Criteria

- ✓ Participant filters the release list by date range and 'Approved' status without facilitator guidance

### Metrics to Capture

- Task completion rate and time on task

- ✓ Participant locates the Export Audit Log button in the toolbar
- ✓ Participant initiates the export without confusion about file format or content scope
- ✓ Task completed within 6 minutes

- Whether participant uses filter controls before or after locating the export button
- Any confusion about what the export will include
- Post-task ease rating (1–7)
- Verbal comments on filter options, export label clarity, and audit log completeness

### TASK 6 Interpret Release Status Badges

*All Personas · 4 min*

**Participant Prompt:** *"Take a look at the release list in front of you. Without clicking anything, can you tell me what you think each of these status labels means — and what you would expect to be able to do with each one?"*

**Steps to Observe**

28. Display the Release List with all five status states visible (Draft, Pending Approval, Approved, Rejected, Rolled Back)
29. Ask participant to describe each status badge in their own words
30. Note which badges are immediately understood vs. require explanation
31. Probe: 'What would you expect to happen if you clicked on a Rejected release?'
32. Probe: 'What is the difference between Approved and Rolled Back in your mind?'

**Success Criteria**

- ✓ Participant correctly interprets at least 4 of 5 status badges without prompting
- ✓ Participant understands that 'Rolled Back' represents a prior approved release that has been superseded
- ✓ Participant does not confuse 'Approved' with 'Deployed to production'

**Metrics to Capture**

- Per-badge comprehension accuracy (correct / partially correct / incorrect)
- Which badge(s) require clarification or generate uncertainty
- Verbatim definitions offered by participant for each badge
- Confusion between approval state and deployment state

## 6. Observation Guide

The observer (a second team member present in all sessions) will use the following framework to capture observations in real time. Observations are recorded in a shared Google Sheet with one row per observation, tagged by task number, participant ID, and category.

| | |
|---|---|
| CONFUSION | Participant pauses, re-reads, or verbalizes uncertainty about where to go, what something means, or what will happen if they act. Note the exact UI element or label that triggered the confusion. |

| | |
|---|---|
| **SUCCESS** | Participant completes a step efficiently and without hesitation, or explicitly verbalizes that something is clear, logical, or matches their expectations. Note what specifically worked well. |
| **WORKAROUND** | Participant attempts to accomplish a task through an unintended path, uses browser back, or expresses a desire for a feature or shortcut not present in the design. Note the attempted path. |
| **MENTAL MODEL** | Participant reveals an assumption about how the system works — correctly or incorrectly. These are often introduced by phrases like 'I would expect...', 'I assume...', or 'Normally in tools like this...'. Capture verbatim. |
| **DELIGHT** | Participant expresses a positive reaction to a specific design element, interaction, or piece of information. Note the element and the participant's exact words. |

## 6.1 Post-Task Probes (read verbatim after each task)

- "How easy or difficult was that on a scale from 1 to 7, where 1 is very difficult and 7 is very easy?"
- "Was there anything confusing or unexpected about what you just did?"
- "Is there anything you would have expected to find or see that wasn't there?"
- "If this were a real system you used at work, how would you feel about using that feature regularly?"

## 6.2 Debrief Questions (ask after all tasks)

- "Overall, what was your impression of the release workflow you interacted with today?"
- "What was the one thing that felt most natural or well-designed?"
- "What was the one thing that felt most confusing or that you'd want to change?"
- "How does this compare to other tools you've used to manage releases or deployments at work?"
- "Is there anything about how your team handles releases today that this tool doesn't seem to support?"

# 7. Success Thresholds & Go / No-Go Criteria

The following thresholds define acceptable vs. below-target performance for each metric. Results that fall below the No-Go threshold on any primary metric will require a design revision and a follow-up study round before the feature proceeds to engineering handoff.

| Metric | Target | Acceptable | No-Go (requires revision) |
|---|---|---|---|
| **Overall Task Completion Rate** | **≥ 85%** | 75–84% | < 75% on any P0 task |
| **Release Initiation (Task 2) — Unassisted Completion** | **≥ 80%** | 65–79% | < 65% complete without assistance |
| **Approval Flow (Task 3) — Unassisted Completion** | **≥ 85%** | 70–84% | < 70% complete without assistance |
| **Rollback Discoverability (Task 4) — Time to Locate** | **< 30 sec** | 30–60 sec | > 60 sec average |
| **Status Badge Comprehension (Task 6)** | **≥ 4 of 5 correct** | 3 of 5 correct | < 3 of 5 correct |
| **Post-Task Ease Rating — Average** | **≥ 5.5 / 7** | 4.5–5.4 / 7 | < 4.5 / 7 average |
| **System Usability Scale (SUS)** | **≥ 75 (Good)** | 65–74 (OK) | < 65 (Poor) |

# 8. Analysis Plan & Deliverables

## 8.1 Analysis Approach

### Quantitative

- Aggregate task completion rates and time-on-task per task and per persona group
- Calculate mean ease ratings per task; flag tasks below 4.5
- Score SUS per participant; average across the full cohort and by persona
- Build error rate heatmap: frequency of incorrect clicks by screen area

### Qualitative

- Affinity map all observation notes using the 5 category framework from Section 6
- Identify recurring patterns across 3+ participants as primary findings
- Tag observations to specific UI elements; build a priority issue list
- Pull verbatim quotes that best illustrate each primary finding

## 8.2 Deliverables

**Study Outputs — delivered within 5 business days of final session**

- Topline Report: 1–2 page summary of critical findings and go/no-go recommendation for PM and engineering leads
- Full Research Report: Findings by task, with supporting quotes, error rate data, SUS score, and prioritized issue list
- Annotated Prototype Clips: 5–8 short video clips illustrating key friction moments and positive reactions, shared in internal Confluence
- Prioritized Issue Backlog: JIRA tickets or Figma annotation comments for each finding, tagged P0/P1/P2 by severity
- Raw Data: Observation spreadsheet, SUS scores, task completion log — archived in study folder for future reference

## 8.3 Timeline

| Phase | Activities | Timing |
|---|---|---|
| **Preparation** | Finalize prototype, recruit screener, set up Qualtrics survey, brief observer team, pilot session with internal team member | **Week 1** |
| **Sessions — Persona A** | 4 sessions with ML engineers / data scientists (60 min each) | **Week 2, Days 1–3** |
| **Sessions — Persona B** | 4 sessions with project leads / tech leads (60 min each) | **Week 2, Days 4–5** |
| **Sessions — Persona C** | 4 sessions with compliance / platform stakeholders (60 min each) | **Week 3, Days 1–2** |
| **Analysis** | Affinity mapping, quantitative analysis, SUS scoring, clip selection, issue prioritization | **Week 3, Days 3–5** |
| **Topline Report** | 1-page summary delivered to PM and engineering lead | **End of Week 3** |
| **Full Report + Backlog** | Complete report, annotated clips, and JIRA tickets delivered to team | **Week 4, Day 3** |

## 9. Risks & Mitigations

### Study Risks

- Participant no-shows: Buffer 2 alternates per persona; allow same-day reschedule up to 24 hours before session
- Prototype dead ends confuse participants beyond the tested area: Facilitator script includes a standard redirect phrase for out-of-scope links
- Think-aloud fatigue in longer tasks: Facilitator uses lightweight prompts ('What are you thinking right now?') every 60 seconds of silence
- Persona B participants unfamiliar with ML releases: Screener includes a minimum familiarity threshold; brief context-setting paragraph read before Task 3

### Validity Risks

- Small sample (n=12) limits statistical generalizability: Study is designed for directional qualitative insight, not statistical significance — findings framed accordingly
- Prototype may feel less realistic than a live product: Participants briefed that prototype is intentionally incomplete; facilitator acknowledges placeholders before tasks
- Facilitator bias toward confirming design hypotheses: Observer takes independent notes; debriefs are structured to probe for negative reactions
- Compliance persona may have low ML context: Tasks are designed to be domain-agnostic for Persona C; no ML knowledge required to complete audit log tasks

*Prepared by Ofir Levy, Senior UX Designer · ofirlevydesign.com · April 2025*

*Amazon SageMaker Unified Studio — Release Center · Usability Test Plan · For internal and portfolio use only*