MASTERARBEIT | MASTER'S THESIS

Titel | Title

Seeing the Stress: Dashboard Design for Post-Hoc Analysis of Police Training Communication

verfasst von | submitted by

Roham Jabbari

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 921
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Informatik
Betreut von | Supervisor:	Univ.-Prof. Torsten Möller PhD
Mitbetreut von | Co-Supervisor:	Laura Koesten MSc Ph.D.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Jakob Uhl and Laura Koesten, for their guidance, insightful feedback, and continuous support throughout this research.

I am also grateful to Michael Oppermann from the Austrian Institute of Technology (AIT) for his expert input and valuable contributions during the development of the dashboard and visualization components.

Finally, I would like to thank my girlfriend, Vira, for her unwavering encouragement, understanding, and patience throughout this journey.

Abstract

Police officers must make split-second decisions under extreme stress, where communication quality can determine life-or-death outcomes. While VR training systems now capture these high-pressure encounters, existing analysis tools fragment communication structure, content, and emotional state into separate views, preventing trainers from understanding how stress manifests in actual dialogue.

This thesis presents an integrated visualization dashboard that synchronizes speech transcripts, topic models, and stress indicators within a unified temporal framework. Informed by prior work in conversation visualization, the dashboard combines structural and affective analysis to enable users to trace precise relationships between stress spikes, topic shifts, and communication patterns, revealing not just what was said, but how stress shaped the conversation.

Evaluation with 15 participants showed strong usability (PSSUQ: 6.6/7) and effective learnability, with task accuracy improving from 76.7% to 96.7% after brief onboarding. Topic visualization created higher cognitive load than other views, but frustration remained low across all tasks.

The work presents a working prototype and evaluation that demonstrates the feasibility of integrated communication visualization for training contexts.

Kurzfassung

Polizeibeamte müssen unter extremem Stress Entscheidungen in Sekundenbruchteilen treffen, bei denen die Qualität der Kommunikation über Leben und Tod entscheiden kann. Obwohl VR-Trainingssysteme diese Hochdrucksituationen mittlerweile erfassen, zerlegen bestehende Analysewerkzeuge Kommunikationsstruktur, Inhalte und emotionale Zustände in getrennte Ansichten und erschweren es Trainern zu erkennen, wie sich Stress in realen Dialogen manifestiert.

Diese Masterarbeit präsentiert ein integriertes Visualisierungs-Dashboard, das Gesprächstranskripte, Themenmodelle und Stressindikatoren in einem gemeinsamen zeitlichen Rahmen synchronisiert. Basierend auf früheren Arbeiten zur Konversationsvisualisierung kombiniert das Dashboard strukturelle und affektive Analyse, sodass Nutzer präzise Zusammenhänge zwischen Stressspitzen, Themenwechseln und Kommunikationsmustern nachvollziehen können und dabei nicht nur erkennen, was gesagt wurde, sondern auch, wie Stress das Gespräch beeinflusste.

Die Evaluation mit 15 Teilnehmern zeigte hohe Nutzbarkeit (PSSUQ: 6,6/7) und gute Erlernbarkeit. Die Aufgabengenauigkeit verbesserte sich von 76,7% auf 96,7% nach kurzem Onboarding. Die Themenvisualisierung verursachte höhere kognitive Belastung als andere Ansichten, aber die Frustration blieb bei allen Aufgaben gering.

Die Masterarbeit präsentiert einen funktionierenden Prototyp und eine Evaluation, die die Machbarkeit integrierter Kommunikationsvisualisierung für Trainingskontexte demonstriert.

Statement on the Use of AI Tools

This thesis was written with the assistance of AI-based tools. Specifically:

- **Claude (Anthropic)** was used for brainstorming, refining prose, checking grammatical consistency, and discussing conceptual framing during the writing process.

All AI-generated content was critically reviewed, verified, and substantially edited by the author. The intellectual contributions, research design, data analysis, and conclusions presented in this work are entirely my own. AI tools were used as writing and coding aids, not as sources of knowledge or original ideas.

Contents

List of Tables

List of Figures

1. Introduction

This chapter introduces the motivation and context of the research on communication analysis for high-stress police training, outlines the problem addressed, and presents the goals and contributions of the thesis. Section 1.1 defines the research problem in the context of high-stress police communication training. Section 1.2 explains why addressing this problem matters for operational outcomes. Section 1.3 describes the research goals and the specific contributions made. Section 1.4 presents the research questions guiding the work. Section 1.5 provides an outline of the thesis structure.

1.1. Problem Statement

In high-stress domains such as policing, emergency response, and military operations, effective communication under pressure is a key factor in team performance [STKSJ12]. Stress impairs working memory and narrows attention [Sta04], which can narrow an individual's perspective from team-level goals to self-focused concerns and degrade team performance [DSJ99]. Simulation-based training can address these challenges, but realizing its potential depends usually on structured debriefing in which participants review both what happened procedurally and how they experienced it emotionally [FG07].

In simulated police operations, officers must communicate effectively while managing substantial time pressure and cognitive demand. Small variations in tone, timing, and topic can influence coordination and task outcomes, yet these details are difficult to reconstruct accurately after an exercise. As a result, important indicators of stress and communication breakdown often remain unnoticed during debriefings. Gaining an overview of such patterns is particularly challenging. Making these complexities actionable and understandable for trainers and trainees requires tools that can surface patterns that would otherwise remain hidden in raw recordings or transcripts.

Visualization research offers techniques for analyzing communication dynamics, including streamgraph-based participation views [BW08] and radial topic-space visualizations [EAGA+16]. These tools reveal who spoke, what was discussed, and when topics shifted. Yet as reviewed in Chapter 2, most systems focus on a single aspect of communication (structure, content, or emotion) and rarely integrate these dimensions. This fragmentation limits their usefulness for reflective learning, particularly in training contexts where understanding how stress influences communication is vital for performance improvement [MUA+24].

A review of prior work (see Chapter 2) highlights a critical gap that motivates this thesis: existing visualization systems treat communication structure, content, and emotional context as separate analytical concerns, preventing trainers and trainees from understanding how stress-induced behavioral changes manifest in actual dialogue. When an officer's speech rate accelerates or their topic focus narrows under pressure, current tools cannot connect these observable patterns to stress indicators or tactical decision points. This disconnect is particularly problematic in police training debriefings, where instructors need to help officers recognize early warning signs of stress-impaired communication that precede critical errors. Without integrated visualizations that span multiple data modalities, post-training reflection risks remaining anecdotal rather than evidence-based.

This limitation underscores the need for a domain-aware system that not only integrates structure, content, and emotion, but also links them tightly enough that users can trace patterns they would miss when these signals are viewed separately.

1.2. Motivation: Why This Matters

The practical stakes of this research extend beyond improved training tools. In operational settings, communication breakdowns under stress contribute directly to preventable failures: delayed tactical responses, conflicting commands, and escalated confrontations. When officers cannot recognize their own stress-induced communication patterns (such as talking over teammates, fixating on irrelevant details, or losing emotional control), they cannot develop strategies to counteract them in the field [DSJ99].

Current debriefing practices rely heavily on instructor recall and subjective interpretation. Video recordings capture what happened, but not why it happened or how it felt [FG07]. Self-reported stress assessments lack temporal precision and miss moment-to-moment fluctuations [Sta04]. By providing an integrated, evidence-based view of communication dynamics synchronized with stress indicators and mission context, this work enables a fundamentally different mode of reflection: one where officers and trainers can trace the relationship between stress indicators, conversational behavior, and operational outcomes with more precision and clarity.

A critical challenge in this domain is translating insights about how stress affects working memory, attention, and communication into forms that non-expert users (police trainers and trainees) can understand and act upon. While research in cognitive psychology and human factors provides detailed accounts of how performance degrades under stress [Sta04], these insights remain inaccessible to practitioners without advanced training. A visualization-based debriefing tool could bridge this gap by presenting cognitive load indicators and their relationship to communication patterns in ways that are intuitive, actionable, and grounded in familiar training contexts.

If successful, this approach could help transform how high-stress communication training is conducted in practice in a number of ways. Officers who understand how their communication degrades under pressure can develop personalized coping strategies. Trainers who can point to specific moments where stress manifested in dialogue can provide targeted, actionable feedback. Organizations that adopt evidence-based debriefing practices can systematically improve team performance in high risk situations. The broader implication is that reflective learning becomes a measurable, improvable skill.

1.3. Research Goals and Contributions

The overarching goal of this thesis is to design, implement, and evaluate an interactive visual analytics dashboard that supports retrospective analysis of police communication training in high-stress scenarios. The work pursues three complementary directions: (1) integrating communication content with emotional and operational context through a synchronized, multi-view visualization architecture, (2) adapting visualization components to domain-specific elements such as participant roles, event markers, and topic categories, and (3) embedding interaction and annotation features intended to support post-training reflection and debriefing.

The main contributions of this thesis are threefold:

1. A **synchronized, multi-view visualization architecture** that integrates multimodal communication data (speech transcripts, topic models, stress indicators) within a shared timeline, enabling cross-modal analysis of how stress manifests in dialogue patterns.

2. A **prototype implementation** demonstrating domain-adapted visualization techniques for police training contexts, including participant role labels, event markers, and synchronized stress-communication overlays.

3. A **formative evaluation** with 15 participants examining usability [Lew92], cognitive load [HS88], and learnability. The results inform design recommendations for communication analytics tools in team-based, high-stress training domains.

This work builds upon and extends prior research in immersive communication analysis, iterating on *Roger, Dashboard for Reflective Analysis* [OUR+25] (under review), which itself expanded on Murtinger et al.'s *Sound of the Police* VR training system [MUA+24]. Roger's dashboard focused on giving trainers a quick sense of who spoke when, linking what was said to the operational situation, supporting comparison across participants, and making it easy to mark and revisit key moments [OUR+25]. This thesis builds on this foundation with a coordinated multi-view design for high-stress training contexts, integrating speech activity, topics, and stress indicators within a unified timeline, and contributes a systematic evaluation of usability and cognitive load.

1.4. Research Questions

Main Research Question

How can an integrated, domain-aware visualization dashboard support post-training reflection and insight into communication dynamics, stress management, and decision-making in high-stress police training scenarios?

The research is guided by the following sub-questions:

- **RQ1 - What do we want to show?** How can we capture and represent the most relevant aspects of high-stress team communication (including timing, content, and emotional context) for meaningful analysis?
- **RQ2 - How do we want to show it?** What visualization techniques and interaction designs best convey the relationships between speech activity, topics, and stress indicators while remaining intuitive and usable?
- **RQ3 - How do we maintain analytical coherence across views?** How can temporal synchronization, contextual linking, and cross-view coordination enable users to trace the relationship between stress indicators, communication patterns, and mission events without losing analytical context during navigation?
- **RQ4 - How well does it work?** How do users perceive the system in terms of usability, learnability, cognitive load, and reflective value during post-training analysis?

1.5. Outline

The remainder of this thesis is structured as follows: Chapter 2 reviews related visualization systems for communication analysis, identifying design trends and gaps. Chapter 3 presents the dashboard design, detailing its conceptual model, visual architecture, and domain-specific adaptations. Chapter 4 describes the system's implementation, including technologies, data integration, and deployment. Chapter 5 reports the evaluation methodology, results, and interpretation. Chapter 6 discusses the findings in relation to research questions, compares the work to related systems, and acknowledges limitations. Finally, Chapter 7 summarizes key contributions and directions for future research.

2. Literature Review

In this chapter, we review recent work on the visualization of communication in two primary modalities: meetings and group chats. The review is organized accordingly, highlighting systems developed to support the analysis of team-based spoken interactions as well as informal, text-based messaging. Each system is discussed in terms of its design focus, visual encoding strategies, and analytical capabilities.

This chapter is organized into three main sections. Section 2.1 reviews systems for **visualizing meetings and team communication**, focusing on tools such as *ConToVi*, *Discussion Flows*, and *MeetingVis*, which model conversational structure, participation, and topic dynamics in spoken interaction. Section 2.2 examines systems for **visualizing group chat and messaging communication**, including *Chat Circles*, *GamIM*, and *Flower-Pop*, which emphasize affective feedback, participation balance, and social expressiveness in text-based interaction. Finally, Section 2.3 synthesizes these findings in a comparative **discussion**, identifying cross-cutting design trends, methodological gaps, and implications for the development of integrated communication visualizations.

2.1. Visualization of Meetings and Team Communication

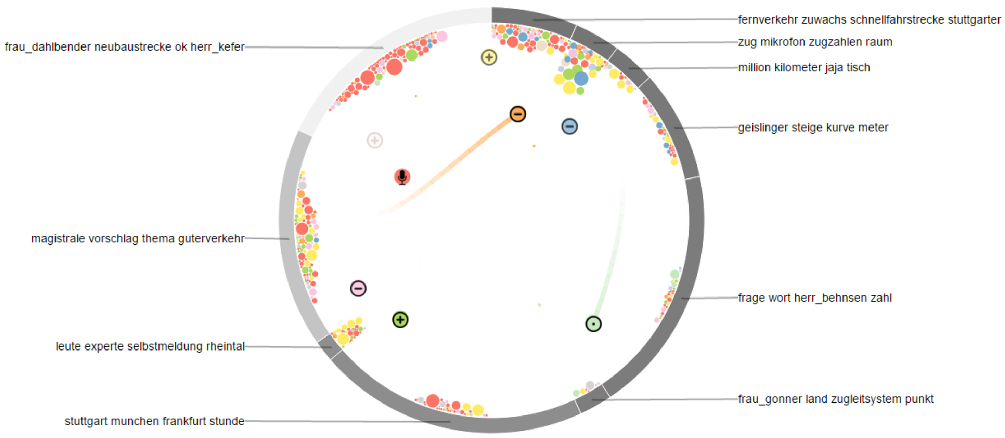


Figure 2.1.: ConToVi's Topic-Space View [EAGA+16] shows utterances as colored dots placed by semantic similarity, with trails visualizing speaker movement across topics. Sedimentation along the border reveals topic persistence, while central glyphs represent speaker roles or actions.

ConToVi [EAGA+16] is a visual analytics system for exploring multi-party conversation dynamics, using a radial “Topic-Space” layout to map utterances based on thematic similarity, as illustrated in Figure 2.1. The system includes animated speaker trails and glyph-based representations of speech turns, designed to help analysts track topic engagement over time. Visually, the radial layout is compelling and conceptually intuitive, especially for capturing the movement of speakers between topics. It enables quick recognition of thematic focus and speaker dominance. However, in practice, the layout can become cluttered as the number of utterances grows, making it harder to follow individual trajectories. Overlapping trails and dense label placement reduce clarity in larger conversations, and the absence of direct visual links between interacting speakers limits the ability to track dialogic exchange. While expert users found the tool helpful for forming mental models of conversation, casual users may struggle with the complexity of multi-view coordination and abstract encodings. The system was primarily intended for communication researchers and discourse analysts seeking to explore topic transitions and speaker engagement patterns in multi-party conversations. Overall, ConToVi’s visuals are rich and expressive, though their complexity and scalability challenges may pose accessibility hurdles for broader audiences.

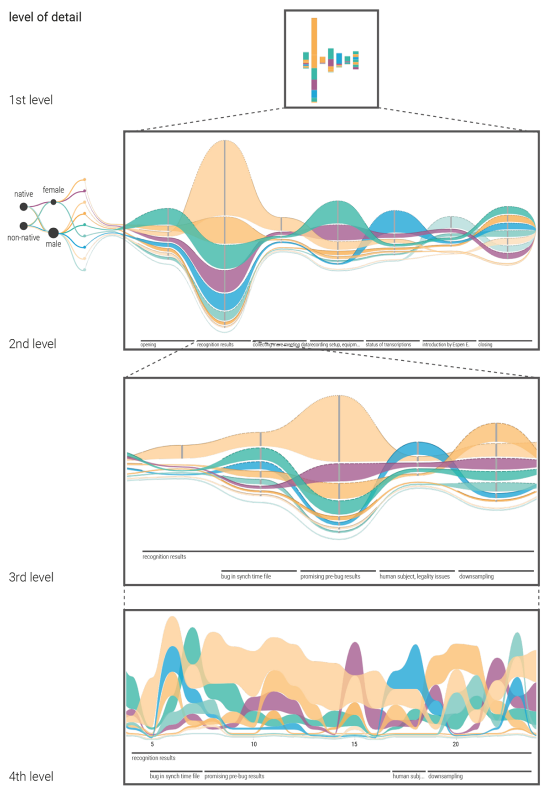


Figure 2.2.: Discussion Flows [WKV21] enables multi-level exploration of speaker engagement across meetings, from a compact glyph overview to detailed agenda-linked timelines.

Discussion Flows [WKV21] is an interactive visualization system for analyzing engagement and participation dynamics across multi-party meetings, as illustrated in Figure 2.2. The system uses a multi-level interface that combines a glyph-based overview for comparing multiple meetings with layered timeline views for in-depth inspection. Its flow-based visual design effectively conveys speaker dominance, turn-taking, and interruptions over time, offering a clear sense of engagement across agenda segments. The minimalistic glyphs and smooth visual transitions contribute to high legibility, even for non-expert users. Demographic filters such as gender or native language provide additional context for detecting participation bias. However, the interface centers heavily on structural aspects of the meetings and omits semantic information such as discussion topics or emotional tone. As the number of meetings or participants increases, the overview view can become visually dense, potentially limiting comparative clarity. The system was built using real meeting corpora and received informal validation through practitioner feedback, underscoring its practical relevance. It was designed for organizational researchers and meeting facilitators who analyze participation equity and engagement patterns across recurring team meetings. Overall, Discussion Flows strikes a strong balance between readability and analytical depth, though future extensions could enhance its expressiveness by incorporating sentiment cues or topic semantics.

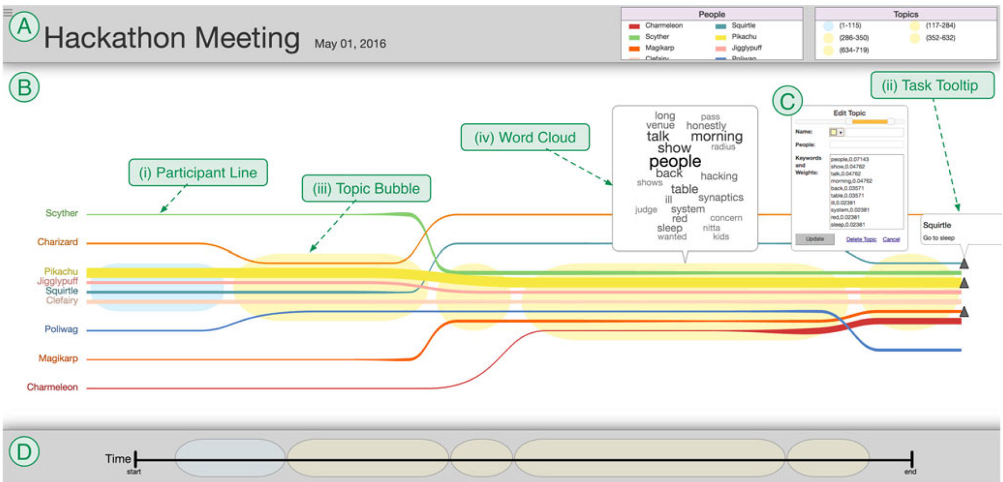


Figure 2.3.: MeetingVis [SBB+18] overview interface showing participant timelines, topic bubbles, and interactive word clouds. Each colored line represents a speaker, while topics are visualized as translucent clusters. The system supports semantic exploration through editable tooltips and detailed keyword distributions.

MeetingVis [SBB+18] provides an interface for reviewing past meetings, emphasizing how participants can reconstruct what was discussed and by whom (Figure 2.3). The interface represents each participant as a horizontal line, with topic bubbles indicating areas of conversational activity and associated word clouds providing semantic context.

Users can hover over tasks to reveal additional metadata and keywords, supporting temporal navigation and reflection. Visually, the design is clean and engaging: the storyline metaphor is effective for mapping involvement over time, and color-coded elements help distinguish users and topics. The word clouds provide helpful cues without overwhelming the display, and the timeline bar offers intuitive segmentation. However, the system's effectiveness depends on accurate automatic speech recognition (ASR), which can introduce noise in informal or overlapping dialogue. Furthermore, it does not provide sentiment analysis or emotional cues, and may not scale well in meetings with high speaker overlap. The intended users were office professionals and design researchers aiming to review meetings for reflection and memory recall, particularly within collaborative or project-based environments. Despite these limitations, MeetingVis stands out for its narrative framing and clear visual mapping of participation, making it useful for collaborative reflection and summarization tasks.

2.2. Visualization of Group Chat and Messaging Communication

Figure 2.4.: Chat Circles [DV02] visualizes synchronous text chat using a spatial layout. Participants are shown as colored circles that grow with each message and fade over time. The visual metaphor encourages ambient awareness and spatial grouping, though it limits clarity for dense or multi-topic conversations.

Chat Circles [DV02] is a visual communication interface that explores spatial metaphors to represent synchronous chat interactions, as shown in Figure 2.4. As described by Donath and Viégas [DV02], each user appears as a colored circle that grows with message length and fades over time. Messages are displayed within circles, and a proximity-based "hearing radius" limits visibility to nearby participants, creating a sense of spatial co-presence and supporting emergent grouping. The interface encourages a sense of shared presence and informal clustering, offering an ambient rather than content-focused visualization. Its design is highly minimalist, which contributes to low cognitive load and makes it easy to grasp at a glance. However, the system does not support content analysis, threading, or sentiment cues, which limits its analytical applications. It also lacks historical tracking and scales poorly in multi-topic environments. Chat Circles targeted casual online chat participants and social computing researchers interested in exploring alternative, socially expressive forms of digital presence. Despite these limitations, Chat Circles was well received in public installations and early lab studies for its novel, ambient interface, particularly in casual or artistic communication settings.

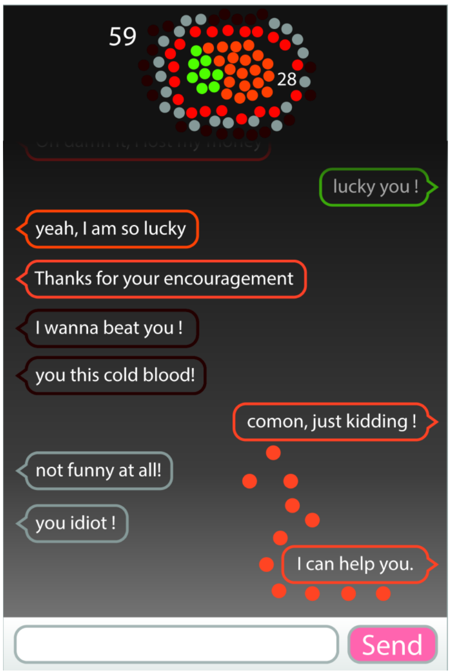


Figure 2.5.: GamIM's interface [PWH14] visualizes the emotional atmosphere of a chat using animated ambient visuals. A "lava lamp" effect accumulates sentiment over time, providing indirect feedback about group mood. The aesthetic design emphasizes emotional awareness without displaying explicit sentiment scores.

GamIM [PWH14] is a messaging interface that visualizes the emotional atmosphere of chat sessions through ambient, animated graphics (Figure 2.5). The interface uses a sentiment classifier to detect the emotional polarity of messages and represents this using a "lava lamp" style bubble animation that changes color and movement based on cumulative group sentiment. This indirect feedback loop encourages users to reflect on the emotional tone of their contributions without relying on explicit metrics or alerts. Visually, the interface is soothing and engaging, with soft transitions and a focus on ambient feedback rather than textual content. While this aesthetic approach is appealing, the underlying sentiment detection is binary (positive or negative), which oversimplifies emotional nuance. Moreover, the sentiment mapping is opaque to users and lacks interpretability, potentially causing confusion. There is no integration of message content or structural indicators, limiting analytical use. GamIM was developed for small informal chat groups and interaction design researchers exploring how ambient emotional feedback could influence communication behavior. Although the tool showed early promise in promoting more reflective communication, it was not formally evaluated. GamIM offers a compelling proof-of-concept for emotional ambient design, but would benefit from finer-grained sentiment models and clearer user feedback mechanisms.

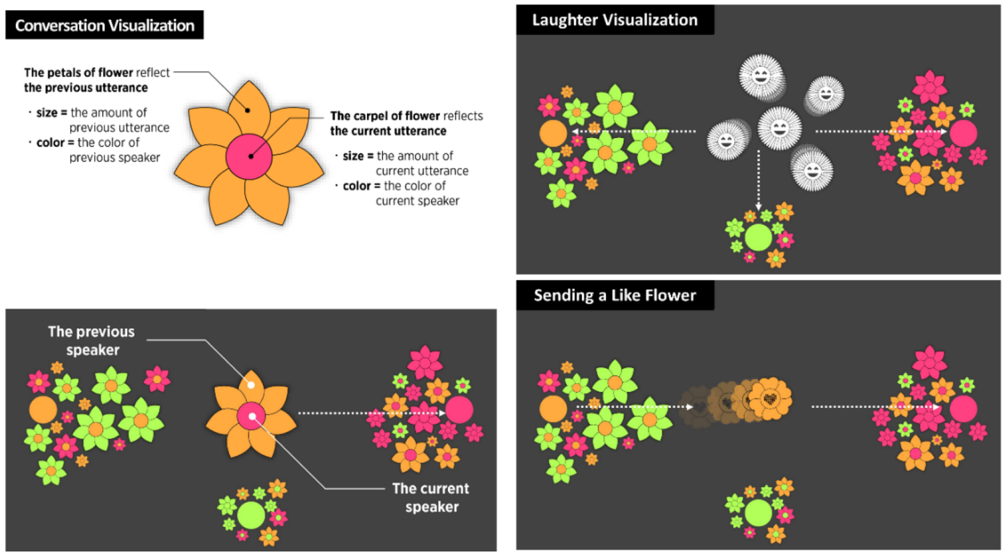


Figure 2.6.: Flower-Pop [LRS+18] uses animated flowers to visualize participation in casual group conversations. Each flower represents a user, with petals growing to indicate speaking turns or shared content. The playful metaphor supports inclusive interaction but abstracts away message detail.

Flower-Pop [LRS+18] is a mobile group chat system that uses animated floral metaphors to encourage participation and engagement in informal conversations, as illustrated in Figure 2.6. Each participant is represented by a blooming flower, with petals reflecting speaking turns, shared photos, or interaction frequency. The system provides multiple facilitation modes, including random speaker prompts and topic rotation, to draw in quieter participants and support balanced interaction. The visual design is playful, engaging, and socially expressive, allowing users to intuitively interpret engagement levels.

However, the metaphor's abstraction can make it difficult to access specific content or assess conversational structure. The system does not include sentiment analysis or topic modeling, limiting its analytical depth. Its appeal lies in promoting inclusivity and playful dynamics in social settings, but it may not scale well for professional use or complex discussions. The intended audience was casual social or educational group chat users, particularly in youth or classroom contexts where participation and inclusivity were key design goals. A field study highlighted increased engagement and positive user reception, though the scope was limited. Flower-Pop succeeds in making group conversation visually appealing and inclusive, while trading off interpretability and analytical utility.

2.3. Summary and Implications

To summarize the reviewed works, Table 2.1 provides a comparative overview of the six selected systems. It includes the communication modality targeted, the main visualization focus, whether emotional or affective dynamics are addressed, and the presence or type of user evaluation. This overview helps highlight the common design goals as well as the diversity of approaches across meeting and group chat visualization systems.

Table 2.1.: Comparison of Reviewed Systems

System	**Communication Modality**	**Visualization Focus**	**Emotional Dynamics**	**User Evaluation**
Discussion Flows [WKV21]	Meetings	Speaker engagement flows	No	Limited
MeetingVis [SBB+18]	Meetings	Meeting narratives (storyline)	No	User study
ConToVi [EAGA+16]	Meetings	Speaker-topic exploration in topic space	No	Expert case study
Chat Circles [DV02]	Group Chat	Spatial chat visualization	Indirect	Informal feedback
GamIM [PWH14]	Group Chat	Emotional atmosphere visualization	Yes	Preliminary feedback
Flower-Pop [LRS+18]	Group Chat	Topic visualization across devices	Partial	Field study

The systems reviewed in this chapter reveal several design trends and recurring patterns in communication visualization. Most approaches focus on either structural or content-based analysis, emphasizing participation dynamics, speaker activity, or thematic flow. Discussion Flows provides layered overviews of speaker engagement across agenda segments, while MeetingVis emphasizes storyline recall through topic bubbles and participant lines. These designs highlight the dominance of timeline-based representation for mapping verbal interaction over time.

Some tools adopt metaphorical designs to enhance interpretability and engagement. ConToVi uses radial trails to visualize speaker movement across topics, while GamIM employs animated ambient visuals for emotional atmosphere, and Flower-Pop uses floral metaphors for participation. These metaphors offer expressive alternatives to conventional charts, though their effectiveness often depends on user familiarity and visual clarity. While these approaches enrich the visual language of communication analysis, few are backed by extensive empirical validation.

Several gaps emerge consistently across the reviewed systems. First, emotional or stress-related modeling remains largely absent. Although GamIM incorporates basic sentiment cues, most visualizations overlook the dynamic affective aspects of communication, how emotional tone shifts, escalates, or diffuses over time. This omission reduces the ability to detect moments of tension, conflict, or empathy.

Second, visual components are rarely integrated across modalities. Message logs, emotional graphs, and operational timelines typically exist as separate views without meaningful cross-linkage, preventing users from tracing relationships between content, structure, and affect. Similarly, many tools are tailored to specific modalities and lack adaptability. Systems designed for meetings rarely generalize well to group messaging, where temporal patterns and interaction dynamics differ.

Third, domain-specific context receives minimal attention. Few systems account for speaker roles, authority relations, or mission phases that are essential for interpreting communication in real-world domains like policing or emergency response. This limits their applicability beyond generic meeting or chat scenarios.

Fourth, evaluation practices vary widely. While some systems are tested in controlled studies or real-world deployments, others rely solely on informal feedback or expert observations. The absence of standardized evaluation frameworks makes it difficult to compare tools or generalize findings across settings.

The reviewed systems reveal three recurrent patterns: strong support for structural and participation analyses (timelines, flow diagrams), frequent use of metaphorical or ambient encodings to improve engagement, and uneven empirical validation. Key gaps persist: stress- or emotion-aware modeling is largely absent, multimodal components (content, structure, affect) are rarely integrated, domain-specific context such as roles and mission phases is often missing, and evaluation practices vary widely. These limitations motivate the present work: a domain-aware dashboard that integrates structural, thematic, and affective signals within a shared temporal framework, aligns visualizations with policing conventions, and supports coordinated multi-view analysis to enable evidence-based post-training reflection.

3. Dashboard Design

This chapter presents the design of the interactive dashboard developed to support post-training reflection in high-stress communication scenarios. Section 3.1 outlines the overarching goals and design objectives that guided the system's development. Section 3.2 describes the shared timeline and multi-modal data signals that form the foundation of the dashboard. Section 3.3 details the architecture and tab designs, explaining how each visualization supports the analysis of communication and stress within a coordinated multi-view system, including the rationale behind key design choices and alternative solutions considered during development.

The design decisions presented here address three of the guiding research questions:

- **RQ1** - *What do we want to show?*, by defining the relevant multi-modal data signals
- **RQ2** - *How do we want to show it?*, by developing tailored visual representations for high-stress communication
- **RQ3** - *How do we maintain analytical coherence across views?*, by establishing synchronized, interactive links between views through a shared temporal and contextual structure

The effectiveness of these design choices is assessed in Chapter 5, where task outcomes, cognitive workload (NASA-TLX), usability (PSSUQ), and qualitative feedback are used to evaluate learnability, interpretability, and the coherence of the integrated multi-view design.

Following the Design Study Methodology (DSM) [SMM12], an established methodology in visualization research, the design process combined iterative problem characterization, abstraction, and visualization design informed by prior domain research [OUR+25, MUA+24]. The design emphasizes the integration of multi-modal signals into intuitive, temporally aligned views that lower barriers to reflection and sense making. This chapter documents the corresponding discover, design, and implement stages of DSM, while later chapters address evaluation with non-domain trainees as a formative usability assessment.

3.1. Goals and Design Objectives

Goals. Building on the design directions established in the Roger dashboard [OUR+25], the system addresses four needs identified through prior domain research: first, making participation patterns visible over time, drawing on timeline-based encodings such as stacked streamgraphs [BW08]; second, grounding each utterance in its operational and emotional context, informed by situation awareness theory [End95], which highlights the role of perceiving, comprehending, and projecting task-relevant information; third, allowing side-by-side comparison of individual and team-level behaviour; and fourth, supporting the capture and revisiting of key moments through embedded annotations, a practice shown to aid both personal reflection and shared review [Mar97].

Design Objectives. To address these goals, the dashboard was guided by five interrelated design objectives.

First, the system needed to support sense-making. High-stress training sessions generate dense, multi-party communication, so the interface presents time-aligned structural features, such as turn-taking and overlaps, together with a stress indicator, enabling trainers to obtain an at-a-glance situational overview without losing temporal context.

Second, the design emphasizes *context-content integration*. Messages, topics, and stress values are bound to a single synchronized timeline, ensuring that every utterance can be interpreted within its operational and emotional frame rather than in isolation. This directly addresses the linking principle of **RQ3**, which concerns maintaining temporal and contextual connection between multi-modal data streams.

Third, the dashboard aims to foster *comparative reflection*. Instructors and trainees can inspect both individual- and team-level traits, such as contribution balance, clarity of expression, or the frequency of questions asked, allowing structured debriefs that move beyond individual utterances toward identifying broader communication patterns.

Fourth, the system was designed for *seamless navigation*. Brushing and linking interactions are applied consistently across all visual views, with synchronized media playback anchoring the analysis in the original recording. This consistency operationalizes another aspect of **RQ3**, enabling fluid transitions between abstract metrics and concrete conversational evidence.

Finally, the dashboard aims to support actionability. Trainers can annotate both events and intervals, capturing noteworthy moments that can later be reused as teaching artifacts. This capability allows trainers to build a library of examples grounded in actual session data, supporting structured debriefing and longitudinal tracking of trainee development.

Together, these objectives translate the overarching goals into concrete design principles that help make the dashboard both analytically powerful and pedagogically useful.

3.2. Shared Timeline and Data Signals

This section defines **what the system shows**, addressing **RQ1**, by outlining the multi-modal data streams integrated into a single synchronized timeline. The data used to populate these signals are derived from the *Sound of the Police* VR training system [MUA+24], with detailed data handling and preprocessing described in Section 4.2.

All components of the dashboard are linked to a **shared session timeline** $t \in [0, T]$, ensuring that structural, content-based, and affective signals can be explored in context.

Speech activity. Speech segments that have been automatically separated and labeled by speaker (a process known as diarization), with precise start-end times, form the backbone of the timeline. These segments support both a streamgraph view showing how much each trainee spoke relative to others (speaking share), and the detection of overlapping turns.

Messages. Each utterance is represented as a timestamped message with associated duration and transcript text. These messages drive a Chat Bubbles visualization and a searchable conversation log.

Stress estimates. Per-trainee stress values are provided as continuous time series aligned to the session timeline. These data feed into both the heat encodings in the Chat Bubbles view and the Stress Lines plots, making fluctuations in affect immediately visible.

Topics. Automatic message-to-topic assignments and trainee-level contribution weights capture the thematic structure of the conversation. These are visualized in a Topic Tab to highlight which trainees contributed to which themes and when.

Derived traits. Higher-level behavioral indicators, such as contribution share, number of questions, clarity proxies, or interruption counts, are computed from the raw speech and message data. These derived metrics populate the Summary Tab and support comparative reflection at the trainee and team levels.

Media. The original audio (inline player) and video (external tab) are tightly coupled to the global timeline, allowing any visual inspection to be grounded in the source recordings.

Annotations. Trainer-created point or interval annotations, enriched with tags and notes, are embedded into the timeline. These annotations appear consistently across all views, supporting externalization of observations and creation of reusable teaching artifacts.

By binding these heterogeneous signals to a single temporal frame, the dashboard directly fulfills the requirement from **RQ3**, maintaining consistent linkage between data modalities so that trainers can fluidly move between abstract visual patterns and concrete evidence in the original recordings.

3.3. Dashboard Architecture and Tab Designs

This section addresses **RQ2**, *How do we want to show it?*, by describing the visual and interaction designs that present the multi-modal data streams defined above. The chosen design emerged from the need to support rapid, evidence-based reflection in high-stress training scenarios such as police academy role-plays or first-responder drills. To reiterate, in such contexts, trainees and instructors must review complex, multi-party interactions under time constraints, often with strong emotional content and shifting team dynamics.

The dashboard follows a modular architecture consisting of three main **tabs** (*Flow*, *Topic*, and *Summary*) and several **shared components** that remain accessible across all views. Each component contributes to a coordinated multi-view system linked through the global timeline defined in Section 3.2, ensuring the interconnectedness emphasized in **RQ3**.

Consistent with the Design Study Methodology (DSM) [SMM12], this section documents both our design choices and the alternatives we considered during the *reflect* and *winnow* stages. In DSM, *reflect* involves evaluating design decisions against project goals, while *winnow* involves narrowing down options by filtering out less promising approaches. By making these decisions explicit, the aim was to contribute not only a working system but also transferable design knowledge for future visualization efforts in high-stress training domains. While presented here as a structured narrative, in practice the process cycled repeatedly between *discover*, *design*, and *reflect* stages as insights from instructors and pilot sessions informed successive refinements.

3.3.1. Overall Structure

Tabs

1. **Flow Tab:** Displays speaking activity, utterance-level details, and per-trainee stress over time. It integrates a Streamgraph (speaking share), Chat Bubbles (utterance duration, overlap, and stress heat), and Stress Lines (individual stress trends).

2. **Topic Tab:** Reveals thematic focus and trainee engagement using a time-based topic streamgraph and a radial Topic Contribution view that visualizes who discussed what, and under which emotional conditions.

3. **Summary Tab:** Aggregates high-level behavioral and affective indicators through a Trait Radar (coach-oriented metrics) and Topic Share Bars (thematic distribution).

Shared Components (shared across all tabs) These components remain accessible across all views, providing continuous access to core functionality. The conversation log offers a searchable transcript with stress badges, where clicking any message highlights the corresponding region across all views and synchronizes playback. This design preserves the verbatim record, which was critical for evidence-based review, reflecting Shneiderman's mantra of "overview first, zoom and filter, then details-on-demand" [Shn03], as well as Pirolli and Card's [PC05] model of sense-making loops. An alternative approach using a pop-up transcript was tested but introduced unnecessary navigation overhead.

Media players include an inline audio player and an optional external video player, both synchronized with the global timeline. Synchronizing audio and video with abstract views anchors interpretation in the original behavioral context. Tight coupling between representations supports more accurate analysis [Rob07], and video-linked data views can support communication analysis [WO99]. Instructors emphasized that this ability to correlate stress spikes or overlaps with tone and body language was essential. Static screenshots were considered as an alternative but lacked the nuance of full playback.

Filters provide controls for trainee selection, time range, stress overlay, and topic-based filtering. Direct manipulation filters and brushing allow focus on subsets of trainees or time ranges while preserving context. Prior work shows that brushing and linking interactions enhance exploratory analysis by reducing cognitive overhead [YaKSJ07]. Predefined bookmarks were considered as an alternative but were too rigid for open-ended exploration.

The annotation tool enables point or interval annotations enriched with tags and notes, with annotations appearing across all views to support structured, evidence-based reflection, a key component of effective simulation-based training [SRHW08]. Annotations support external cognition, allowing trainers to offload and share insights. This principle is well established in research on annotation practices [Mar97], and annotation practices can facilitate collaborative review [Mar98]. External note-taking was rejected as an alternative due to the risk of desynchronization with the dashboard.

All components are linked via brushing and highlighting interactions, ensuring consistent cross-view synchronization, again addressing the integration principle from RQ3. The global timeline serves as the backbone of the interface, maintaining alignment between transcripts, stress data, and video playback. This coordinated multiple view design was adapted following established principles for linking related visualizations [Rob07]. By synchronizing views, trainees can move fluidly between perspectives without reorienting themselves in time, which proved essential in time-pressured debriefs. A single composite visualization that merged speech, topics, and stress was tested as an alternative, but pilot trainers found it overwhelming.

3.3.2. Flow Tab

Figure 3.1.: Flow Tab showing (top) Streamgraph of speaker activity, (middle) Chat Bubbles with stress heat, and (bottom) per-trainee Stress Lines. A brushed interval updates all views and seeks media playback.

Streamgraph (speaker activity). The Streamgraph (Figure 3.2) encodes speaking intensity over time using stacked, color-coded areas for each trainee [BW08]. It reveals dominance, turn-taking, and pauses at a glance. Hovering over the graph exposes the speaker identity and their proportional share, while brushing a region sets the global selection and synchronizes all linked views.

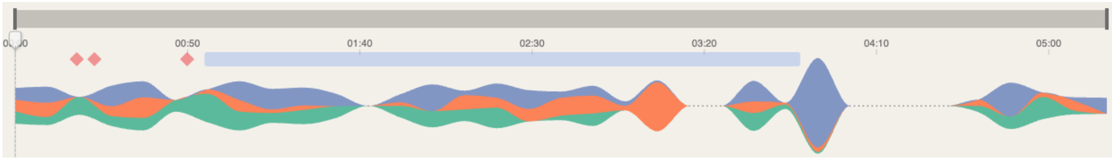


Figure 3.2.: Streamgraph of speaking activity. Hover reveals trainee share; brushing selects an interval and synchronizes all views.

The streamgraph was selected because of its ability to show participation balance and turn-taking patterns over time. Streamgraphs effectively depict changing distributions and highlight dominance shifts [BW08]. A stacked-bar timeline was considered, but binning introduced artificial boundaries that obscured smooth transitions and speaker overlap, reducing interpretability.

Chat Bubbles (overlap, duration, content, stress). The Chat Bubbles view (Figure 3.3) positions utterances along the timeline according to their timestamp, with lanes or slight jitter separating trainees. Bubble size encodes the duration or length of the utterance, while overlaps indicate simultaneous speech. A background heat layer represents the local stress level at the time of speaking.

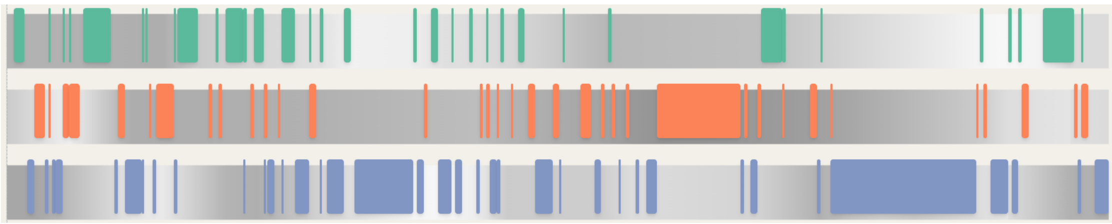


Figure 3.3.: Chat Bubbles view showing overlap, duration, and stress heat for each utterance.

When hovering over a bubble (Figure 3.4), a tooltip reveals the full message text of the utterance. Clicking a bubble seeks the media playback to that exact moment.

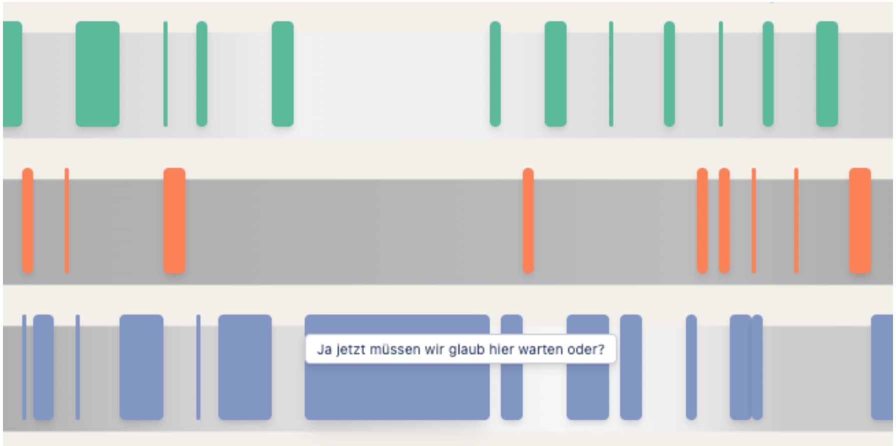


Figure 3.4.: Hover interaction reveals the message text of the utterance.

The bubble-based encoding leverages the familiar structure of messaging interfaces, reducing cognitive load by aligning with trainers' existing mental models [Nor13]. Text-only transcripts or stress as a separate chart were tested as alternatives, but integrating both into one view reduced context switching and improved recall during pilot sessions.

Stress Lines (per-trainee). The Stress Lines view (Figure 3.5) plots one line per trainee, showing stress values over time.

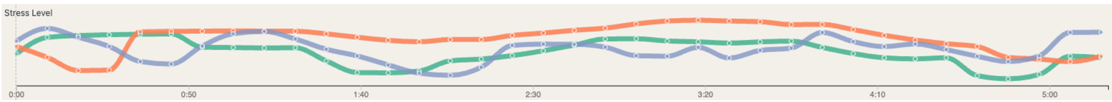


Figure 3.5.: Per-trainee stress lines.

When hovering over a stress line (Figure 3.6), a tooltip displays the exact stress value and timestamp, allowing for precise inspection.

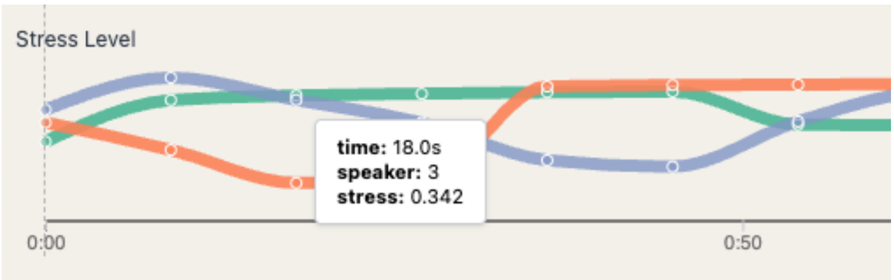


Figure 3.6.: Hover interaction showing the exact stress value and timestamp for a given point on the line.

Stress lines were chosen because line graphs effectively reveal temporal trends and spikes while maintaining compactness and clarity [Tuf01]. A stacked area chart was rejected as an alternative because it obscured individual trajectories by conflating them into a cumulative trend.

3.3.3. Topic Tab

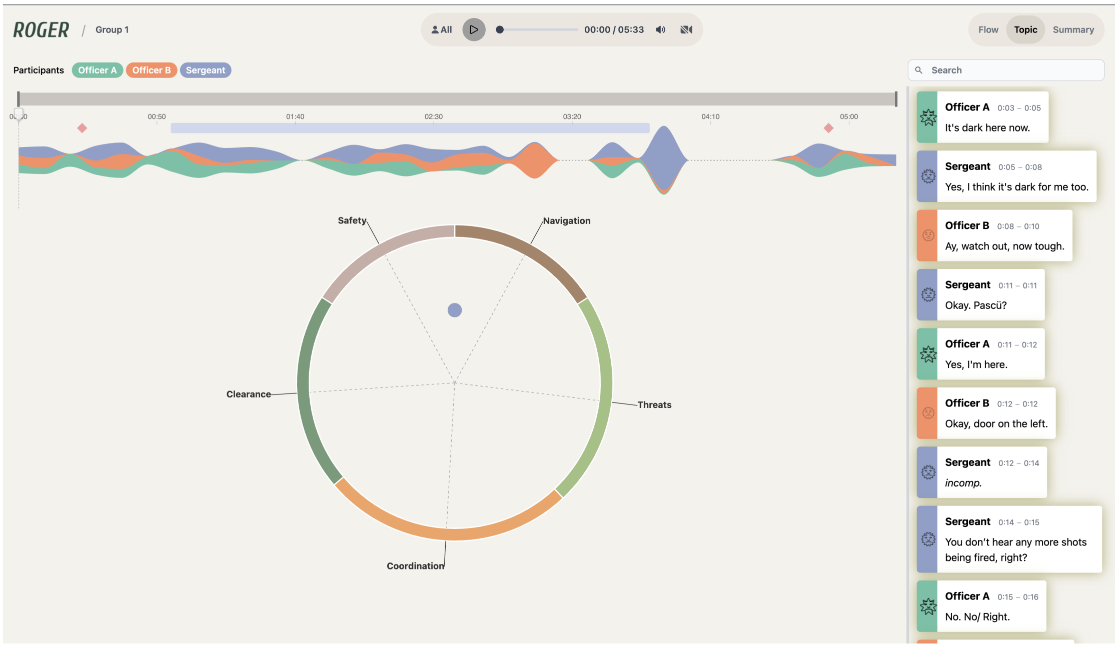


Figure 3.7.: Topic Tab showing trainee contributions per topic with time-of-mention indicators.

The Topic Tab integrates temporal and relational perspectives on thematic structure. A temporal streamgraph shows the evolution of topics over time, while a radial Topic Contribution view (adapted from *ConToVi* [EAGA+16]) maps trainees (central nodes) to topic segments in a circular layout. Each topic is represented as a segment in a donut, while trainees appear as nodes in the center whose proximity to segments reflects their relative engagement. This pairing reveals both when topics surfaced and who drove them, supporting reflection on topical equity and focus.

Hover interactions connect the two perspectives: hovering over a topic segment highlights related utterances in the streamgraph, while hovering over a trainee node reveals their contribution proportions and stress context (Figure 3.8). This integrated design supports both temporal and relational reflection on thematic communication patterns.

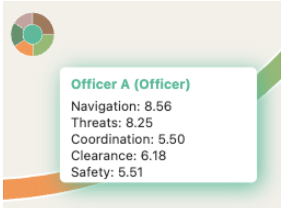


Figure 3.8.: Hover interaction displaying topic contribution values for a specific trainee, along with stress context.

We explored stacked bars, matrices, and simple topic timelines as alternatives, but they either obscured trainee-topic relationships or required frequent context-switching between separate temporal and relational views. The integrated streamgraph combined with radial view provided clearer, at-a-glance insight into both temporal development and relational engagement patterns.

3.3.4. Summary Tab

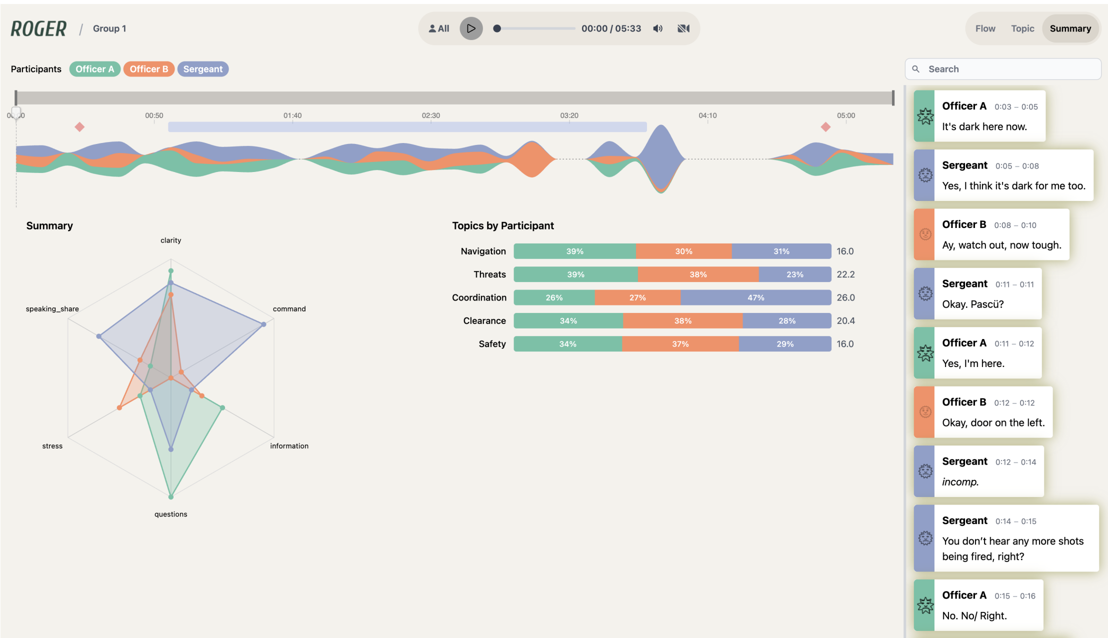


Figure 3.9.: Summary Tab combining Trait Radar (left) and Topic Share Bars (right). This tab provides a compact overview of comparative performance and stress dynamics.

The Summary Tab provides an at-a-glance overview of individual and team traits, combining trait-oriented metrics with thematic distribution. It abstracts away from fine-grained events toward higher-level summaries suitable for debriefs.

Trait Radar. Radar axes summarize coachable traits (e.g., questions asked, clarity, speaking share, interruptions, on-topic rate, average stress). Multiple trainees can be overlaid for comparison, and hovering over a vertex reveals the trainee name and exact value for that metric.

Figure 3.10.: Trait Radar comparing trainees; hovering over a vertex reveals the participant name and value.

Radar charts were chosen for their ability to display multiple traits in a compact form that supports direct visual comparison through superposition [GAW+11]. While radar charts have known interpretability limitations when axes overlap or when more than a few profiles are compared, pilot sessions indicated that instructors found the multivariate overlay intuitive for identifying imbalances between two or three trainees. Parallel coordinates were considered as an alternative but were judged less accessible for non-technical trainers in our context.

This visualization uses length encoding, which Cleveland found to be among the more perceptually accurate channels for comparing proportions [CM84].

3.3.5. Design Reflection

High-stress interactions are marked by rapid turn-taking, emotional volatility, and shifting situational awareness. Research on simulation debriefing emphasizes that effective learning requires structured phases addressing emotional reactions, systematic analysis of events, and connection to future practice [FG07, DFS+13]. Our design extends these principles to communication visualization by making temporal relationships and affective patterns visually explicit rather than relying solely on verbal reconstruction.

The dashboard addresses these needs by integrating structural, content, and affective signals in a single workspace. Trainers can navigate quickly and flexibly without losing temporal context, moving between macro-level patterns such as flow and dominance and micro-level details such as individual utterances and stress peaks. Synchronized media playback and annotation tools enable evidence-based feedback grounded in concrete moments from the training session.

These design rationales and rejected alternatives illustrate the iterative, non-linear character of visualization design emphasized in DSM [SMM12]. The choices reflect a balance between analytical power, cognitive accessibility, and practical deployment constraints in high-stress training environments.

Having established the design rationale and visual architecture of the dashboard the following chapter describes its technical realization. Chapter 4 details the technology stack, data processing pipeline, deployment infrastructure, and engineering decisions that translate these design principles into a functional, accessible prototype suitable for both controlled evaluation and future replication by other researchers.

4. Implementation

The implementation of the dashboard was guided by the design requirements outlined in Chapter 3, with a focus on interactivity, responsiveness, and deployment feasibility for both controlled usability studies and reuse or adaptation by future researchers. The system was implemented using modern web technologies, ensuring a smooth development workflow, maintainability, and ease of deployment.

This chapter outlines the technical realization of the dashboard across six key aspects. Section 4.1 describes the technology stack used to build the interface and visualization components. Section 4.2 details the data sources and preprocessing pipeline that enable the integration of multimodal communication data. Section 4.3 explains the deployment and hosting setup, followed by a description of the version control and reproducibility framework. Section 4.5 discusses the usability study integration and interaction logging used during the evaluation. Finally, Section 4.6 presents the engineering rationale underlying key implementation choices.

4.1. Technology Stack

The front-end was implemented using the **Svelte** framework, chosen for its lightweight reactivity model and low runtime overhead compared to virtual DOM based approaches.[1] Because Svelte compiles components into plain JavaScript during build, it avoids maintaining a virtual DOM at runtime and loads faster, which is helpful in environments with unstable or slow network conditions. Its simplicity and responsiveness also make it a good fit for rapid prototyping of interactive visualizations, which matches the exploratory goals of this dashboard.

The build tooling was handled by **Vite**, a modern frontend build tool optimized for rapid development and hot module replacement.[2] This allowed for fast iteration during the prototyping and testing phases, particularly when refining interactive visualization components.

Visualizations were implemented using a combination of **D3.js** for low-level, highly customized rendering and **Plotly.js** for certain higher-level charting needs. The choice of technology was driven by the need for:

- *Fine-grained control* over bespoke visualizations such as the streamgraph and chat bubbles.

[1]Svelte documentation: `https://svelte.dev`
[2]Vite documentation: `https://github.com/vitejs/vite`

- *Declarative charting* for components like the radar plot, where development speed and responsiveness were prioritized.

4.2. Data Source and Handling

We used structured datasets exported from the *Sound of the Police* VR training system [MUA+24], which records speech segments, transcripts, and physiological stress estimates during simulated police operations.

For our purposes, the structured datasets from this system were exported as `.csv` files and reformatted into JSON to be consumed by the dashboard. The datasets included speech segments, transcript text, and stress estimates aligned to a common timeline. Importantly, we did not implement a live backend or real-time data streaming pipeline; instead, our focus was on prototyping the visualization layer of the dashboard. This allowed us to test design ideas rapidly without being constrained by engineering overhead. As such, the current version of the dashboard should be considered a *prototype*, emphasizing design exploration rather than production deployment.

4.3. Deployment and Hosting

For usability study deployment and public accessibility, the dashboard was hosted on **Firebase Hosting**. This platform was selected for its seamless integration with static site deployment, built-in HTTPS, and global content delivery network (CDN).

The deployed version of the prototype can be accessed at: `https://roger-thesis.web.app/`

The deployment workflow consisted of:

1. Building the production bundle with Vite.
2. Deploying to Firebase Hosting via the Firebase CLI.
3. Conducting browser compatibility testing across Chrome, Firefox, and Edge to ensure cross-platform usability.

This setup ensured consistent performance and accessibility during the user study and allows external reviewers or future researchers to explore the interactive prototype directly.

4.4. Version Control and Reproducibility

The full source code, along with configuration files and example datasets, is hosted in a public **GitHub** repository.[3] This enables replication of the system for further research, as well as adaptation for related domains. The repository includes:

[3] `https://github.com/RohamJabbari/roger`

- Source code for all dashboard components.
- Sample JSON data for demonstration and testing.
- Deployment scripts and environment configuration templates.
- Documentation for installation, data preparation, and running the dashboard locally.

4.5. Usability Study Platform Integration

For the evaluation, participants accessed the dashboard directly through the Firebase-hosted instance via web links. This approach eliminated the need for software installation while ensuring consistent performance and versioning across all sessions. Basic client-side logging was implemented to capture interaction events (e.g., tab switches, filter applications, annotation creation), but these logs served only to verify task completion and were not included in the formal analysis.

4.6. Rationale for Engineering Choices

The engineering choices reflect a balance between prototyping speed, usability, and long-term reproducibility.

The prototype scope was deliberately limited to front-end interactivity and dashboard design, omitting real-time data streaming and backend services. This constraint made it possible to validate the design principles in a realistic but controlled setting without the overhead of complex infrastructure. By focusing on the visualization layer, we could iterate rapidly on design alternatives and gather feedback on usability and analytical value before committing to production-level engineering.

The choice of Svelte and Vite reflects modern priorities for lightweight, high-performance interactive visualization tools. Svelte, in particular, is well-suited for prototypes due to its low overhead and rapid iteration cycle, enabling quick refinements during the design and testing phases. Vite complemented this by providing hot module replacement and fast builds, reducing friction during development.

Cross-system usability was supported by avoiding platform-dependent technologies and deploying as a static web app. This approach makes it possible for the dashboard to run on most modern systems with a browser, improving accessibility for both participants and future researchers. The simplicity of the deployment model also reduces technical barriers for adaptation in other research contexts.

Finally, hosting the codebase publicly supports reproducibility and reuse, aligning with FAIR principles for research artifacts [WDA+16]. By providing not only the source code but also sample data and deployment documentation, we enable other researchers to inspect, reuse, and build upon the system design and implementation.

With the technical foundation established, the following chapter shifts focus from implementation to evaluation. Chapter 5 describes the formative user study conducted to assess the dashboard's usability, cognitive load, and capacity to support reflective insight during post-training analysis. The evaluation investigates whether the design decisions and technical implementation successfully translate into a system that non-expert users can effectively learn, operate, and employ for meaningful reflection on high-stress communication scenarios.

5. Evaluation

This chapter presents the evaluation of the proposed dashboard, focusing on usability, cognitive workload, and reflective value in post-training analysis. It begins by outlining the evaluation objectives (Section 5.1) and describing the participant sample (Section 5.2), followed by the procedure (Section 5.3) and tasks (Section 5.4) used to assess learnability and interaction flow. The chapter then introduces the measurement instruments (Section 5.5), including the Post-Study System Usability Questionnaire (PSSUQ), NASA Task Load Index (NASA-TLX), and post-session open-ended feedback. Subsequently, the results and analysis (Section 5.6) present quantitative and qualitative outcomes across usability, workload, task performance, and participant perceptions. Finally, the chapter concludes with findings and interpretation (Section 5.7) and an overarching discussion (Section 5.8), summarizing implications for training design and lessons learned from the dashboard's redesign.

5.1. Evaluation Objectives

The evaluation was designed to assess the dashboard's ability to support post-training reflection, usability, and cognitive efficiency in high-stress communication analysis scenarios. Our focus was threefold:

1. **Post-Training Reflection:** Determine whether the dashboard enables users to replay and inspect emotionally or cognitively charged moments, fostering insight into team dynamics, decision-making, and communication under stress.

2. **Usability and Interaction:** Measure how intuitive, clear, and learnable the interface is, with emphasis on navigation between visualizations, annotation tools, and filtering features.

3. **Cognitive Load:** Evaluate the mental effort required to interpret and operate the dashboard's visualizations, and identify components that impose higher workload.

5.2. Participants

A total of 15 participants took part in the study. Unlike the intended target groups (police trainees, emergency responders, and professional trainers), the actual sample consisted of non-domain participants recruited from a convenience population. This choice reflects the prototyping and formative evaluation stage of the work, where the primary goal was to assess usability, clarity, and interaction design rather than domain-specific training effectiveness. It also partially reflects access constraints to the target expert population. As a result, the findings should be interpreted as evidence of general usability and learnability rather than as validation of the system's impact in operational training contexts.

ID	Age	Gender	Occupation	Experience
501	25	Female	Business Analyst	4–6 years
502	21	Male	Student	1–3 years
503	29	Male	Founder / Process Improvement Manager	1–3 years
504	25	Female	Student	4–6 years
505	29	Male	Trainee (Auszubildender)	<1 year
506	32	Male	Operations Manager	4–6 years
507	29	Male	Student	4–6 years
508	31	Female	Team Lead	1–3 years
509	31	Female	Student	4–6 years
510	28	Male	Software Engineer	<1 year
530	24	Female	Data Integration Engineer	4–6 years
531	40	Male	Delivery Driver Dispatcher	4–6 years
532	40	Male	Student	>10 years
533	42	Female	Student	4–6 years
534	28	Male	Architect	1–3 years

Table 5.1.: Demographic information of study participants.

The sample included 9 males and 6 females, with ages ranging from 21 to 42 years (mean age: 30.3). Participants came from diverse professional backgrounds (see Table 5.1), including students, engineers, managers, and an architect. Their reported work experience ranged from less than one year to over ten years. While not directly representative of the intended end-users, this group was able to easily understand the tasks and provide valuable feedback on usability and design, which informed the refinement of the dashboard.

5.3. Procedure

Participants joined the study remotely via video call. Each session began with an introduction to the study purpose, followed by informed consent and a brief demographic questionnaire.

The study followed a within-subjects design with two phases, using equivalent but distinct task sets to prevent learning effects:

1. **Unguided Interaction:** Participants completed analytical tasks without prior explanation of the dashboard. This phase assessed first impressions, discoverability, and initial learnability.

2. **Guided Interaction:** After a brief walkthrough of each dashboard feature, participants completed an equivalent set of tasks. This phase assessed usability once the system's features were known.

Observations were made throughout, noting hesitation points, navigation patterns, and instances of confusion. Sessions concluded with questionnaires and an opportunity for open-ended feedback. Each session lasted approximately [40] minutes. Participants did not receive compensation.

5.4. Tasks

Participants completed a structured sequence of analytical tasks designed to evaluate discoverability, interpretability, and efficiency across the three main dashboard tabs (*Flow*, *Topic*, and *Summary*). The tasks simulated realistic post-training analysis activities, such as identifying key conversational events, inspecting stress patterns, comparing participant contributions, and interpreting topic distributions. Together, they reflected the types of sensemaking and reflection the dashboard is intended to support in operational training debriefs.

Each participant performed the task set in two phases. In the **unguided phase**, they explored the dashboard independently to assess its immediate learnability and clarity. In the **guided phase**, they completed a new but equivalent set of questions, similar in structure and difficulty but referring to different time points, participants, or topics, after receiving a short walkthrough of each feature. This design allowed us to compare performance and workload between first-use discovery and informed use.

In the *Flow* tab, tasks focused on locating and tagging emotionally charged events, retrieving stress values, and interacting with the annotation and filtering features. In the *Topic* tab, participants analyzed thematic focus and contribution balance, using search and filtering to observe topic-specific changes. In the *Summary* tab, they compared participants' behavioral and topical metrics, including stress balance, speaking share, and question frequency.[1]

5.5. Measures

To comprehensively evaluate the system, we employed four complementary measures capturing task performance, cognitive workload, usability, and qualitative feedback:

- **Task Outcomes**, recorded as binary completion (correct/incorrect) for each analytical task. These provided objective performance indicators for comparing unguided and guided phases.

[1] The full set of task instructions and tab overviews, as presented to participants in the study interface, are provided in Appendix A.1.

- **NASA Task Load Index (NASA-TLX)** [HS88], a multidimensional rating tool that measures subjective workload. While the standard NASA-TLX includes six dimensions, we report four subscales: mental demand, performance (perceived success), effort, and frustration. Physical demand was excluded because dashboard interaction is a non-physical task (all participants rated this dimension at or near zero). Temporal demand was excluded because tasks were untimed and participants could work at their own pace.[2]

- **Post-Study System Usability Questionnaire (PSSUQ)** [Lew92], which assessed perceived *system usefulness*, *information quality*, and *interface quality*. This standardized instrument is designed for complex, multi-component systems and provides diagnostic insight into usability and design effectiveness.[3]

- **Open-Ended Questions**, which collected qualitative reflections on usability, interpretability, and the value of dashboard components. Responses provided context to the quantitative results and revealed user perceptions of system clarity and usefulness.[4]

Quantitative measures (PSSUQ and NASA-TLX) were administered after both the unguided and guided phases, while qualitative feedback was gathered at the end of the guided phase. The following section presents the corresponding results for each measure, followed by overall interpretation and discussion.

Statistical approach. Given the small sample size (N=15) and the formative nature of this evaluation, we report descriptive statistics only. Comparisons between conditions and task types are presented as observed differences without inferential statistical testing, which would require larger samples to achieve adequate power. The results should therefore be interpreted as indicative patterns that inform design refinement rather than as generalizable population estimates.

5.6. Results and Analysis

This section presents the empirical results derived from the evaluation study, combining quantitative performance measures with subjective workload and usability ratings, as well as qualitative participant feedback. The results are organized into four parts: **Task Outcomes**, which capture objective performance indicators such as completion rate and accuracy; **Workload (NASA-TLX)**, which assesses perceived cognitive and physical demand; **Usability (PSSUQ)**, summarizing participants' satisfaction across key usability dimensions; and **Open-Ended Feedback**, which provides qualitative insights into perceived strengths, challenges, and design implications of the dashboard.

[2] See Appendix A.3 for the complete NASA Task Load Index (NASA-TLX) instrument.
[3] See Appendix A.4 for the complete Post-Study System Usability Questionnaire (PSSUQ) instrument.
[4] See Appendix A.5 for the full set of post-session open-ended questions.

5.6.1. Task Outcomes

As shown in Figure 5.1, task completion rates improved substantially between the unguided and guided phases across all three task categories, indicating a strong effect of brief onboarding on task performance.

In the unguided phase, participants completed 76.7% of tasks correctly on average. After receiving a brief walkthrough of the dashboard, this rose to 96.7%, a 20 percentage point improvement. The pattern held across all task types: Flow tasks improved from 80.0% to 96.7%, Topic tasks from 70.0% to 100.0%, and Summary tasks from 80.0% to 93.3%.

The Topic tasks showed the largest gain, jumping 30 percentage points after guidance. This aligns with qualitative feedback indicating that the radial Topic Chart required the most interpretation effort during initial exploration. Once participants understood how to read the visualization, their performance reached ceiling levels.

These results demonstrate that brief onboarding (approximately 5 minutes) substantially improved task accuracy, suggesting the dashboard can be effectively adopted in similar analytical settings with minimal instruction.

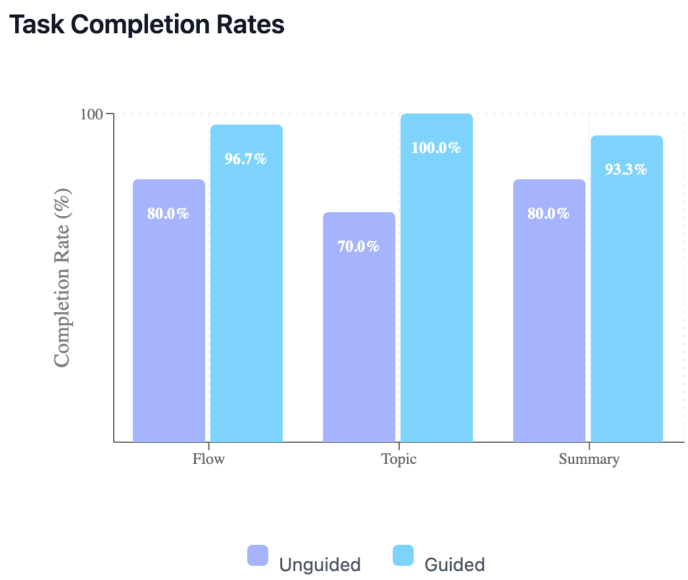


Figure 5.1.: Task completion rates by phase (Unguided vs. Guided) across Flow, Topic, and Summary task categories. All three categories show substantial improvement after brief guidance, with Topic tasks exhibiting the largest gain.

5.6.2. Workload (NASA-TLX)

Figure 5.2 compares self-reported workload scores across the three task categories (*Flow*, *Topic*, and *Summary*). Overall, perceived workload was low to moderate, with clear differences between task types that align with the visual complexity of each dashboard component.

Topic tasks imposed the highest cognitive demand, with mental demand averaging 38.9 (out of 100), effort at 32.7, and the lowest perceived success rate at 68.3. This pattern aligns with participant comments that the radial Topic Chart required more interpretation and imposed higher cognitive load than other views.

Flow tasks showed moderate workload levels: mental demand averaged 25.8, effort 15.7, and perceived success 79.9. The Flow tab's timeline-based visualizations (streamgraph and chat bubbles) were more immediately interpretable, likely due to their resemblance to familiar interfaces.

Summary tasks were perceived as easiest, with mental demand at only 13.1, effort at 6.3, and the highest perceived success at 83.9. The aggregated metrics and straightforward bar charts in the Summary tab required minimal interpretation effort.

Notably, frustration remained minimal across all task types: 9.1 for Flow, 11.3 for Topic, and just 2.5 for Summary. This indicates that despite differences in cognitive demand, participants rarely felt discouraged or overwhelmed. The pattern suggests users were cognitively engaged rather than struggling with the interface itself.

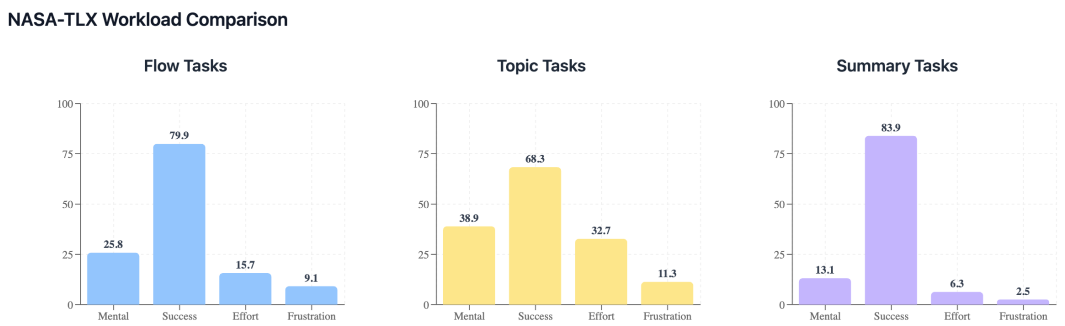


Figure 5.2.: NASA-TLX subscale scores across task categories. Bars represent mean scores (0–100) for Mental Demand, perceived Success, Effort, and Frustration. Topic tasks show notably higher cognitive demand while Summary tasks required minimal effort.

The threefold difference between Topic and Summary workload scores is substantial. However, given that Topic tasks also showed the largest improvement in completion rates after guidance (70% to 100%), this elevated workload reflects interpretive complexity rather than poor usability. Once participants understood the radial encoding, they performed at ceiling levels despite the higher cognitive investment required.

These workload levels indicate moderate cognitive engagement without high frustration, suggesting that higher effort did not coincide with increased frustration.

5.6.3. Usability (PSSUQ)

Figure 5.3 summarizes the PSSUQ results, showing consistently high usability ratings across all subscales. Participants reported strong agreement with positive statements about the system's usefulness (mean = 6.36 out of 7), information quality (mean = 5.69), and interface quality (mean = 6.33), with an overall satisfaction score of 6.60. These results suggest that participants found the dashboard intuitive, informative, and pleasant to use, even without prior domain experience.

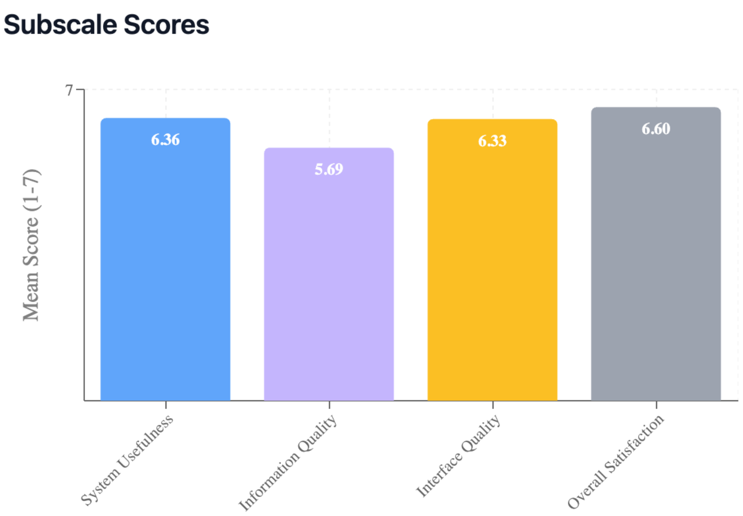


Figure 5.3.: PSSUQ subscale scores. Mean ratings (1-7) for System Usefulness, Information Quality, Interface Quality, and Overall Satisfaction. Higher scores indicate better perceived usability.

Figure 5.4 provides a detailed breakdown of individual question scores. The highest rated items were Q6 ("I believe I could become productive quickly using this system") and Q11 ("The information was effective in helping me complete the tasks"), both scoring 6.73. Q5 ("It was easy to learn to use this system") also scored highly at 6.60. Together, these results indicate that participants felt confident in their ability to learn and use the dashboard effectively.

The lowest rated item was Q7 ("The system gave error messages that clearly told me how to fix problems"), which scored only 3.20, well below the neutral midpoint of 4.0. This represents a significant gap: the prototype did not include error messages or feedback prompts for invalid actions, leaving users without guidance when interactions failed. Q8 (error recovery, 5.93) and Q9 (documentation clarity, 5.73) also scored lower than other items, pointing to the same deficiency in user guidance features.

The lower subscale score for Information Quality (5.69 compared to 6.36 for System Usefulness) is driven primarily by the Q7 deficiency. This represents a clear priority for future development: implementing meaningful error feedback and interaction guidance.

Question Details

QUESTION	DESCRIPTION	MEAN ± SD	RANGE	RATING	PERFORMANCE
Q1	Overall, I am satisfied with how easy it is to use this system	**6.00** ±0.52	5 - 7	excellent	86%
Q2	It was simple to use this system	**6.40** ±0.71	5 - 7	excellent	91%
Q3	I was able to complete the tasks and scenarios quickly using this system	**6.20** ±0.65	5 - 7	excellent	89%
Q4	I felt comfortable using this system	**6.20** ±0.75	5 - 7	excellent	89%
Q5	It was easy to learn to use this system	**6.60** ±0.49	6 - 7	excellent	94%
Q6	I believe I could become productive quickly using this system	**6.73** ±0.44	6 - 7	excellent	96%
Q7	The system gave error messages that clearly told me how to fix problems	**3.20** ±1.22	1 - 4	poor	46%
Q8	Whenever I made a mistake using the system, I could recover easily and quickly	**5.93** ±1.06	4 - 7	good	85%
Q9	The information (such as online help, on-screen messages, and other documentation) provided with this system was clear	**5.73** ±1.69	1 - 7	good	82%
Q10	It was easy to find the information I needed	**6.40** ±0.61	5 - 7	excellent	91%
Q11	The information was effective in helping me complete the tasks and scenarios	**6.73** ±0.44	6 - 7	excellent	96%
Q12	The organization of information on the system screens was clear	**6.13** ±0.88	4 - 7	excellent	88%
Q13	The interface of this system was pleasant	**6.60** ±0.61	5 - 7	excellent	94%
Q14	I liked using the interface of this system	**6.40** ±1.25	2 - 7	excellent	91%
Q15	This system has all the functions and capabilities I expect it to have	**6.00** ±0.89	4 - 7	excellent	86%
Q16	Overall, I am satisfied with this system	**6.60** ±0.49	6 - 7	excellent	94%

Figure 5.4.: Individual PSSUQ question scores. Each row shows the question text, mean score, standard deviation, response range, qualitative rating, and performance percentage. Q7 (error messages) stands out as the main area for improvement.

Several participants mentioned that visual consistency across tabs contributed to ease of learning, supporting the effectiveness of the coordinated design approach. Overall, these results affirm that the dashboard achieved strong baseline usability despite being a research prototype.

5.6.4. Open-Ended Feedback

Participants provided qualitative reflections through open-ended questions about usability, clarity, and perceived value of dashboard components. Responses were coded using an inductive approach by the author, with categories refined iteratively across multiple passes. This section presents their responses organized by theme, supplemented by observation notes taken during the study sessions.

First Impressions and Initial Confusion

Participants had mixed first impressions of the dashboard. Several found it visually appealing but initially overwhelming. Participant 501 noted: *"At first I was confused as there are quite some elements. Certain elements seemed to be more important than others."* Participant 534 echoed this: *"The page's color palette was attractive, but there were too many charts and graphs, and their clutter gave a sense of complexity."* Participant 507 was more direct: *"At first I was thinking why it's so complicated."*

However, others found the dashboard approachable from the start. Participant 502 described it as *"very colorful but futuristic"* and noted that *"it followed the general conventions of software."* Participant 508 found it *"nice, colorful, and not too cluttered. Nicely and simply designed."*

The Topic Tab: Most Challenging Component

The Topic tab emerged as the most cognitively demanding component across both open-ended responses and observation notes. Participants consistently struggled with the radial donut chart layout. Participant 501 explained: *"Topic chart was the hardest one to grasp as it's not a familiar chart."* Participant 507 stated bluntly: *"Topic graphs, I don't get them to be honest! They are really complicated for me."*

Observation notes confirmed this pattern. Four out of 15 participants missed the big donut chart entirely during the unguided phase. The small pie charts within the Topic view were considered less useful than bar charts, and participants frequently confused *"overall"* values with *"over time"* values. Three participants also confused topics with personality traits, suggesting the terminology needed clarification.

Despite these difficulties, participants found the Topic view valuable once understood. Participant 501 acknowledged: *"On the other hand, topic chart was also helpful to have a general understanding what the topics in question are."* Critically, observation notes recorded that after guided explanation, all participants understood the visualization completely.

The Flow Tab: Search Discoverability Issues

The Flow tab was generally well received for its timeline-based design, but a significant discoverability problem emerged: 6 out of 15 participants (40%) missed the search tool entirely. Observation notes documented that some participants used browser search instead of the dashboard's built-in search, while others scrolled through the entire chat log manually or even listened to the recording to find specific content.

Participant 530 highlighted broader confusion with the Flow tab: *"At first, the Flow page is not too intuitive, especially in understanding the functionality of choosing the timestamps, which will update the view."* Participant 531 also found it initially confusing, though the design eventually proved helpful.

Once participants discovered the search functionality, they completed tasks quickly. The marking system was discovered by most participants without guidance, and filters were used effectively. Participant 510 summarized: *"The dashboard was user friendly and I was excited when I was discovering its features."*

The Summary Tab: Intuitive but Limited Interaction

The Summary tab received the most positive feedback for its straightforward presentation. Participant 502 noted: *"The summary was the easiest because it showed useful and concise information on the screen."* Participant 509 found the radar chart particularly helpful, calling it *"Perfection."*

Observation notes identified some issues: the time filter did not affect the summary view as participants expected, and three participants confused topics with traits. However, once found, the radar chart was quickly understood and the bar charts were considered intuitive.

The Guidance Effect

The difference between unguided and guided phases was dramatic. Participant 507 described it as *"night and day,"* explaining: *"After the guide, I understood where I need to look for and where to look for them."* Participant 533 noted: *"In flow and topic tab, I didn't understand the interface in unguided phase, but after the guidance I found the relationships between charts and colors and I understood its concept."*

This pattern appeared consistently across responses. Participant 504 estimated that *"80% was okay and I knew instantly how to use it but of course it was easier when I knew what to exactly look for."* Observation notes confirmed that 100% of participants understood the visualizations after receiving brief explanations.

Valued Features

Several features received consistent praise. The search functionality was valued by those who discovered it. Participant 506 appreciated *"the chat log and its search."* The stress level visualizations were frequently mentioned as helpful. Participant 534 specifically noted *"the stress level bar and the voice volume/speech activity graph"* as useful discoveries.

The synchronized timeline and playback features helped participants make connections they would have missed otherwise. Participant 532 listed multiple discoveries: *"The topic and Trait classification. The stress level line chart helps a lot to track stress level of each participant in comparison to icons used in the chat log."*

Participants also suggested improvements: more visible labels, additional timestamp ticks, hover tooltips showing participant names, and clearer separation between sections. One bug was reported: after filtering by participants, chat bubbles changed position but stress hue did not update accordingly.

Quantitative Summary

The coded themes from open-ended responses, shown in Figure 5.5, confirm the patterns described above. The Topic Component received the most mentions (22), followed by Flow Component (20) and Guidance Helpful (17). Visualization Issues (13) and Confusing Elements (11) also appeared frequently, pointing to specific areas for improvement.

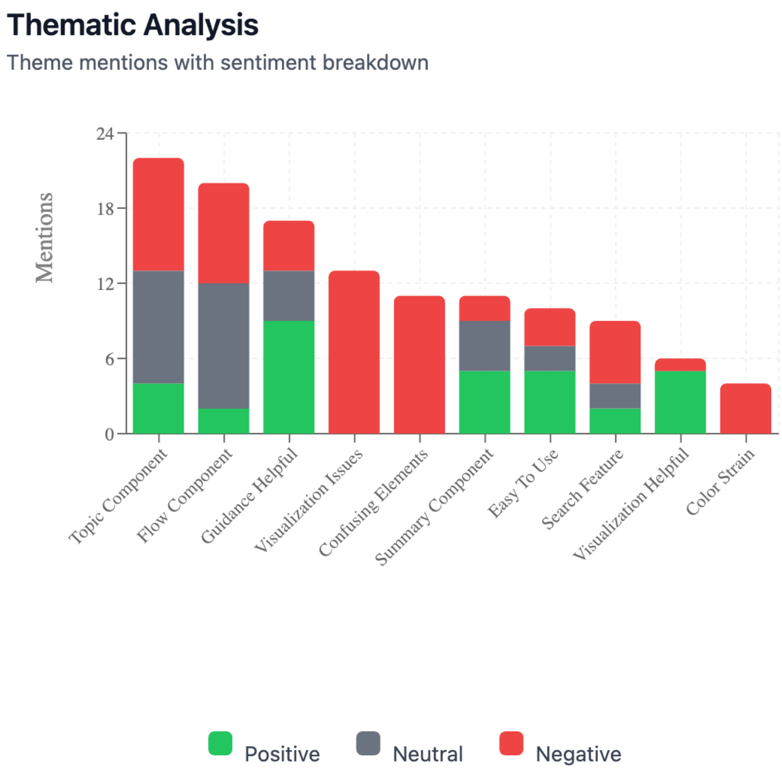


Figure 5.5.: Coded themes from open-ended responses with sentiment breakdown. Topic and Flow components were discussed most frequently, while "Guidance Helpful" reflects the strong learnability effect observed between phases.

Figure 5.6 shows the overall sentiment distribution. Of the coded responses, 53% were positive in tone, 40% were negative (primarily concerning specific component issues rather than overall usability), and 7% were neutral. This relatively balanced distribution indicates that while participants valued certain features, they also encountered notable difficulties, particularly with the Topic Chart and search discoverability.

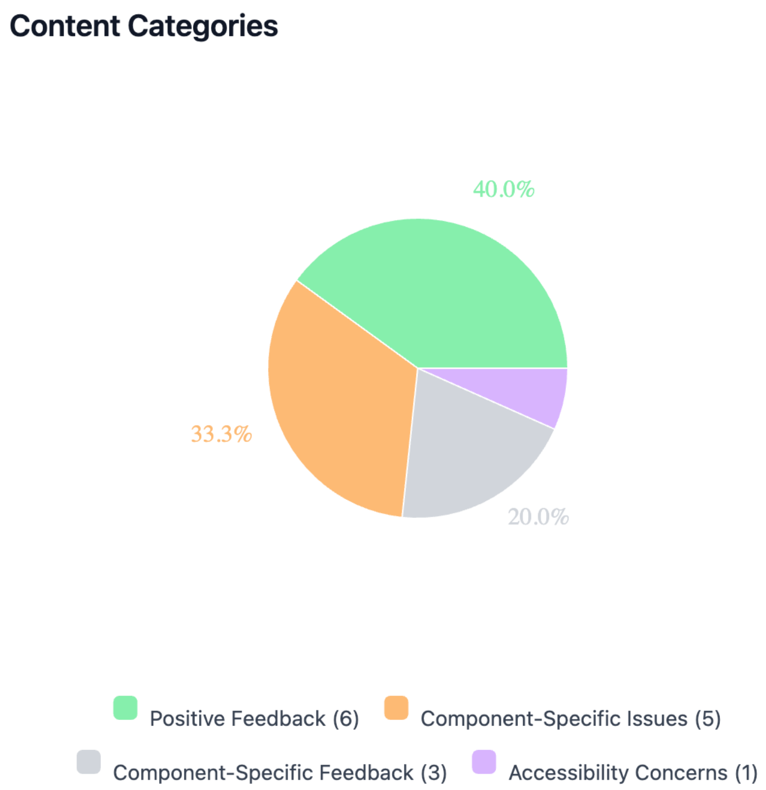


Figure 5.6.: Sentiment distribution across open-ended responses. The majority of feedback was positive despite component-specific challenges.

To supplement the analyses presented above, a separate interactive dashboard was developed to visualize the evaluation results. This companion tool provides an exploratory overview of task outcomes, workload profiles, usability ratings, and thematic feedback across participants. It can be accessed at `https://roger-survey-eval.web.app/` and serves as an interactive extension of the evaluation presented in this chapter.

5.7. Findings and Interpretation

Across all instruments, results indicate that participants perceived the dashboard as highly usable and cognitively manageable once familiar with its layout and interactions.

Task outcomes showed a clear learnability effect. Completion rates improved from 76.7% in the unguided phase to 96.7% in the guided phase. The Topic tab showed the largest improvement, jumping from 70% to 100% after brief guidance. This pattern confirms that the dashboard's complexity is learnable rather than inherently problematic.

NASA-TLX ratings revealed that cognitive demand varied by visualization type. Topic tasks yielded the highest weighted workload (28.7), followed by Flow tasks (17.7), with Summary tasks imposing the lowest demand (9.5). Mental demand for Topic tasks averaged 38.9 compared to 25.8 for Flow and just 13.1 for Summary. These differences align with qualitative feedback: participants consistently described the radial Topic Chart as cognitively heavier due to its unfamiliar layout and color-based contribution mapping.

Importantly, frustration remained minimal across all task types (9.1 for Flow, 11.3 for Topic, 2.5 for Summary). This indicates that the higher cognitive demand in Topic tasks reflected interpretive complexity rather than poor usability. Participants were engaged in meaning-making, not struggling with the interface itself.

PSSUQ scores reflected strong satisfaction across all subscales. System Usefulness scored 6.36 out of 7, Interface Quality 6.33, and Overall Satisfaction 6.60. The lower Information Quality score (5.69) reflects a significant prototype gap: the absence of error messages and system feedback (Q7 scored only 3.20/7). This is a clear priority for future development rather than a fundamental comprehension issue.

The qualitative findings add depth to these patterns. Observation notes revealed specific discoverability problems: 40% of participants missed the search tool entirely, and 27% initially missed the big donut chart in the Topic view. However, 100% of participants understood the visualizations after receiving brief explanations. Participants valued the stress level visualizations, synchronized timeline, and search functionality (once discovered), while consistently identifying the Topic Chart and time marker selection as areas needing improvement.

The combination of high task completion after guidance, low frustration despite moderate cognitive load, and strong usability ratings suggests a balance between interpretive complexity and baseline usability. The challenges participants encountered were localized to specific components and consistently resolved through minimal onboarding.

5.8. Summary

The evaluation results provide empirical evidence addressing RQ4 (How well does it work?) through three complementary lenses: task performance, cognitive workload, and perceived usability.

Task outcomes demonstrated a clear learnability effect. Participants achieved 96.7% accuracy in the guided phase compared to 76.7% in unguided exploration, a 20 percentage point improvement. The Topic tab showed the most dramatic gain (70% to 100%), confirming that even the most challenging visualization becomes accessible with brief instruction.

Workload analysis revealed that cognitive demand varied by visualization complexity. Topic tasks imposed roughly three times the workload of Summary tasks (28.7 vs 9.5

weighted NASA-TLX), with mental demand averaging 38.9 for Topic compared to 13.1 for Summary. However, frustration remained below 12% across all task types, indicating that interpretive effort reflected engagement with complex material rather than poor usability.

Usability ratings were consistently strong. PSSUQ scores exceeded 6.0 out of 7 for System Usefulness (6.36), Interface Quality (6.33), and Overall Satisfaction (6.60). The lower Information Quality score (5.69) reflects a significant gap in error feedback (Q7: 3.20/7), a clear priority for future iterations.

Qualitative feedback identified specific improvement areas: search tool visibility (missed by 40% of participants), Topic Chart discoverability (missed by 27%), and time marker interaction clarity. At the same time, participants valued the stress visualizations, synchronized timeline, and overall visual design. The most consistent finding was the effectiveness of guidance: every participant who struggled with a component in the unguided phase understood it after brief explanation.

Together, these findings indicate that the dashboard successfully supports post-training reflection while maintaining acceptable cognitive load and strong usability. The results inform both the design recommendations and future work discussed in the following chapter.

6. Discussion

This chapter synthesizes findings from the design and evaluation phases and reflects on their broader implications for multimodal communication visualization in high-stress training. Section 6.1 interprets the empirical results in relation to the research questions and compares the dashboard to related systems. Section 6.2 presents design recommendations based on the evaluation findings. Section 6.3 acknowledges methodological and practical limitations encountered during the study. Section 6.4 outlines directions for future research that could extend the system's analytical and pedagogical potential.

6.1. Interpreting the Findings

This thesis set out to design, implement, and evaluate an integrated visualization dashboard for analyzing communication in high-stress training scenarios. The work was motivated by a critical gap identified in Chapter 2: existing visualization systems fragment communication analysis by treating structure, content, and emotional context as separate concerns, preventing trainers and trainees from understanding how stress manifests in actual dialogue. The dashboard developed in this thesis directly aims to address this gap by presenting speech activity, topics, and stress indicators together in coordinated views that share a common timeline, tailored to the needs of high-stress training debriefs.

6.1.1. Answering the Research Questions

The four research questions posed in Chapter 1 guided the design and evaluation of the dashboard. The findings provide answers to each question while revealing nuances that inform future work in this domain.

RQ1: What do we want to show? Building on the literature reviewed in Chapter 2, the design identified structure, content, and affect as three key dimensions for post-training reflection. The dashboard was built around these dimensions, combining speech diarization (who spoke when), topic modeling (what was discussed), and stress estimation (under what emotional state) within a single temporal frame. This approach demonstrated how multimodal indicators can jointly represent the cognitive and social dynamics of high-stress communication. While participant responses were mixed (53% positive, 40% negative), several participants explicitly noted the value of being able to connect structure, topic, and stress within a single view.

RQ2: How do we want to show it? Building on the design rationale presented in Chapter 3, the dashboard introduced a coordinated set of visualizations that balance expressiveness with interpretability. The Flow view combined streamgraphs and chat bubbles to represent participation patterns and turn-taking; stress cues were embedded directly in the chat bubbles rather than shown in a separate view, reducing context-switching and leveraging familiar message-like encodings. The Topic view adopted a radial layout adapted from ConToVi [EAGA+16] to represent participant-topic relationships compactly, though this imposed higher cognitive demand than timeline-based representations. The Summary view aggregated behavioral and thematic metrics to support cross-participant comparison without requiring users to navigate temporal detail. Three coordinated views were chosen over a single composite visualization to allow users to focus on different analytical perspectives while maintaining linkage through the shared timeline.

Evaluation results reflected these trade-offs. The Flow tab's timeline-based design was generally well received, though some participants initially struggled with time-selection interactions and search discoverability. The Topic view imposed higher cognitive load but participants were able to interpret it after brief guidance. The Summary view showed the lowest workload, though guided completion rates (93.3%) were slightly lower than other tabs, possibly due to ambiguity in how time filters affected aggregated metrics.

Overall, the findings show that different visual forms (timeline-based, message-based, radial, and summary-aggregated) each contribute complementary perspectives. When linked through shared temporal coordination and clear interaction patterns, they collectively enable users to navigate between detail and overview while maintaining analytical coherence.

RQ3: How do we maintain analytical coherence across views? A key contribution of this work lies in the coordinated multi-view architecture that binds all data sources to a shared timeline. Brushing, linking, and synchronized media playback created a continuous analytical flow between abstract metrics and concrete evidence, directly addressing the fragmentation observed in systems like Discussion Flows [WKV21] and MeetingVis [SBB+18], which provide rich visualizations but limited cross-modal integration. Participants reported that this integration was useful, valuing the synchronized timeline and playback features and noting that these helped them make connections they would have missed otherwise, though some initially found the time-selection interactions unintuitive.

RQ4: How well does it work? The formative evaluation provided empirical evidence on usability, learnability, and cognitive load. The substantial improvement in task accuracy between unguided and guided phases (76.7% to 96.7%) demonstrates effective learnability, suggesting that brief onboarding could support adoption in contexts where users have limited time for system familiarization. High PSSUQ scores for System Usefulness (6.36/7) and Interface Quality (6.33/7) indicate functional coherence and clarity, while the minimal frustration levels across all task types (below 12 on a 0–100 scale) suggest that cognitive demands remained manageable even when working with complex visualizations.

6.1.2. Contributions to Communication Visualization

This work contributes the field of communication visualization in three specific ways. First, it demonstrates that integrating structural, thematic, and affective dimensions within a unified design approach can support reflective review. Building on ConToVi's topical structure analysis [EAGA+16], the dashboard extends the radial topic view with stress context and embeds it within a coordinated multi-view environment, enabling users to explore relationships between what was said, how stress levels fluctuated, and when critical moments occurred. Participants reported that the integrated views were helpful for connecting stress, topic, and conversational structure, though direct comparison with alternative systems was not conducted.

Second, the work contributes empirical evidence on the cognitive trade-offs inherent in complex multimodal visualization. The finding that Topic tasks imposed notably higher workload than Flow tasks, yet resulted in successful task completion once users received brief guidance, provides actionable insight for future designers. It suggests that interpretive complexity should not be avoided but rather supported through progressive disclosure, contextual help, or adaptive onboarding. This nuances the conventional wisdom that simpler visualizations are always preferable, showing instead that complexity can be productive when it serves clear analytical goals and is accompanied by appropriate scaffolding.

Third, the domain-aware design approach, including integration with data from VR training systems like Sound of the Police [MUA+24] and support for domain-specific labels such as participant roles and event markers, demonstrates a pathway for adapting to professional training contexts. While previous systems have primarily targeted generic meetings or casual chats, this dashboard addresses the analytical needs of high-stress training debriefs, where understanding the interplay between emotional state and communication patterns is relevant for training feedback.

6.1.3. Implications for Practice

The findings offer preliminary implications for how communication analytics tools might be designed and deployed in training environments. The strong learnability effect observed between unguided and guided phases suggests that minimal onboarding (approximately 5 minutes) can substantially improve task accuracy. Designers of similar systems should therefore consider planning for brief system walkthroughs rather than expecting intuitive discovery, particularly for visualizations like the Topic Chart that encode complex relational information.

The cognitive load observed in Topic tasks was the highest among the three tabs, yet participants completed tasks successfully once they received brief guidance. This indicates that participants can handle sophisticated visualizations when they serve clear analytical goals. Rather than simplifying visualizations to minimize cognitive load, designers may focus on ensuring that complexity is justified by analytical value and supported by clear interaction patterns and contextual guidance.

Participants valued the synchronized timeline and playback features, noting that these helped them make connections they would have missed otherwise. The search functionality was also appreciated by those who discovered it, though its low visibility (missed by 40% of participants) underscores the importance of clear affordances. Designers of communication analytics tools may prioritize fluid movement between abstract patterns and concrete evidence, enabling users to navigate from high-level summaries to specific utterances and back again without losing context.

Finally, qualitative feedback revealed that participants frequently commented on component-specific issues, including interpretation difficulties and suggestions for improvement. This highlights the importance of iterative refinement with actual users. Even well-designed systems benefit from observing where users hesitate, what they misinterpret, and which features they find most valuable. These observations suggest that ongoing usability assessment and incremental improvement may be valuable considerations for future deployments rather than treating visualization design as a one-time implementation task.

6.1.4. Bridging Theory and Application

This work sits at the intersection of visualization research, cognitive psychology, and applied training design. It illustrates how theoretical principles from each field can inform the design of practical tools aimed at real-world needs. The dashboard draws on Norman's principles of user-centered design [Nor13] by leveraging familiar metaphors (message bubbles, timeline scrubbing) while introducing novel analytical capabilities. It builds on situation awareness models [End95] by ensuring that users can maintain temporal and contextual orientation while navigating between detail and overview. The evaluation employed NASA-TLX [HS88] to assess whether the resulting designs imposed manageable cognitive demands, with results indicating that workload remained acceptable even for complex visualizations.

The positive evaluation results suggest that this integration was successful, but they also reveal tensions that future work must address. The higher cognitive load for Topic tasks indicates a trade-off between analytical expressiveness and cognitive accessibility. The prototype limitations acknowledged by participants (missing error feedback, initially unintuitive interactions) show that technical polish matters even when conceptual design is sound. The small, non-domain sample highlights the gap between formative usability testing and ecological validation in operational settings.

These tensions are productive rather than problematic. They clarify the design space for communication visualization, showing where additional research is needed, which features require refinement, and how systems might be adapted to different user populations or training contexts. The dashboard developed in this thesis should thus be understood not as a final solution but as a research platform that demonstrates feasibility, identifies key challenges, and establishes a foundation for future innovation.

6.1.5. Addressing Gaps Identified in Related Work

Returning to the synthesis in Section 2.3, this work targets the four major gaps the four major gaps identified across existing communication visualization research. First, the dashboard tackles the absence of emotional and stress-related modeling by integrating continuous stress indicators with conversational structure and content, moving beyond the ambient sentiment feedback of GamIM [PWH14] toward temporally-aligned affective context that supports identification of stress-related communication patterns.

Second, it addresses the lack of cross-modal integration observed in prior work. While ConToVi [EAGA+16] tracks topic engagement over time through animated trails, it lacks integration with affective indicators and provides limited direct links between interacting speakers. Similarly, Discussion Flows [WKV21] isolates participation metrics from conversational content and omits emotional information entirely. This dashboard demonstrates that structure, theme, and affect can be meaningfully unified through coordinated views bound to a shared timeline, enabling users to trace relationships that remain invisible when these dimensions are visualized separately.

Third, the design addresses the neglect of contextual layers noted in Chapter 2. Unlike generic meeting or chat visualizations, the dashboard incorporates domain-specific elements such as participant role labels, event markers, and topic categories relevant to training scenarios. The design demonstrates the feasibility of adapting communication visualization to professional training contexts.

Fourth, the systematic evaluation using standardized instruments (PSSUQ, NASA-TLX), controlled comparison of guided versus unguided interaction, and qualitative feedback through open-ended questions addresses the evaluation gap identified in Table 2.1. While most reviewed works relied on informal feedback or expert case studies, this thesis provides empirical evidence on usability, cognitive load, learnability, and user perceptions that can inform future research in high-stress communication visualization.

Together, these contributions move toward an integrated framework where trainers and trainees can examine how stress manifests in dialogue, addressing limitations identified in current visualization techniques and debriefing practices.

6.1.6. Addressing the Original Motivation

Returning to the motivation presented in Chapter 1, the findings suggest that visualization can enable a more structured and evidence-oriented form of reflection in high-stress training. By linking stress indicators with conversational behavior and temporal context, the dashboard provides a means for examining how stress co-occurs with communication events. Integrating stress data with structural and thematic information creates opportunities to surface stress-related communication patterns that current debriefing practices often identify only retrospectively or inconsistently.

However, the vision articulated in the motivation, where reflective learning becomes a measurable and improvable skill rather than an abstract goal, remains partially realized. The evaluation suggests that participants were able to use the dashboard for structured review tasks and reported that it supported evidence-based discussion of the data, but whether these capabilities translate to improved operational performance remains unknown. The prototype demonstrates feasibility and usability, but ecological validation in actual training pipelines is needed to assess whether insights gained through dashboard analysis lead to actual behavioral change.

This gap between showing that something works and proving it has real impact is common in design research. Tools that perform well in controlled studies may encounter problems when used in practice. This work provides a starting point, but fully realizing its potential will require working with training organizations over time, studying actual learning outcomes, and refining the system based on real-world use.

6.2. Design Recommendations

The evaluation revealed specific usability issues that can be addressed through targeted design refinements. This section presents concrete recommendations organized by dashboard component, drawing on task performance data, NASA-TLX workload ratings, and qualitative feedback from participants.

6.2.1. Flow Tab Refinements

The most critical issue in the Flow tab was search discoverability. With 40% of participants missing the search tool entirely, this feature needs increased visual prominence. Despite being positioned at the top of the chat log with a search icon, participants either overlooked it or defaulted to alternative strategies like browser search (Ctrl+F), manual scrolling, or listening to the recording. Possible improvements include adding placeholder text (e.g., "Search messages..."), increasing the visual contrast of the search bar, or highlighting it briefly when users first enter the Flow tab.

The time brush interaction caused confusion for multiple participants. Four participants struggled with the relationship between selecting a time range and how it updates the view. Participant 530 noted: "It was not too intuitive which part of the dashboard is the main controller, and which parts are actually linked." The core problem is that brushing serves multiple purposes (time filtering and event marking), making it unclear what action the user is performing.

The recommended solution separates these functions into explicit actions. For time filtering, the brush would be replaced with a "Filter by Time" button that opens a selection mode, allowing users to either click start and end points on the timeline or type specific time values directly. For event and period markers, a dedicated section would display existing markers, with an "Add Marker" button that initiates marker creation. Users could then click on the timeline or visualizations to place the marker, or specify a time range manually. This separation makes each action intentional rather than ambiguous, and provides clear entry points for users who might otherwise miss these features.

Component labels and visual separation were insufficient throughout the dashboard. Participant 501 mentioned not understanding the stress line chart initially because the "Stress Level" label was hard to notice. More broadly, observation notes indicated that participants sometimes struggled to distinguish where one component ended and another began. Adding clear labels to all components and enclosing them in visually distinct containers (with backgrounds, borders, or spacing) would improve readability and help users understand the dashboard's structure at a glance. During the session, Participant 534 suggested "separated sections with backgrounds and borders" as an improvement.

6.2.2. Topic Tab Refinements

The Topic tab imposed the highest cognitive load (weighted NASA-TLX of 28.7) and showed the lowest unguided completion rate (70%). While 100% of participants understood it after guidance, the initial learning curve is too steep for unguided use.

The primary issue is that topic contribution data is only visible through hover interactions. Users must hover over each participant or topic segment individually to see values, with no way to compare multiple participants side by side. Participant 507 stated: "Topic graphs, I don't get them to be honest! They are really complicated for me."

The recommended solution adds persistent information panels on the sides of the visualization. These panels would display contribution data for each trainee and overall session totals at all times, updating when users interact with the timeline or filters. This approach makes differences immediately visible without requiring users to hover and mentally retain values for comparison. The donut chart would remain as a visual overview, but the panels provide the detail needed for actual analysis.

The small pie charts within the Topic view were also considered less useful than bar charts. Participant 501 noted: "It's hard to say which topic dominates as in pie charts the comparison is not so exact." Replacing the small pie charts with horizontal bars would improve comparability and align with the Summary tab's design language.

Three participants confused topics with traits, indicating a terminology problem. Clearer labeling and a brief legend explaining the distinction would reduce this confusion. The distinction between "overall" values and "over time" values also confused two participants. Adding explicit labels such as "Session Total" versus "Timeline View" would clarify this.

6.2.3. Summary Tab Refinements

The Summary tab received the most positive feedback and lowest workload (9.5 weighted NASA-TLX), but observation notes identified areas for improvement.

The time filter did not affect the Summary view, contrary to participant expectations. Either the Summary should respond to time range selections (showing metrics for the selected period only), or the interface should clearly indicate that Summary shows session-wide aggregates regardless of time selection.

6.2.4. Cross-Component Issues

Several issues spanned multiple tabs:

Participant names in the stress line chart differed from participant names elsewhere (noted by Participant 532). Consistent naming conventions across all components would reduce confusion.

A bug was reported where filtering by participants caused chat bubbles to change position but stress hue did not update accordingly. This synchronization issue should be fixed.

Hover interactions were sometimes insufficient. Participants suggested adding persistent info boxes instead of hover-only tooltips, showing minute and second timestamps in tooltips, and making participant markers larger on hover.

6.2.5. Onboarding Considerations

The improvement between unguided (76.7%) and guided (96.7%) task completion suggests that minimal onboarding can effectively address most usability challenges. Rather than redesigning complex visualizations to be immediately intuitive (which may sacrifice analytical power), the system could gradually introduce features through first-use tooltips, a brief interactive tutorial, or contextual help icons that explain unfamiliar encodings.

In addition to initial onboarding, persistent help should remain accessible throughout the interface. Adding small information buttons (e.g., an "i" icon) next to each component would allow users to access explanations at any time, not just during their first session. This is particularly useful for the Topic Chart and other complex visualizations where users may need a refresher after returning to the dashboard.

This approach preserves the dashboard's analytical depth while lowering the initial learning barrier. As Participant 507 noted, the difference was "night and day" after receiving brief explanations, and 100% of participants understood all visualizations after guidance.

6.2.6. Summary of Recommendations

Table 6.1 summarizes the key recommendations, prioritized based on how frequently issues were observed and their apparent impact on usability.

Priority	Component	Recommendation
High	Flow	Increase search bar visual contrast and add placeholder text
High	Flow	Replace time brush with explicit "Filter by Time" button
High	Flow	Add separate "Add Marker" button and marker panel
High	Topic	Add persistent side panels showing data without hover
High	Topic	Replace small pie charts with horizontal bars
Medium	All	Add component labels and visual containers for separation
Medium	Topic	Clarify "overall" vs "over time" labels
Medium	Topic	Clarify terminology (topics vs traits)
Medium	Summary	Make time filter behavior explicit
Medium	All	Add first-use onboarding tooltips
Medium	All	Add persistent help buttons (i icons) for each component
Low	All	Consistent participant naming across views
Low	Flow	Fix stress hue synchronization bug when filtering

Table 6.1.: Design recommendations prioritized by issue frequency and impact.

6.3. Limitations

While the study demonstrates the dashboard's promise and usability, it comes with several limitations. First, the evaluation was conducted with a small convenience sample of non-domain participants rather than professional users from police or emergency training contexts. This choice was appropriate for early-stage usability testing but limits how well the findings apply to real-world settings and whether observed behaviors would generalize. Future evaluations should therefore involve domain experts to assess operational relevance and workflow integration.

Second, the stress and topic data used in the prototype were pre-processed and static, rather than derived from real-time data processing pipelines. As a result, the system's ability to handle noisy or incomplete multimodal input was not assessed, though this was outside the scope of the present work. Dynamic integration and uncertainty visualization will be critical for real-world deployment.

Third, while the study included both unguided and guided conditions, it focused primarily on short-term learnability. Long-term retention, collaborative usage, and cross-session comparison were not explored. Finally, some visualization components (particularly the radial Topic Chart) proved cognitively demanding, and terminology confusion (e.g., participants conflating topics with traits) indicated that clearer labeling and contextual help are needed.

These limitations are typical of early-stage visualization research and should be addressed in future iterations of the system.

6.4. Future Work

Building on this foundation, several directions can extend both the functionality and impact of the dashboard.

Adaptive onboarding and progressive guidance could significantly improve the initial user experience. Introducing built-in tutorials and tooltips that adapt to user actions would help new users understand the interface without interrupting analysis. This directly addresses the learnability gap observed between unguided and guided sessions, potentially eliminating the need for structured training sessions before users can engage with the system effectively.

Automated event and pattern detection represents another promising direction. Integrating machine learning models to flag relevant conversational moments (such as sudden stress peaks, topic transitions, or dominant speaking shifts) would reduce manual effort and guide attention during review. This capability could transform the dashboard from a passive visualization tool into an active analytical assistant that proactively surfaces insights.

Expanded emotion and behavior tracking would provide a richer picture of cognitive and emotional states. Moving beyond stress to include features such as frustration, hesitation, or overlapping speech would capture more nuanced aspects of team dynamics and individual performance under pressure. This expansion could reveal subtle indicators of cognitive overload or team coordination breakdown that current stress metrics alone might miss.

Configurable and domain-specific views would enhance the dashboard's applicability across contexts. Allowing trainers and analysts to define custom metrics and layouts tailored to their operational domain (such as tactical coordination for police or triage communication for medical teams) would make the system more flexible and relevant to diverse training scenarios. Doing so might allow aligning the tool more closely with existing training frameworks and evaluation criteria.

Long-term and comparative analysis capabilities would support tracking improvement over time. Enabling comparison of multiple sessions would allow trainers to evaluate training outcomes, identify recurring communication bottlenecks, and measure the effectiveness of specific interventions or training protocols. This temporal perspective is currently missing but critical for understanding whether training actually improves operational performance.

Collaborative and multi-user review features would turn individual reflection into team learning. Implementing shared annotation and synchronized playback for in-person or remote debriefs would allow multiple people to interpret the same evidence together, supporting shared understanding during post-training discussions.

Finally, broader validation and deployment in real-world settings remains essential. Conducting large-scale field studies with domain experts would evaluate ecological validity, scalability, and long-term adoption in actual training pipelines. Such studies would reveal whether the benefits observed in controlled evaluation settings translate to operational contexts, and would identify practical barriers to implementation that can only be discovered through sustained real-world use.

7. Conclusion

This thesis presented an integrated visualization dashboard for analyzing high-stress, multi-party communication in training contexts. The system targets a central gap in existing tools: the separation of structural, thematic, and affective analysis, which makes it difficult for users to understand how stress unfolds in dialogue.

The dashboard brings these dimensions together through coordinated multi-view visualization. Speech activity, topic dynamics, and stress indicators are aligned on a shared timeline, while the Flow, Topic, and Summary views offer complementary perspectives ranging from moment-level inspection to higher-level comparison. Features such as synchronized media playback, transcript search, and annotation tools further support evidence-based reflection, though evaluation revealed that some features require improved discoverability.

The formative evaluation showed that users could learn and operate the system effectively. Task accuracy improved from 76.7% to 96.7% after brief guidance, and PSSUQ ratings indicated strong usability across dimensions (overall satisfaction: 6.60 out of 7). Although the Topic view imposed the highest cognitive demand among the three tabs, frustration remained low, suggesting that the system strikes a workable balance between analytical expressiveness and accessibility.

Beyond evaluating a single prototype, this work contributes to communication visualization research more broadly. It suggests that multimodal integration can support reflective insight, that visual complexity is manageable when paired with minimal onboarding, and that domain-specific design choices can be relevant to professional training contexts. The documented design rationales and rejected alternatives provide additional transparency and transferable guidance for future systems.

Overall, the dashboard represents a step toward more rigorous and evidence-driven post-training reflection. By making relationships between stress and communication behavior visible, it offers a practical path for structured post-training reflection. Realizing the full potential of this approach will require field deployment and longitudinal validation, but the results of this thesis establish a solid foundation and point toward promising opportunities for advancing visualization approaches for high-stakes training contexts.

Bibliography

[BW08] Lee Byron and Martin Wattenberg. Stacked graphs – geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008.

[CM84] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[DFS+13] Sharon Decker, Mary Fey, Stephanie Sideras, Sandra Caballero, Leland Rockstraw, Teri Boese, Ashley E Franklin, Donna Gloe, Lori Lioce, Carol R Sando, Colleen Meakim, and Jimmie C Borum. Standards of best practice: Simulation standard VI: The debriefing process. *Clinical Simulation in Nursing*, 9(6):S26–S29, 2013.

[DSJ99] James E Driskell, Eduardo Salas, and Joan Johnston. Does stress lead to a loss of team perspective? *Group Dynamics: Theory, Research, and Practice*, 3(4):291–302, 1999.

[DV02] Judith Donath and Fernanda B Viégas. The chat circles series: Explorations in designing abstract graphical communication interfaces. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, pages 359–369. ACM, 2002.

[EAGA+16] Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. ConToVi: Multi-party conversation exploration using topic-space views. *Computer Graphics Forum*, 35(3):431–440, 2016.

[End95] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995.

[FG07] Ruth M Fanning and David M Gaba. The role of debriefing in simulation-based learning. *Simulation in Healthcare*, 2(2):115–125, 2007.

[GAW+11] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[HS88] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

[Lew92] James R Lewis. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society Annual Meeting*, volume 36, pages 1259–1260. SAGE Publications, 1992.

[Lew95] James R Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[LRS+18] Moon-Hwan Lee, Yea-Kyung Row, Oosung Son, Uichin Lee, Jaejeung Kim, Jungi Jeong, Seungryoul Maeng, and Tek-Jin Nam. Flower-pop: Facilitating casual group conversations with multiple mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–24, 2018.

[Mar97] Catherine C Marshall. Annotation: From paper books to the digital library. In *Proceedings of the Second ACM International Conference on Digital Libraries*, pages 131–140, New York, NY, 1997. ACM.

[Mar98] Catherine C Marshall. Toward an ecology of hypertext annotation. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, pages 40–49. ACM, 1998.

[MUA+24] Markus Murtinger, Jakob Carl Uhl, Lisa Maria Atzmüller, Georg Regal, and Michael Roither. Sound of the police – virtual reality training for police communication for high-stress operations. *Multimodal Technologies and Interaction*, 8(6):46, 2024.

[Nor13] Don Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, 2013.

[OUR+25] Michael Oppermann, Jakob Carl Uhl, Georg Regal, Markus Murtinger, and Manfred Tscheligi. ROGER: Visualizing voice records to enhance team communication trainings for high-stress situations. In *Proceedings of the 18th International Symposium on Visual Information Communication and Interaction*, VINCI '25, Linz, Austria, 2025. Association for Computing Machinery. Submitted for publication.

[PC05] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. 2005.

[PWH14] Ke-Chen Pong, Chi-An Wang, and Shuo Hsiu Hsu. GamIM: Affecting chatting behavior by visualizing atmosphere of conversation. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pages 2497–2502. ACM, 2014.

[Rob07] Jonathan C Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the Fifth International Conference* *on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, pages 61–71. IEEE, 2007.

[SBB+18] Yang Shi, Chris Bryan, Sridatt Bhamidipati, Ying Zhao, Yaoxue Zhang, and Kwan-Liu Ma. MeetingVis: Visual narratives to assist in recalling meeting context and content. *IEEE Transactions on Visualization and Computer Graphics*, 24(6):1918–1929, 2018.

[Shn03] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*, pages 364–371. Elsevier, 2003. Reprint of Shneiderman (1996).

[SMM12] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.

[SRHW08] Eduardo Salas, Michael A Rosen, Jonathan D Held, and Jeff J Weissmuller. Performance measurement in simulation-based training: A review and best practices. *Simulation in Healthcare*, 3(1):33–41, 2008.

[Sta04] Mark A Staal. Stress, cognition, and human performance: A literature review and conceptual framework. Technical Memorandum NASA/TM-2004-212824, NASA, 2004.

[STKSJ12] Eduardo Salas, Scott I Tannenbaum, Kurt Kraiger, and Kimberly A Smith-Jentsch. The science of training and development in organizations: What matters in practice. *Psychological Science in the Public Interest*, 13(2):74–101, 2012.

[Tuf01] Edward R Tufte. Graphical excellence. *The Visual Display of Quantitative Information*, pages 13–52, 2001.

[WDA+16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.

[WKV21] Tao Wang, Mandy Keck, and Zana Vosough. Discussion flows: An interactive visualization for analyzing engagement in multi-party meetings. In *EuroVis 2021 – Short Papers*, pages 85–89. The Eurographics Association, 2021.

[WO99] Steve Whitaker and Brid O'Conaill. Video interaction for research and education. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, pages 272–273. ACM, 1999.

[YaKSJ07] Ji Soo Yi, Youn ah Kang, John Stasko, and Julie A Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.

A. Appendix

This appendix contains all supplementary materials that support the evaluation and analysis presented in Chapter 5. It includes the complete instruments and task sets used in the user study, providing transparency and reproducibility for future research. Specifically, the appendix is organized as follows:

- Section A.1 lists the full **experimental task sets** used for each dashboard tab (*Flow*, *Topic*, and *Summary*), which participants completed during both the unguided and guided phases.
- Section A.2 presents the **demographic questionnaire** that captured participant background information such as age, role, and experience.
- Section A.3 includes the complete **NASA Task Load Index (NASA–TLX)** instrument used to assess cognitive workload, along with scoring details.
- Section A.4 provides the full **Post-Study System Usability Questionnaire (PSSUQ)** form used to evaluate system usability across three subscales.
- Section A.5 reproduces the **open-ended post-session questions** that collected qualitative feedback about usability, interpretability, and perceived value.

Together, these materials provide the methodological foundation for the evaluation study, ensuring that all aspects of task design, measurement, and data collection are fully documented for replication or extension in future work.

A.1. Experimental Tasks

During the evaluation, participants were asked to complete a sequence of realistic analytical tasks corresponding to each dashboard tab. Each task set was presented through the study interface, which provided brief orientation text and instructions. The goal was to assess initial discoverability, navigation, and comprehension across the *Flow*, *Topic*, and *Summary* tabs. The full text of the task prompts is reproduced below for reference.

A.1.1. Flow Tab Tasks

Instructions: Please take some time to check the Flow tab and then try to answer the following questions. Please refrain from asking for help at this stage.

1. Can you find the timestamp where the conversation involves talking about a knife?

2. Can you identify a moment that seems important and tag it for later reflection?
3. Can you add a time period annotation with some customization?
4. How much is the stress value of the Sergeant at 2:33 (153s)?

A.1.2. Topic Tab Tasks

Instructions: Please take some time to check the Topic tab and then try to answer the following questions. Please refrain from asking for help at this stage.

1. Who is most concerned about safety close to the end of recording at the "Tango discharge 2" event marker?
2. Which topic has the highest share overall?
3. Use the search tool to filter the conversation for the term "Knife." What changes do you observe in the Topic component?
4. Which topic has the highest share after filtering with "Knife"?

A.1.3. Summary Tab Tasks

Instructions: Please take some time to check the Summary tab and then try to answer the following questions. Please refrain from asking for help at this stage.

1. Which participant appeared to be the most stressed overall?
2. Which participant was more concerned about "Safety" overall?
3. Who asked the most questions during this operation, and how many did they ask (raw value)?
4. Who had the most contribution to the "Navigation" topic?

A.2. Demographic Questionnaire

The demographics form captured participant background relevant for interpreting results in high-stress communication analysis. Items included role (e.g., police trainee, trainer, EMT), years of experience, prior exposure to high-stress simulations, and familiarity with analysis dashboards.

The full set of demographic questions was:

1. **How old are you?**
   (Open text response)
2. **What gender do you identify as?**
   (Choose one)

- Male
- Female
- Non-binary / third gender
- Prefer not to say

3. **What is your current occupation or job title?**
   (Open text response)

4. **How many years of experience do you have in this role?**
   (Choose one)

   - Less than 1 year
   - 1-3 years
   - 4-6 years
   - 7-10 years
   - More than 10 years

A.3. NASA Task Load Index (NASA-TLX)

We used the *Raw TLX* (unweighted) [HS88] to measure participants' subjective workload after the interaction phases. The questionnaire was administered on a 0-100 scale, with sliders in 5-point increments for each dimension. The full instrument is shown in Figure A.1.

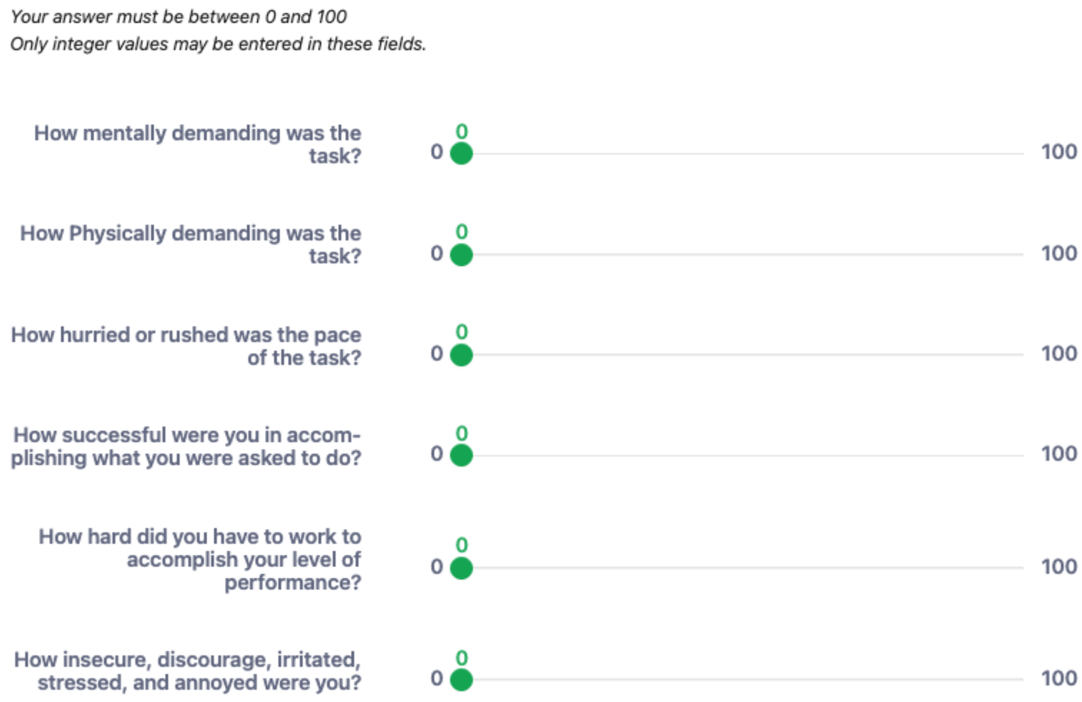


Figure A.1.: NASA-TLX instrument used in the study.

The instrument consisted of six dimensions, each rated independently:

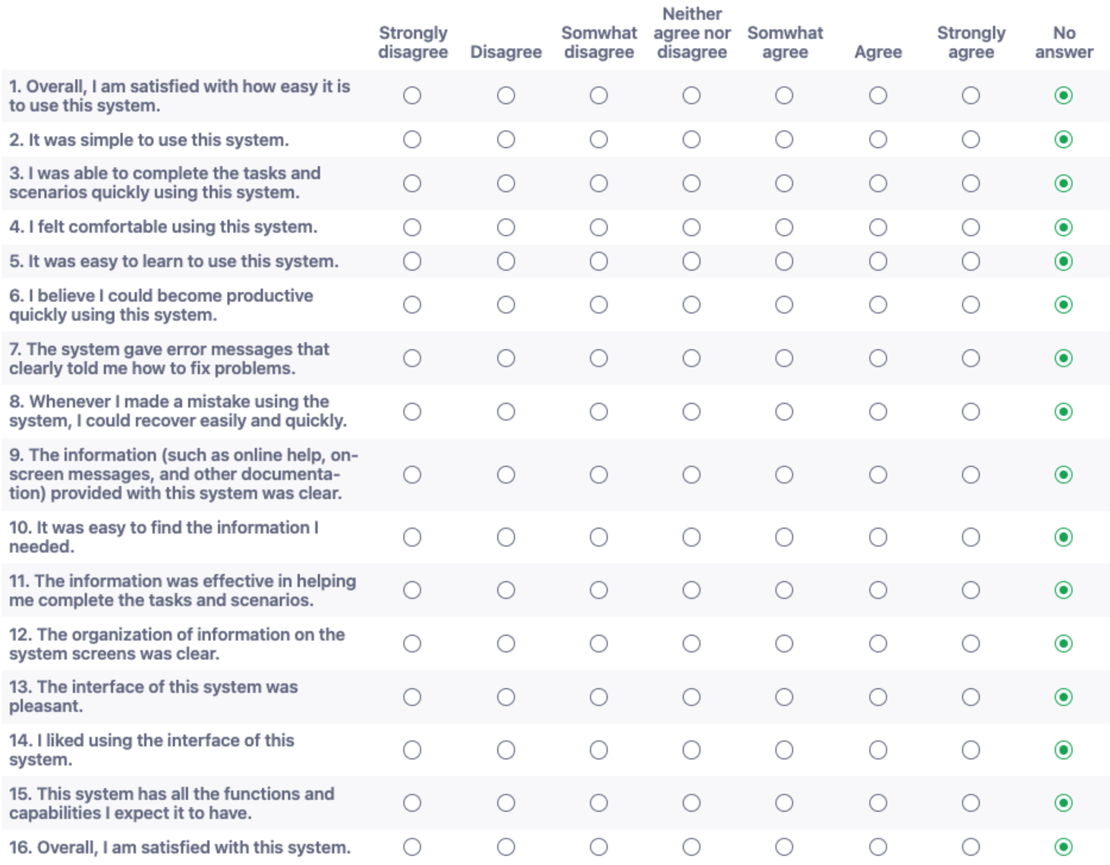
1. **Mental Demand:** How mentally demanding was the task?
2. **Physical Demand:** How physically demanding was the task?
3. **Temporal Demand:** How hurried or rushed was the pace of the task?
4. **Performance:** How successful were you in accomplishing what you were asked to do?
5. **Effort:** How hard did you have to work to accomplish your level of performance?
6. **Frustration:** How insecure, discouraged, irritated, stressed, or annoyed were you?

For the Raw TLX score, we averaged the six ratings to obtain an overall workload value between 0 and 100; no weighting was applied.

A.4. Post-Study System Usability Questionnaire (PSSUQ)

We administered the *Post-Study System Usability Questionnaire* (PSSUQ) after participants completed both the unguided and guided phases. Items were rated on a 7-point

Likert scale where **1** = **Strongly Disagree** and **7** = **Strongly Agree**; lower scores indicated better usability. The full instrument is shown in Figure A.2.



	Strongly disagree	Disagree	Somwhat disagree	Neither agree nor disagree	Somwhat agree	Agree	Strongly agree	No answer
1. Overall, I am satisfied with how easy it is to use this system.	○	○	○	○	○	○	○	◉
2. It was simple to use this system.	○	○	○	○	○	○	○	◉
3. I was able to complete the tasks and scenarios quickly using this system.	○	○	○	○	○	○	○	◉
4. I felt comfortable using this system.	○	○	○	○	○	○	○	◉
5. It was easy to learn to use this system.	○	○	○	○	○	○	○	◉
6. I believe I could become productive quickly using this system.	○	○	○	○	○	○	○	◉
7. The system gave error messages that clearly told me how to fix problems.	○	○	○	○	○	○	○	◉
8. Whenever I made a mistake using the system, I could recover easily and quickly.	○	○	○	○	○	○	○	◉
9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.	○	○	○	○	○	○	○	◉
10. It was easy to find the information I needed.	○	○	○	○	○	○	○	◉
11. The information was effective in helping me complete the tasks and scenarios.	○	○	○	○	○	○	○	◉
12. The organization of information on the system screens was clear.	○	○	○	○	○	○	○	◉
13. The interface of this system was pleasant.	○	○	○	○	○	○	○	◉
14. I liked using the interface of this system.	○	○	○	○	○	○	○	◉
15. This system has all the functions and capabilities I expect it to have.	○	○	○	○	○	○	○	◉
16. Overall, I am satisfied with this system.	○	○	○	○	○	○	○	◉

Figure A.2.: PSSUQ form used in the study.

The standard PSSUQ is a validated usability questionnaire originally developed at IBM [Lew92]. It has appeared in slightly different forms (with 18 or 19 items in early drafts). In this work, we adopt the refined **16-item version** introduced by Lewis (1995) [Lew95], which is now the most widely cited and recommended form. This version organizes items into three subscales:

Scoring. We computed the overall mean across Items 1-16 (lower is better). Subscale scores were calculated as follows:

- **System Usefulness (SU):** Items 1-6
- **Information Quality (IQ):** Items 7-12
- **Interface Quality (IF):** Items 13-15

A.5. Post-Session Open-Ended Questions

At the end of the study, participants were presented with a set of open-ended questions to gather qualitative feedback on usability, comprehension, component value, and perceived usefulness of the dashboard. These were not conducted as interviews but were written responses collected directly in the evaluation questionnaire.

A.5.1. Questions

1. What were your first impressions of the dashboard? (write as much as you want)
2. How did your experience change between the unguided and guided phases? What were your first impressions of the dashboard?
3. Were there any parts of the dashboard that you didn't understand at first?
4. Which components did you find easiest to use? Which ones were more difficult or confusing?
5. Did you feel that the interface layout helped or hindered your understanding of the data?
6. Which components required the most effort to interpret or use?
7. Were there moments where the visualizations helped you discover something you wouldn't have noticed otherwise?
8. Were there any views you rarely used or avoided? Why?
9. Would you use this tool again for analyzing stressful team communication sessions?
10. Any other thoughts, frustrations, or suggestions you'd like to share?