박 사 학 위 논 문
Ph.D. Dissertation

복잡한 정보의 신뢰성, 이질성, 조건성 탐색 지원을 통한 이해 증진

Enhancing Understanding and Engagement with Complex Information through Exploration of Reliability, Heterogeneity, and Contingency

2026

고 은 영 (高 銀 永 Ko, Eun-Young)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

박 사 학 위 논 문

복잡한 정보의 신뢰성, 이질성, 조건성 탐색 지원을 통한 이해 증진

2026

고 은 영

한 국 과 학 기 술 원

전산학부

복잡한 정보의 신뢰성, 이질성, 조건성 탐색 지원을 통한 이해 증진

고 은 영

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2025년 12월 5일

심사위원장	김 주 호	(인)
심 사 위 원	장 정 우	(인)
심 사 위 원	Joseph Seering	(인)
심 사 위 원	Steven Dow	(인)
심 사 위 원	Saiph Savage	(인)

Enhancing Understanding and Engagement with Complex Information through Exploration of Reliability, Heterogeneity, and Contingency

Eun-Young Ko

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Daejeon, Korea
December 5, 2025

Approved by

Juho Kim
Associate Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

DCS 고은영. 복잡한 정보의 신뢰성, 이질성, 조건성 탐색 지원을 통한 이해 증진. 전산학부 . 2026년. 106+vi 쪽. 지도교수: 김주호. (영문 논문)
Eun-Young Ko. Enhancing Understanding and Engagement with Complex Information through Exploration of Reliability, Heterogeneity, and Contingency. School of Computing . 2026. 106+vi pages. Advisor: Juho Kim. (Text in English)

Abstract

Simplification is a common strategy to make complex information easier to understand. However, it can sometimes lead to the loss of important context and meaning, resulting in misunderstanding or superficial interpretation. This loss of nuance can lead to superficial understanding and poor decision-making, particularly in domains where information directly shapes judgment, such as health information, online discourse, and public policy.

This dissertation examines how interactive systems can help users engage with three critical dimensions of information complexity: reliability (understanding how claims are supported by evidence and what remains uncertain), heterogeneity (recognizing how opinions and interpretations differ across individuals), and contingency (grasping how outcomes vary depending on conditions and contexts). Through three systems, ReviewAid, PRISM, and PolicyScope, I demonstrate how interfaces that support disaggregation of simplified information through explorable representations can promote deeper engagement and understanding.

ReviewAid helps readers evaluate health news by surfacing the chain of information from articles back to original research, making evidential uncertainties visible and encouraging critical examination of scientific claims. PRISM addresses the reduction of nuance in online discussions caused by binary reactions, introducing user-generated labels to capture diverse and precise reactions. PolicyScope supports citizens in understanding complex policy effects by presenting structured, stakeholder-based comparisons that help users explore diverse and uncertain potential effects.

Collectively, this dissertation demonstrates that the dimensions of complexity—reliability, heterogeneity, and contingency—can become a medium for engagement when interactive systems make them explorable. Rather than eliminating nuance, these systems reveal hidden structure, support systematic comparison, and help users develop more informed and nuanced understanding.

Keywords Oversimplification, interactive systems, science communication, online discourse, policy communication, human-computer interaction, computer-mediated communication, cognitive engagement

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Communicating information to the public often requires simplifying complex ideas. Scientific findings are condensed into headlines and short news stories, policy proposals are communicated through concise press release and policy briefs. Even in online spaces, systems that collect public reactions often reduce diverse opinions into a small set of simplified signals, such as likes, upvotes, or brief labels. Simplification is a widely used strategy that improves accessibility and understanding of complex information.

However, this simplification carries a significant risk when critical underlying mechanisms, context, or conditions are removed. Oversimplification arises not only from design choices but also from cognitive and media constraints. Communication formats emphasize brevity, clarity, and speed, which can unintentionally hide the structure needed for deeper understanding. Digital platforms further amplify this tendency by prioritizing compact signals and high-level summaries. When this structural information is stripped away, audiences tend to rely on surface cues or pre-existing heuristics [1], increasing the likelihood of misunderstanding and potentially hindering grounded public understanding.

This thesis understands oversimplification as arising most critically from the loss of three key structural dimensions: reliability, heterogeneity, and contingency. These dimensions reflect core components of organized complexity, and losing them can make information appear more certain, uniform, or one-dimensional than it truly is. Reliability addresses how claims are supported by evidence and what uncertainties exist, yet is often reduced to overly definitive messages in science communication. Heterogeneity captures how interpretations and impacts differ across perspectives, but is frequently compressed into binary or polarized framings. Contingency reveals the conditions under which outcomes change, yet this dimension is often omitted in policy communication in ways that make effects appear fixed or universal.

The core contribution of this thesis is demonstrating that interactive systems can act as scaffolds for complexity, restoring reliability, heterogeneity, and contingency without overwhelming the user. We challenge the assumption that complexity must be eliminated for the public to understand an issue. Instead, we propose that understanding is achieved when the inherent structure of complexity is clearly and progressively revealed. Interactive systems offer a uniquely powerful approach to this problem because they can reveal structure without removing content. Unlike static summaries, interactive representations can present information in layers, highlight relationships, and give users control over when and how to explore additional detail [2]. This flexibility allows systems to preserve complexity in an interpretable form, supporting user-paced exploration and richer sensemaking across domains.

Chapter 2 provides the dissertation's conceptual background. Drawing on research in science communication, online discourse, and public policy, it explains how simplification removes key structural elements of information. Organizing prior work around the concepts of reliability, heterogeneity, and contingency, the chapter reviews how oversimplification arises and how interactive systems have begun to address these gaps. Building on this foundation, the dissertation investigates how reliability, heterogeneity, and contingency can be made accessible to users by examining concrete cases of oversimplification in science communication, online discourse, and policy communication.

Chapter 3 discusses communication of reliability in science communication, focusing on how uncertainty and methodological nuance in scientific research are frequently lost in health news. The chapter

presents ReviewAid, a system that supports readers in critically evaluating health articles by surfacing key factors of scientific uncertainty and common patterns of misinterpretation.

Chapter 4 addresses oversimplification of heterogeneity in online discourse, where diverse viewpoints are often collapsed into binary or low-dimensional reaction metrics such as likes or upvotes. This chapter introduces PRISM, an interface that enables richer audience expression through user-generated labels, allowing multifaceted reactions to become visible to others.

Chapter 5 investigates oversimplification of contingency in policy communication, where policy outcomes are often presented as fixed rather than condition-dependent. The chapter presents PolicyScope, a system that helps the public explore complex policy issues through structured, stakeholder-based representations that reveal how effects may vary across contexts and conditions.

Collectively, these three systems demonstrate how interactive approaches can help people engage with information more thoughtfully by restoring three key dimensions that simplification often obscures: reliability in scientific communication, heterogeneity in online discourse, and contingency in policy communication. In all three projects, we applied a user-centered design approach. We interviewed users to understand their needs and challenges, designed systems based on these insights, and validated the effectiveness of the proposed solutions through user studies. Together, these systems illustrate how considerate design that reveals reliability, heterogeneity, and contingency can support more informed engagement with complex information.

Thesis Statement Interactive systems that facilitate exploration of reliability, heterogeneity, and contingency behind simplified information can support users' understanding of and engagement with complex information. These systems deepen engagement by scaffolding structured assessment, enabling nuanced expression of perspectives, and supporting active construction of understanding.

Chapter 6 reflects on the themes that emerge across the three systems and examines how they collectively inform the design of tools for supporting public understanding of complex information. The chapter discusses what the findings suggest about users' engagement, the role of structure in interpretation, and the conditions under which interactive systems can facilitate more thoughtful reasoning.

Chapter 8 summarizes the contributions of the dissertation and outlines directions for future research. It reflects on the potential of interactive systems to mitigate oversimplification and identifies opportunities to extend the design principles and methodological approaches developed in this work.

Contribution

This dissertation makes the following contributions:

1. A conceptual view of information complexity that identifies reliability, heterogeneity, and contingency as core structural dimensions whose loss leads to oversimplification.
2. Design insights on how different types of complexity benefit from different design approaches for revealing structure and supporting deeper engagement.
3. Three interactive systems that instantiate these design insights by addressing distinct forms of simplification.
4. Empirical findings showing that making reliability, heterogeneity, and contingency visible encourages active engagement and leads to clearer, more grounded understanding of complex information.

Chapter 2. Background and Related Work

This chapter provides the conceptual and empirical foundation for the dissertation. It reviews prior literature on the problem of oversimplification, in which the necessary act of simplifying information for accessibility can remove critical structural elements and hinder accurate understanding. To frame this issue, the chapter adopts the three dimensions that guide this dissertation: *Reliability* (how claims are supported and what uncertainties exist), *Heterogeneity* (how interpretations and impacts differ across perspectives), and *Contingency* (the conditions under which outcomes may differ).

In this chapter, we establish the conceptual context for how oversimplification emerges in science communication, online discourse, and public policy, and examine how prior research has addressed these challenges. Then, we discuss prior interactive systems that have attempted to restore the dimensions of *Reliability*, *Heterogeneity*, and *Contingency* and clarify the remaining gaps that this dissertation addresses.

The chapter is organized as follows. Section 2.1 introduces the conceptual problem of oversimplification, outlining the trade-offs inherent in simplifying complex information. Section 2.2 discusses the three structural dimensions in detail, reviewing why **Reliability**, **Heterogeneity**, and **Contingency** matter and how prior systems have attempted to recover them. Section 2.3 synthesizes this literature, highlighting the limitations of existing approaches and articulating the contribution and necessity of designing interactive systems that more systematically mitigate oversimplification.

2.1 Communicating Complex Information to the Public

Effective public communication requires the simplification of complex ideas. In fact, simplification is inevitable because humans have fundamental limits in how much information they can process. Bounded rationality [3] explains that people have limited working memory and cannot fully handle the volume and complexity of real-world information. Dual process theory [4] explains how fast, efficient System 1 processing often takes over when information is complex or ambiguous. In this sense, simplification in communication becomes a practical way to reduce cognitive load and keep information manageable.

Although this is essential for accessibility, it comes with a risk of oversimplification, where underlying mechanisms, assumptions, or perspective differences are lost. For instance, scientific findings are reduced to definitive headlines [5, 6, 7], controversial discussions are framed as two opposing sides [8, 9, 10], and policy proposals are summarized as short lists of pros and cons [11, 12]. These simplifications have substantive consequences. When important contextual details are removed, people tend to rely on surface cues or pre-existing heuristics [1, 4, 13], increasing the likelihood of misunderstanding [14]. Scientific news without caveats can fuel misperceptions of consensus and overstate certainty [15, 16]. In online platforms, binary or aggregated reactions can "flatten'' opinions, obscuring subtle disagreements and contributing to perceived polarization [9, 8]. Similarly, when policy explanations omit the conditions under which a policy operates, they overlook how its effects vary across stakeholder groups [11, 17], making outcomes appear more uniform and deterministic than they actually are [18].

However, complexity itself is not incompatible with public understanding. Research in information design and sensemaking shows that understanding depends less on the amount of information than on how clearly its structure is revealed [2]. People are capable of working with complex material when they

can see how pieces relate to one another, what the main components are, and how to move through the space of information. Studies in visual analytics demonstrate that interactive representations can help users form coherent mental models of large or heterogeneous information spaces [19]. Techniques such as progressive disclosure [20], hierarchical organization, and interactive visualization support this by presenting information in layers, allowing users to understand the overall structure before examining specific details. Prior work also shows that making relationships, dependencies, or flows explicit can reduce perceived complexity and support more accurate reasoning [21, 22, 23], even when the underlying content remains intricate. Recent work further suggests that communicating complexity can enhance intellectual humility and improve perceptions of epistemic trustworthiness, underscoring the value of preserving structure rather than removing it [24]. In this sense, the goal is not to hide or compress complexity but to scaffold it by providing cues, structure, and navigational guidance that make the information interpretable without flattening it.

Building on this, the next section discuss three key structural dimensions whose loss is central to the problem of oversimplification: **reliability**, **heterogeneity**, and **contingency**.

2.2 Structural Dimensions of Information Complexity

Prior work suggests that the challenge of complexity is not solely due to the volume of information. A critical, yet often overlooked, factor is the lack of visible structure [4, 25]. The challenge for public communication lies in making these underlying structures perceivable and navigable. This leads to a core question: which aspects of structure are most important for public understanding? In many communication settings, certain forms of structure consistently disappear when information is simplified. Identifying these recurring points of loss helps clarify what must be made visible for complexity to remain interpretable.

In this thesis, I propose that information complexity in public communication can be understood through three structural dimensions: depth, breadth, and contingency. These dimensions represent distinct aspects of informational organization that are frequently lost through simplification. While elements of depth, breadth, and contingency have been discussed separately across different domains, I use these three dimensions to characterize the forms of oversimplification that appear throughout the systems presented in this thesis.

2.2.1 Reliability: How Claims Are Supported

I define *reliability* as how claims are supported by evidence and what uncertainties exist. It reveals the connection between a claim and its underlying evidence: what data support it, what assumptions bridge evidence to interpretation, and what limitations or caveats remain. Reliability addresses the structural understanding of how information is grounded rather than simply presenting the claim itself. Reliability differs from accuracy: accuracy concerns whether a claim is correct or incorrect, while reliability concerns how the claim is supported. It focuses on the evidential structure and what uncertainties remain, regardless of whether the claim ultimately proves true or false.

Public communication often presents conclusions without showing the reasoning, evidence, or limitations behind them. In science communication, this occurs when variability and caveats are omitted and findings are reduced to definitive headlines [26]. Similar patterns appear in policy communication, where expected outcomes are described without explaining the mechanisms that produce them. Research

shows that such shallow presentations can overstate certainty and hide the tentative nature of knowledge [5, 27]. When people encounter only polished conclusions, they develop unrealistic expectations and struggle to evaluate competing arguments or interpret new evidence [7]. Moreover, when promised outcomes fail to materialize under unanticipated conditions, the initial claims lose credibility, paradoxically undermining the trust that simplification was meant to preserve [16]. This loss of reliability can leave audiences confident in their understanding while lacking the underlying rationale, an illusion of explanatory depth [28].

2.2.2 Heterogeneity: How Interpretations and Impacts Differ

heterogeneity as how interpretations and impacts differ across perspectives. It reveals which stakeholders hold distinct positions, what dimensions they prioritize, and how their interpretations diverge. Heterogeneity asks "*who* sees this differently and *how*?" to expose the plurality that single framings obscure. While the term "diversity" is often used in similar contexts, I use heterogeneity to emphasize that perspectives are not the same, rather than focusing on comprehensiveness or balance across viewpoints. The focus is on the fact of difference itself.

Heterogeneity can be lost across various forms of communication. News coverage presents topics through a single narrative frame, or at best offers dichotomous perspectives that flatten the actual spectrum of viewpoints. Online platforms compress this further through reaction mechanisms that reduce complex opinions to likes or emoji. Comment sections allow expression but rarely structure it in ways that make the range of interpretations perceivable.

This narrow framing has consequences across domains. In science communication, presenting phenomena through a single disciplinary lens obscures how economists, sociologists, and psychologists interpret the same patterns differently, making competing explanations appear invalid rather than complementary. In policy evaluation, reliance on aggregate metrics can hide how the same intervention creates different outcomes across groups. In political discourse, research shows that exposure to diverse perspectives can reduce polarization [29], yet platforms often fail to surface this diversity, leading to skewed perceptions [30]. Without access to multiple perspectives, people may dismiss valid alternatives as uninformed rather than recognizing them as grounded in different priorities or interpretive frames.

2.2.3 Contingency: Conditional Variability

I define *contingency* as how outcomes and effects vary across conditions and circumstances. Unlike heterogeneity, which captures different *interpretations* of the same information, contingency captures how actual *results* differ across contexts. While reliability addresses how claims are supported by evidence (*how* is this grounded?"), contingency maps the boundary conditions within which effects occur (*when and where* does this happen?"). It reveals that causal effects are not universal but conditional—the same intervention can produce different outcomes depending on population characteristics, implementation details, or environmental factors. Contingency differs from uncertainty: uncertainty reflects incomplete knowledge about what will happen, while contingency reflects the fact that different outcomes occur under different conditions, even when those outcomes are known.

Public communication often presents outcomes as universal when they are actually conditional. Policy proposals are described as uniformly beneficial or harmful, ignoring how impacts vary across demographics, regions, or implementation contexts. Educational advice is offered as one-size-fits-all, despite variations in learning contexts and student backgrounds. Product recommendations are given

without consideration of individual circumstances. Scientific findings are reported as general truths without specifying the populations studied or conditions under which effects were observed [31].

This loss of contingency has significant consequences. In policy communication, effects are inherently heterogeneous—the same intervention produces different outcomes for different populations under different conditions [11]. Yet debates often proceed as if outcomes were uniform, leading to shallow disputes over whether a policy works rather than nuanced discussions of when, for whom, and under what circumstances it works. Without contingency, people may assume deterministic outcomes, overlook important variability, and struggle to assess whether general claims apply to their specific situations. When promised benefits fail to materialize in particular contexts, this can erode trust and fuel skepticism toward evidence-based recommendations.

2.2.4 An Integrated View of Information Complexity

While reliability, heterogeneity, and contingency have been studied separately across different domains, they share a common characteristic: they represent structural properties of information that become invisible through oversimplification. Moreover, these three dimensions are often interconnected. Understanding why different stakeholders hold distinct positions (heterogeneity) often requires examining the underlying evidence and assumptions they prioritize (reliability). Recognizing that outcomes vary by context (contingency) reveals why different groups experience the same policy differently (heterogeneity). A claim's degree of certainty (reliability) may depend on how well it generalizes across conditions (contingency).

Rather than hiding this complexity through simplification, the goal is to make its structure visible and navigable. When reliability, heterogeneity, and contingency are preserved, users can identify what matters, understand why outcomes differ, and see how various perspectives connect. This forms the foundation of this dissertation: interactive systems can help people meaningfully engage with complex information by revealing and structuring these dimensions rather than flattening them.

The three systems presented in subsequent chapters each address different combinations of these dimensions: **ReviewAid** emphasizes reliability by revealing how claims are supported and what uncertainties exist in scientific claims; **PRISM** emphasizes heterogeneity by surfacing how interpretations differ in public reactions; and **PolicyScope** addresses both heterogeneity and contingency by enabling exploration of how policy effects vary across stakeholders and conditions.

2.3 Interactive Systems for Deep Understanding and Sensemaking

As discussed in the previous section, the challenge of complexity is not solely due to the volume of information. A critical, often overlooked factor is the lack of visible structure [4, 25]. A large body of HCI, information visualization, and social computing research has explored how interactive systems can help users see this structure and reason about information more effectively. In this section, I review prior work that aligns with these goals, organized into three themes that correspond to the structural dimensions introduced earlier. Domain-specific prior work for each project will be discussed in the related work section in corresponding chapters.

2.3.1 Reliability: Revealing Evidential Structure and Uncertainty

Many systems focus on helping users understand the context, variability, and uncertainty behind a claim. These systems aim to restore the reliability that is often lost in simplified messages.

Several projects use annotations, visual cues, or short explanations to show where a fact comes from and how certain it is [32]. Much of this work focuses on how systems can present incomplete or uncertain information in ways that help people form calibrated judgments, rather than avoid or mistrust it [33, 34]. Building on this idea, other systems look beyond presentation and introduce guided prompts that help users reflect on assumptions, limitations, or missing context instead of accepting information at face value.

Recent HCI research further emphasizes that guided prompts can foster both the breadth and depth of users' reflections, enabling them to explore diverse perspectives and elaborate on the reasoning behind their judgments [35]. Such frameworks highlight that supporting uncertainty in interactive systems is not only about revealing quantitative uncertainty but also about structuring opportunities for users to reflect on the qualitative dimensions of their understanding [36]. By prompting active engagement with uncertainty, these systems help users move beyond passive information consumption and toward deeper, more nuanced reasoning.

A core design strategy is to reveal structure gradually. Following Shneiderman's "overview first, zoom and filter, then details-on-demand" principle [20], many interfaces provide a simple starting point and allow users to explore deeper layers only when needed. This reduces information overload while still preserving important context. However, most systems stop at presenting uncertainty or provenance. They expose structure but do relatively little to prompt users to reason with that structure.

Recent work suggests the importance of going further. Studies on uncertainty visualization in AI systems show that making uncertainty visible not only improves calibrated trust but also encourages users to reflect on context, limitations, and assumptions rather than accepting outputs at face value [37]. Similarly, reflection-focused HCI research highlights that guided prompts, structured questioning, and other interactive techniques can help users engage more actively with complex information and form more grounded interpretations [38]. Together, this work points to the need for systems that reveal structure and also support users in working with that structure and motivates the approach taken in ReviewAid, which aims to help users actively examine the uncertainty underlying scientific claims, rather than merely displaying it.

2.3.2 Heterogeneity: Supporting Pluralistic Expression and Viewpoint Exploration

Heterogeneity is often lost when complex viewpoints are compressed into a small number of categories or a single aggregate signal. Many systems, particularly in social and civic contexts, summarize diverse opinions into simplified metrics that hide important distinctions [39]. Prior work has examined how interface design can better capture this diversity of perspectives and make it visible to others. This is essential because public understanding depends not only on the availability of information but also on the ability to see how different people make sense of the same issue.

A first area of research looks at how to collect a wider range of opinions. People often hold nuanced views, but systems provide only limited ways to express them. Binary or numeric signals collapse many types of responses into a single metric, obscuring important differences. To counter this, researchers have introduced more expressive reaction mechanisms, such as multi-dimensional labels, fine-grained sentiment

markers, or short textual statements, that allow users to articulate more specific viewpoints [40, 41]. More recent work explores how generative AI systems can support iterative refinement and personalization of user viewpoints through interactive dialogue [42]. These designs aim to reflect the diversity of opinions rather than flatten them into a uniform score.

A related stream of work focuses on how to present diverse viewpoints once they have been collected. Techniques such as clustering, structured summaries, and argument maps help organize large sets of perspectives so that users can see patterns without losing nuance [43, 44]. Systems such as StarryThoughts visualize the landscape of public opinions through semantic clustering and interactive exploration, helping users navigate a wide range of viewpoints without collapsing them into a single consensus [45].

Recent work also examines how AI-driven systems can facilitate exposure to contrasting perspectives: HearHere visualizes the political stance of news articles and comments to help users access viewpoints outside their echo chambers [46], while other systems use LLM-generated multi-persona debates to expose users to contextually grounded alternative viewpoints and reduce confirmation bias [47]. Such systems illustrate how presenting viewpoints in ways that make contrasts and trade-offs visible supports a broader understanding of public opinion [29], though recent HCI scholarship emphasizes that interface designs must also attend to positionality and ensure that marginalized perspectives remain visible [48, 49].

Deliberation tools offer yet another approach. Systems such as ConsiderIt separate pros and cons to make diverse viewpoints easier to compare [44], while argument visualization tools map claims and counterarguments to support structured disagreement [43]. Although these systems primarily target deliberative contexts, they illustrate the broader value of interface designs that help users see how perspectives differ and where tensions lie.

Taken together, this line of work shows how systems can preserve, organize, and surface diverse viewpoints, an aim that aligns with PRISM's focus on capturing nuanced reactions before they are flattened through aggregation.

2.3.3 Contingency: Facilitating Conditional Reasoning and Exploratory Analysis

Contingency is lost when outcomes are presented as fixed or uniform, even though they often depend on specific conditions, contexts, or interactions. Prior work has explored how interactive systems can help users understand how outcomes change when underlying factors vary, supporting more flexible and conditional reasoning.

A major line of work focuses on exploratory visualization and scenario-based tools. Systems in policy informatics and data journalism allow users to manipulate variables and observe how outcomes shift in response [50]. These tools support "what-if" exploration, helping users see that effects are conditional rather than deterministic. By allowing direct manipulation of inputs, they make variability visible and give users a way to reason through multiple scenarios instead of relying on a single authoritative output.

Research in policy informatics further highlights the challenge of communicating heterogeneous effects: the idea that a single intervention can affect different stakeholders or regions in different ways [11]. These effects reflect both heterogeneity (different groups experience different impacts) and contingency (those impacts vary depending on conditions). These contingent factors often reflect concrete differences in geography, demography, or public infrastructure. While some systems incorporate scenario planning to illustrate such differences, many do not provide a clear design structure for systematically comparing

how stakeholder groups are affected under varying conditions. As a result, non-experts often struggle to understand why outcomes diverge or who may be affected differently.

The importance of conditional variation extends beyond policy domains. HCI research on context-aware systems and situated interaction emphasizes that outcomes depend on specific circumstances, whether involving user context, environmental factors, or temporal constraints [51]. Work in visualization and decision support similarly shows that people form more accurate judgments when they can explore how results vary across scenarios or assumptions. Across these threads, a consistent insight emerges: systems that hide variability tend to promote overly deterministic interpretations, while systems that reveal conditional structure support more grounded reasoning.

Despite these advances, few interfaces provide non-experts with a structured way to examine conditional effects in a multi-stakeholder setting. Helping people see who is affected, how, and under what circumstances remains an open design challenge, one that motivates the approach taken in PolicyScope.

Chapter 3. ReviewAid: A Scaffolded Approach to Supporting Readers' Evaluation of Health News

This chapter presents ReviewAid, an interactive system that helps readers critically evaluate health news by surfacing the reliability of scientific claims. This chapter focuses on reliability, by inviting health news readers to evaluate how claims are supported by evidence and what uncertainties exist in that support. Through two studies, I demonstrate how interfaces that make evidential structure visible can encourage readers to adopt a more critical stance toward health news, moving beyond passive acceptance of simplified claims toward active examination of underlying research. This chapter is partly based on a paper published at ISLS 2022 [52]. All uses of "we", "our", and "us" in this chapter refer to coauthors of the aforementioned paper.

3.1 Introduction

"Sugary Drinks are Linked to Cancer Onset"[53], *"Breast-Feeding May Cut Breast Cancer Recurrence Risk"*[54], and *"Even Moderate Air Pollution May Lead to Lung Disease"*[55]. These are titles of news stories that deliver health-related research to the public. As health news affects readers' everyday decisions or behaviors[56], it is important for readers to get an accurate and comprehensive understanding of research findings and their application. However, health news stories often present findings as simplified assertions that hide the information readers need to judge their reliability [57, 58].

Scientists and experts in science journalism are trying to improve public communication of science. Some science journalists are aiming to improve the production pathway (e.g.,[59, 60]), while watchdog journalists (e.g.,[61, 62, 63]) try to help health news readers by providing evaluations of news stories. However, with limited expert resources, such approaches face challenges in terms of scalability and sustainability [1].

We tackle this issue by supporting readers to evaluate health news themselves by breaking down simplified claims into their underlying components—the quality of research, the strength of evidence, and the appropriateness of interpretations. Previous research in crowdsourcing and civic engagement has successfully engaged the general public in complex tasks commonly done exclusively done by experts (e.g., policymaking[64], budgeting[65], and science[66]) by lowering barriers to participation. In this project, we design a system that supports readers to collaboratively evaluate and review health news stories using structured criteria that make reliability assessment more accessible.

Assessing health news stories is a complex task for readers without expert knowledge as it necessitates the use of both scientific and media literacy. One needs to gauge the quality of scientific evidence and, at the same time, evaluate how properly the information is delivered and interpreted. Therefore, using explicit evaluation criteria, which tackle both media and scientific literacy, is widely recommended[67, 68]. Evaluation criteria scaffold readers' assessment of health news and, at the same time, can serve as a basis for their evaluation. We apply this idea and design a system that supports non-experts' review of health news stories using evaluation criteria.

We conducted a formative study to better understand the challenges that readers face while evaluating health news stories using evaluation criteria. Based on the formative study results, we identify

[1] HealthNewsReview.org stopped their periodical publication at the end of 2018.

design guidelines for a system supporting non-experts' review of health news: 1) provide evaluation criteria adjusted for non-experts, 2) provide explicit and detailed criteria, 3) guide readers to disentangle media and research aspects, 4) help readers apply each criterion with the context of a specific news story, and 5) utilize the differences in individual preferences to reduce readers' burden when reviewing.

As one instantiation of the identified design guidelines, we present ReviewAid, a system that allows readers to collaboratively review health news stories with evaluation criteria designed for non-experts. Readers can review health news stories with criteria of their choice, and the reviews are shared as a collective evaluation. ReviewAid scaffolds the evaluation by prompting readers to evaluate 1) if the media deliver enough information (media coverage), 2) if the research is conducted well (validity of the research), and 3) if the media provide enough interpretation or explanation of the research. To support users in understanding and applying each criterion with more relevance to the specific story being reviewed, ReviewAid presents example questions for each evaluation step.

We evaluate the effect of ReviewAid in two studies. In the first study, we measure the effect of the scaffolded evaluation process in an individual setting where no collaboration exists. We conducted a between-subjects study (n=66) where each participant evaluated a health news story and wrote a review with high-level criteria (Baseline), high-level criteria with detailed subcriteria (SubCriteria), or SubCriteria with the scaffolded evaluation process (ReviewAid). Results show that participants using ReviewAid wrote a quality review, held a more critical view, and reported a increased level of self-efficacy.

In the second study, we explored a more practical use of ReviewAid in a collaborative setting. We ran a five-day-long study with 24 participants to understand readers' behaviors and experiences. We used a between-subjects (SubCriteria and ReviewAid) design to measure the effect of the scaffolding framework under the presence of collaboration. Our observations illustrate participants' considerations when doing distributed work, and the opportunities and costs that arise in a collaborative setting. Also, the result shows that participants using ReviewAid wrote more high-quality reviews and reported a increased level of self-efficacy.

The contributions of this project are as follows:

- Formative study results that reveal design guidelines for systems to support non-expert readers' evaluation of health news.
- ReviewAid: a web interface that supports online readers' collaborative review of health news stories with evaluation criteria designed for non-experts and a scaffolded evaluation process.
- Results from two studies that showed the effectiveness of the scaffolded evaluation process in both individual and collaborative settings.

3.2 Background and Related Work

We first provide background on challenges in science communication and summarize existing efforts to improve the public communication of science. Then we review previous research on involving non-experts in quality evaluation of information and scaffolding complex tasks.

3.2.1 Challenges in public communication of scientific research

The innate complexity is one of the most important challenges in public communication of scientific research [69, 70]. Details of scientific research should be summarized and interpreted so that the public

can engage, read, and understand [58] under the risk of omitting important information or misinterpreting the finding. Likewise, the conditional nature of scientific research limits the scientists' knowledge and control about a subject matter, providing limited validity to some studies [71, 72, 73, 74, 75]. However, communicating the limitation of research is a highly demanding task for journalists as it requires a huge amount of time and effort and sometimes costs readers' engagement and trust[76, 77].

When in comes to health-related topics, public communication gets more challenging. As health-related research often involves human or living subjects, where researchers have less control on the condition, it tends to have a higher scientific uncertainty[78, 16, 27]. However, facing various constraints such as time to write a quality story, expert knowledge, or even reader's attention [58], media often omit important information (e.g., not reporting the study was an animal study), misinterpret the findings (e.g., interpreting relative risk as absolute risk), and provide inaccurate implications for the research (e.g., generalizing findings on old adults to the general adult population)[79].

3.2.2 Efforts to improve the public communication of science

Science journalists aim to improve the quality of published media by supporting the production pathway[59, 60] or forming initiatives on a better journalistic practice[80, 81]. Smith et al.[58] proposed ways to support the process with techniques in social computing. Despite these efforts, however, there exist health news stories[82, 79] and scientific publications[83] that deliver distorted and exaggerated interpretations, or inaccurate information.

Some groups of experts are taking a reactive approach by assessing health news stories and share their evaluation with the public. In HealthNewsReview.org[61], journalists and medical practitioners rate and review news stories on medical interventions or health-related research. SciCheck of FactCheck.org[62] provides fact-checking of misleading scientific claims. Science Feedback[63] provides credibility scores and comments on news stories on climate change and health. However, with limited expert resources, these approaches are neither scalable nor sustainable.

A more scalable approach is assisting readers, or non-experts, to evaluate science news stories. SciLens[84] developed an AI-powered quality indicator of the context of news stories (e.g., the number of visitors on the news website, length of the story, or number of quotes) and showed that the developed indicator helps non-experts evaluate scientific news stories. In this project, we present a tool that supports non-expert readers to evaluate the content of science news stories – how good or bad the provided information is and why.

3.2.3 Involving non-experts in assessment of information quality

Involving non-experts in assessing the quality of information is one of the most widely used approaches[85, 86], together with methods that use algorithmic indicator[87, 88, 89] or experts[62, 90]. However, the method for gathering non-experts' assessment should be designed carefully to guarantee the reliability of the evaluation[91, 92, 93].

One important decision that needs to be made is the scope of evaluation. In many fact-checking processes[92, 94], crowds provide primary inputs on what to fact-check and simple ratings on it while experts make the final decisions or reviews. On the other hand, Zhang et al.[95] proposed credibility indicators that can be shared by players (experts, crowd, or algorithms) and developed quality indicators that can be annotated without domain expertise.

The design of the evaluation process is also important. Previous research has introduced algorithmic

supports for non-experts' evaluation of information quality[84, 96]. Providing detailed and explicit metrics is also an effective strategy[95, 93]. However, research showed that evaluation of the content, not context, is hardly supported by algorithmic tools, and non-experts need training or other technical support for content evaluation[95, 93]. In this work, we design evaluation criteria to assess the content of health news stories that are adjusted to non-experts and provide a scaffolded evaluation process.

3.2.4 Scaffolding complex tasks for non-experts online

Scaffolding is an instructional technique that guides students to accomplish complex tasks by breaking them into several subtasks. We summarize below previous work that introduced the scaffolding technique for general non-experts in an online environment. Using scaffolding has been a common technique to support crowdworkers in providing high-quality design critique [97, 98, 99, 100, 101, 102]. Critiki[97] scaffolded the critique by prompting set of questions in order while CrowdCrit[98] used pre-authored design principles that crowdworker can refer. Likewise, Voyant[101] identified the types of design feedback that are expected from novices and scaffolded the feedback process with them.

Another line of research applied scaffolding to help the crowd in doing more objective and scientific work. Docent[66] introduced the Learn-Train-Ask framework to help crowdworkers in collaborative generation of scientific questions. Wang et al.[103] developed the CrowdSCIM workflow, which customized a system originally designed for students' analysis of historical sources to a crowdsourcing setting with no instructor. Reviewing health news stories is challenging as the issues of scientific research and miscommunication are mixed. In this work, we present a scaffolded evaluation process that guides readers to disentangle these issues and better evaluate each.

3.3 Formative Study

We conducted a formative study to better understand how people use existing evaluation criteria in reviewing health news stories. We sought to 1) understand how existing evaluation criteria benefit online readers and 2) discover challenges that online readers face when following evaluation criteria to review health news stories.

3.3.1 Existing evaluation criteria for health news stories

Scholars and journalists have developed several evaluation criteria for science news stories. Criteria-based evaluation can be seen as an analytic evaluation process where assessment is done for each criterion. Previous work has shown that analytic evaluation leads to more reliable evaluation (e.g., less affected by first impressions and generates constructive feedback, than holistic evaluation where the evaluation is done on overall quality[104, 105, 106]. Below we briefly introduce two kinds of evaluation criteria (one from science education and one from journalism) and discuss their implications.

Evaluation criteria used in science education

Researchers and practitioners in science education have developed sets of evaluation criteria to teach students how to conduct a scientific inquiry when reading media reports on scientific findings. They aim to teach student what to consider in evaluating a scientific claim and what can threaten the validity of the claim. They often come as a list of detailed questions that students should answer while reading the story[68, 67].

Strengths of providing explicit evaluation criteria (see the left column of Table 3.1), compared to teaching the concept of scientific inquiry steps in an abstract form [107, 108, 68, 109] (see the right column of Table 3.1 for examples), is that it is easier for students to engage with the news article and interrogate it. These serve as checklists that students can use to assess the quality of news stories based on how many questions are answered satisfactorily[67].

Existing evaluation criteria are designed to be used in the classroom setting, where instructors guide students and correct them as they conduct evaluation. In online news reading environments, however, instructors are not available or expected to guide users. Therefore, additional support is needed to understand readers' potential misuse of criteria and provide preventive or reactive measures.

Table 3.1: Example prompts in each type of guideline. Evaluation criteria (left) and inquiry steps (right)

Type	Evaluation criteria	Inquiry steps
Examples	(From Always Ask[67]) - What were the subjects of the study? - How was the experiment carried out? - How certain are the scientists about their conclusions?	(From Elements of Science Critical Reading[68]) - Identify the main idea of the text. - Identify the writer's purpose. - Identify data and evidence given in the text.

Existing evaluation criteria are designed to be used in the classroom setting that instructors guide students and correct them if students conduct an inappropriate evaluation. In the online news reading environment, however, the presence of such instructor is not expected. Therefore, understanding potential misuse and providing preventive or reactive measures are necessary.

Evaluation criteria used in journalism

Compared to those designed for students, criteria developed in journalism put more emphasis on how the study findings and their implications are interpreted. For example, the evaluation criteria used in HealthNewsReview[61] include questions on the cost and harms of the medical intervention (if applicable), how strong the study findings are, or how reliable and novel the finding is. Some journalists (e.g., [110, 111]) or organizations (e.g., [112, 113]) also provide a similar list of questions.

Evaluation criteria provided by journalists or practitioners guide readers to ask how news stories deliver scientific research, interpret it, and discuss it in a broader manner. However, as miscommunication in health news stories occurs in a sophisticated manner, it is difficult to assess how well the media delivered the research without knowing the scientific methodology and understanding of the study being delivered. This remains a question as to how well the non-expert reader can use these criteria to evaluate health news stories.

3.3.2 Participants and Procedure

We recruited eight participants by making a call for participation in an online community at a technical university in Korea. Participation was limited to those who are fluent in English. Five of them were undergraduates and three were graduate students. Six participants said that they had no prior experience in research. Each participant received KRW 10,000 (approx. $8) for participating in an hour-long study session.

We asked participants to evaluate one health news story by using two sets of criteria – 'Always Ask'[67] (AA) and 'HealthNewsReview'[114] (HNR) shown in Table 3.2 We used both kinds of evaluation criteria to get a comprehensive understanding of non-experts' challenges in conducting an evaluation. AA is evaluation criteria designed for students and comes with detailed and explicit subcriteria for each high-level criterion. HNR is used by experts to review health news stories and has more focus on how the story frames the research and discusses its implication to the audience. Although HNR is used by experts, it is of a similar level of difficulty to other guidelines developed by journalists for the general public.

Table 3.2: Evaluation criteria used in the formative study

Source	**High-level Criteria**
Always Ask (Jarman et al. [67])	1. Context of the study (Who did research? Who funded? Where reported?) 2. How research was conducted (Subjects, sample size, procedure, time period) 3. Basis for conclusion (Data collected, evidence justifying conclusions, certainty) 4. What other scientists think (References to other studies, support from others) 5. Context of media report (Who wrote, outlet interests, campaign associations) 6. Importance of study (Implications, applications, how to respond)
HealthNewsReview ([114])	1. Adequately discuss costs of intervention 2. Adequately quantify benefits of intervention 3. Adequately explain/quantify harms of intervention 4. Grasp the quality of evidence 5. Avoid disease-mongering 6. Use independent sources and identify conflicts of interest 7. Compare new approach with existing alternatives 8. Establish availability of treatment/test/product/procedure 9. Establish true novelty of approach 10. Avoid relying solely on news release

Participants first read a news story and then were asked to evaluate the story and provide the rationale by using the two evaluation criteria in order. The order of AA and HNR was counter-balanced. We asked them to think-aloud during the evaluation process. After the session, we conducted a short interview asking about their overall experience and difficulties during the evaluation process. All sessions were conducted individually.

3.3.3 Result

Benefits

Learning the evaluation criteria Most participants said that the evaluation criteria helped them learn what to consider. P1 and P3 noted that they could realize that asking those questions is very important only after seeing them. Participants also said that the criteria contain points of view that they had not thought about, while different criteria were new to different participants. P7 said that the criteria made him care more about the numbers, which he rarely did when reading health news stories. On the other hand, P1 said he cares about the numbers but had not thought about the social context around the research.

Critical evaluation of news story Participants said they could hold a more critical view of the

news story by assessing it with evaluation criteria. P4 said that he first thought the story delivers quality information but realized it does not as it could not answer many questions in the criteria. P7 said that the criteria prevented him from judging the news story based on his gut feelings.

Challenges

Despite their positive comments on having evaluation, participants said conducting the actual evaluation with the criteria is difficult. They reported challenges across various steps, from understanding each criterion to applying the criteria to evaluate the news story.

Lack of prior knowledge or external information Participants said they could not evaluate some criteria because they require prior knowledge or external information. Knowledge in the topic or even knowledge of previous research in the domain was needed to answer questions that ask about the value of delivered research (e.g., *"Does this story establish the true novelty of the approach?"* in HNR or *"What is the importance of this study?"* in AA). P2 said that *"Some criteria were meaningless to me as there was nothing that I can do with my level of knowledge."* P4 said, *"It seems that I need to conduct an extensive investigation to really judge the novelty of this research, and I don't think I can correctly answer the question even after the investigation."*

Unable to conduct a grounded evaluation with broad criteria Participants often failed to provide a rationale for their evaluation. This happened a lot when participants were evaluating the news story with HNR, which does not give explicit subcriteria for each criterion. P8 said, *"I think I conceptually understood what each (HNR) criterion is asking for, but this does not mean I know what to consider. I can give a score based on my impression, but I cannot explain why I gave that score."* Most participants said they are more confident with their evaluation given to AA than HNR as the subcriteria helped them understand concrete questions that they can ask and develop their thoughts. P3 noted that *"... evaluating the story with the broad criteria feels much easier as I don't need to think that much, but I have no confidence with my score on them."*

Unable to apply the criteria to a specific news article Participants found it difficult to apply the criteria, even the detailed ones, to the specific news article they were reading. Some participants said it is unclear how each criterion could be evaluated for the news article, which occurred frequently to criteria regarding the research itself. P7 noted that *"I can see that some information ...[concerning] the experiment [is missing] but cannot think of what it is specifically."* Some participants said that they were not confident with their evaluation for criteria that require subjective inference, such as "Does the story seem to grasp the quality of the evidence?" (HNR) or "Does the evidence appear to justify the conclusions?" (AA). P8 said, "It was hard to determine what is enough and what is not. It'd be great to have some examples of what can be considered to be sufficient or problematic."

Confusing the quality of news story with the quality of scientific research covered As discussed in Section 2, the research, as well as how it is delivered, affects the quality of health news story. There were some difficulties in evaluating the health news story raised by this innate complexity when participants were asked how reliable the research is in the absence of information. P1 and P4 explicitly mentioned that it is tricky as those criteria ask them to evaluate the research when there is not enough information provided. Some participants actually mistook the lack of information for lack of validity in the research itself. For example, three participants (P2, P3, and P8) blamed the research assuming that it did not control potential confounding variables or used unreliable measures to collect the data.

Effort required for exhaustive evaluation Despite the benefits of having evaluation criteria, participants said that conducting an exhaustive evaluation is demanding. Participants spent 5-10 minutes

to evaluate the news story for each of AA and HNR. P1 noted that *"...If I am given these evaluation criteria while I'm reading health news, I would like to evaluate the story for a couple of criteria but not all of them."* However, we observed that each participant has different preferences over criteria. For example, P1 liked the criterion "Is there information about what other scientists think?" (AA4) the most, while it was the least preferred criterion for P4.

3.4 Design Guidelines for Systems Supporting Non-expert Readers' Evaluation of Health News

Providing evaluation criteria helps online health news readers get a critical perspective on the information delivered. However, our formative study result shows that challenges arise in evaluation due to the difficulty of understanding the criteria and the complexity of science communication. Also, we learned that readers find it demanding to evaluate news stories for a set of criteria. To address such challenges, we have identified design guidelines for systems supporting online readers' evaluation of health news.

G1. Provide evaluation criteria adjusted to non-experts.

Criteria that ask about the value (e.g., novelty or implication) of the delivered research require external information or expert knowledge. Participants in our formative study said that they feel helpless when they are asked to evaluate such criteria. This does not mean, however, those criteria should be excluded. Having such criteria is still valuable as they suggest and teach readers what to consider. Rather, the benefits of such criteria can be maintained without discouraging the readers by adjusting the scope and target of the evaluation. For example, rather than asking readers to evaluate the novelty of the research, readers can still check whether the news story has explained or discussed how novel the research is.

G2. Provide detailed and explicit evaluation criteria.

Findings from our formative study revealed that having broad or implicit evaluation criteria can lead readers to perform a cursory evaluation, resulting in unwanted outcomes such as reinforcing the initial impression of the news story. Participants also said that detailed criteria could serve as a basis for their judgment and increased confidence in their evaluation. By providing detailed and explicit evaluation criteria, a system can help readers get a clearer understanding of the criteria and conduct a grounded evaluation with higher confidence.

G3. Guide readers to disentangle the complexities in science communication.

Health news stories are a result of the sequential effort of multiple players (scientists, PR department, or science journalists) in the production pathway. Issues in each production stage can be added and entangled together in a health news story. It is important for readers to clearly disentangle the source of the problem because the consequences of evaluation may vary significantly depending on the source. For example, suppose a patient reads a low-quality health news story on a potentially effective intervention. If the patient concludes that the research is invalid, the consequence can be just neglecting it. If the patient concludes that the media did a poor job of delivering the research, the consequence will be finding another health news story or consulting with his/her doctor. Our formative study showed readers can easily confuse the source of issues between the research and the medium. Therefore, a system should help readers disentangle and distinguish the issues raised from the research and the medium.

G4. Provide easier ways to apply evaluation criteria with the context of an individual news story.

Applying evaluation criteria to a specific health news story is challenging even when the reader clearly understands the questions being asked. In our formative study, participants said they could point out unmet criteria, but they could not explicitly say what should be included or improved in the news story. Therefore, a system can assist readers by supporting the actual application, or instantiation, of evaluation criteria with the context of individual health news stories.

3.5 ReviewAid: Scaffolding Readers' Collaborative Review of Health News

As one instantiation of the identified design guidelines, we designed a system called ReviewAid that supports online readers' collaborative review of health news. In ReviewAid, readers can collaboratively review a health news story by following the evaluation criteria designed for non-experts. ReviewAid provides subcriteria for each criterion and guides readers to evaluate each subcriterion following a scaffolded evaluation process. Below we explain 1) evaluation criteria designed for non-expert readers and 2) how ReviewAid scaffolds evaluation of each criterion.

3.5.1 Evaluation criteria designed for non-expert readers (G1, G2)

ReviewAid uses evaluation criteria that we developed by restructuring three existing criteria, namely AA, HNR, and a taxonomy developed in Korpan et al.[115]. Criteria were designed so that they support 1) scientific inquiry on the content of research and 2) how the research finding is connected to the other research and the real world. Table 3.3 shows the criteria designed for ReviewAid. Five of them (1-5) are on the content of research, and the remaining two (6-7) are on its certainty (connection to other research) and application (connection to the real world).

Criteria 1-5 are introduced to scaffold the scientific inquiry process explicitly. Unlike questions such as 'Does the story seem to grasp the quality of the evidence?' or 'How was the study conducted?', these criteria are organized around the specific components of scientific research. Criterion 6 brings the concept of scientific uncertainty[71, 72, 73] and lets readers question the certainty of research findings. Lastly, criterion 7 asks questions on how the research finding is applied to the real world and why it is important.

To help readers better understand each criterion and base their evaluation on, ReviewAid provides two to four subcriteria for each criterion. Readers evaluate the article for each subcriterion and then assess the high-level criterion based on it. Subcriteria give concrete questions to be asked to assess the story for each criterion. Lists of subcriteria for each criterion were reconstructed based on the taxonomy developed in Korpan et al.[115]. The list of criteria and subcriteria used in ReviewAid are presented in Table 3.3.

We designed evaluation criteria for ReviewAid by following the design guidelines G1 and G2. Note that there can be many different sets of criteria following G1 and G2, and the scaffolded evaluation process (which will be described below) can be used for any evaluation criteria regardless of the specification.

3.5.2 Scaffolded evaluation process (G3, G4)

Disentangling media and research aspects (G3)

When a reader selects a subcriterion to evaluate, ReviewAid guides readers to evaluate the news article on 1) how well the story provides information (coverage), 2) how reliable the research is (reliability),

Table 3.3: Evaluation criteria and subcriteria used in ReviewAid

#	High-level Criterion	Subcriteria
1	**Participants/Subject:** Who did the research study? Who were the participants?	- What/who were the subjects? - How many subjects did they study? - How did they select/recruit the subjects?
2	**Research Design:** How and why was this research designed and how was it done?	- What question was the study designed to answer? - When and how long did they conduct the study? - What factors were controlled?
3	**Measure:** How were the factor (putative cause) and effect defined and measured?	- How did they define and deliver/measure the factor? - How did they define and measure the effect?
4	**Data & Statistics:** How were the raw data and statistical results?	- How does the data look like? - How significant is the effect? - How large is the effect?
5	**Social Context:** Is there any social factor that may have influenced the research?	- Who conducted the research? - Who funded the research? - Who promoted the research?
6	**Theory & Related Research:** Does the finding align with other research?	- Is there any theory that can explain the result? - Does this finding align with previous research? - What do other researchers think about this?
7	**Application & Implication:** What is the implication and how should I relate this to the real world?	- Will the same effect hold in general? - How can/should I get or change the factor? - What are the side-effects of the factor? - How important is this finding to our society?

and 3) how well the story comments on the reliability of the research (explanation). The system prompts the evaluations of these aspects in order of coverage, reliability, and explanation. For the criterion on the contextual information (6 and 7), ReviewAid asks for the explanation aspects only. Figure 3.1-A shows the evaluation steps ordered by the aspects.

Also, to prevent potential confusion between the media coverage and research validity, the step for research reliability was skipped if the reader answers there is no information on the subcriterion. This was to prevent potential confusion, such as mistaking the lack of information for a limitation in reliability. Likewise, the step for an explanation was skipped when the reader thought the research was reliable, and no additional comment or explanation is needed.

Evaluation results for each aspect, for each subcriterion, are summarized in color as in Figure 3.1-C. Readers write a review for the high-level criteria while seeing the summary result. Readers can refer to their evaluation of each subcriterion by clicking on it.

Bridging each criterion and the news story with examples (G4)

For each aspect of evaluating coverage, reliability, and explanation, ReviewAid presents example questions or comments related to each aspect. These examples are from questions or comments left on other health news stories. Phrases that are specific to the news stories that each question or comment is raised on were indicated separately (in purple color) as shown in Figure 3.1-B. This is designed to give readers a better sense of how they can evaluate the subcriteria and aspects by seeing how those subcriteria are contextualized with individual news articles.

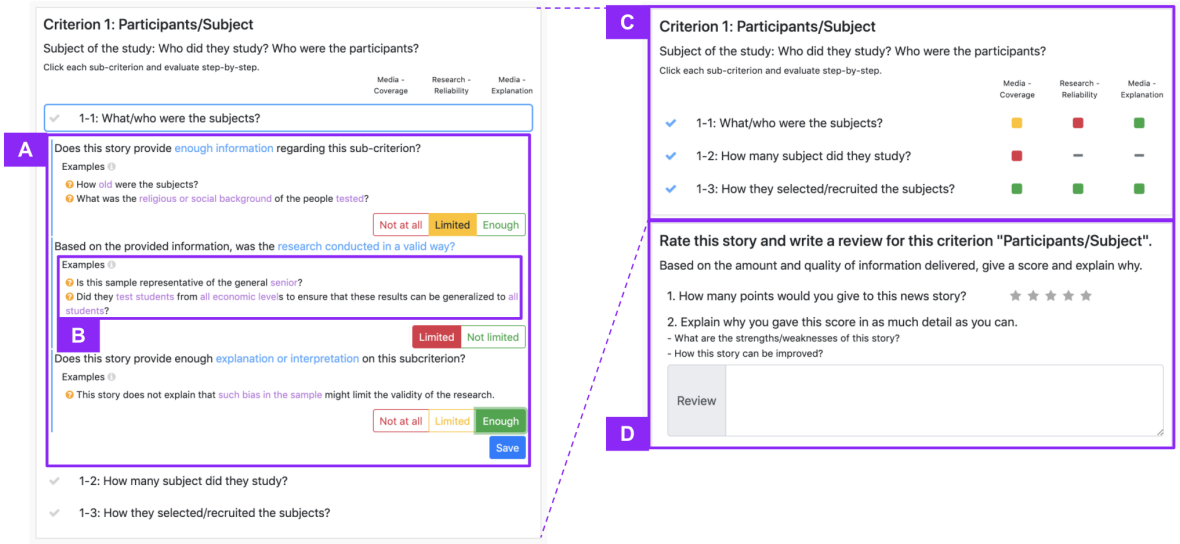


Figure 3.1: Overview of scaffolded evaluation process in ReviewAid. (A) For each subcriterion, the user evaluates the story for each of media coverage, research validity, and media interpretation aspects. (B) Example comments or questions raised in other news stories are shown. Terms that are specific to the source news story are colored in purple. (C) Results of the scaffolded evaluation are summarized and (D) the user gives a score and writes a review based on this.

The examples used in the ReviewAid are constructed by collecting questions from multiple sources. First of all, we used questions from Korpan et al.[115], and HealthNewsReview.org [61]. Also, we collected questions by running a survey on Amazon Mechanical Turk that asked workers to generate five or more questions on health news stories. One researcher categorized questions into subcriteria and the three aspects (coverage, reliability, and explanation) and prepared 1-3 examples for each subcriterion and aspects pair.

Summarizing evaluation on each aspect

We added one step to the scaffolded evaluation process that asks the user to evaluate the story by aspects - media coverage, research validity, and media interpretation. This summary evaluation for each aspect is shown to other readers as in three colored flags (shown in Figure 3.7-(D)). At this stage, users can internally adjust weights to each subcriterion based on its importance and summarize the assessment for each aspect of the high-level criterion.

3.6 Study 1: Evaluation of ReviewAid in an Individual Setting

We conducted two studies to assess the effect of ReviewAid. In Study 1, we measure the effect of ReviewAid in an individual setting: each participant reviewed a health news story with all seven criteria by oneself. We explore the individual setting to measure the effect of using the scaffolded evaluation process in a controlled setting. In Study 2, we explore a collaborative use of ReviewAid: each participant can choose news stories and criteria to review while reviews are shared as a collective evaluation. Study 2 is designed to demonstrate the feasibility of ReviewAid in a more realistic an practical setting. We

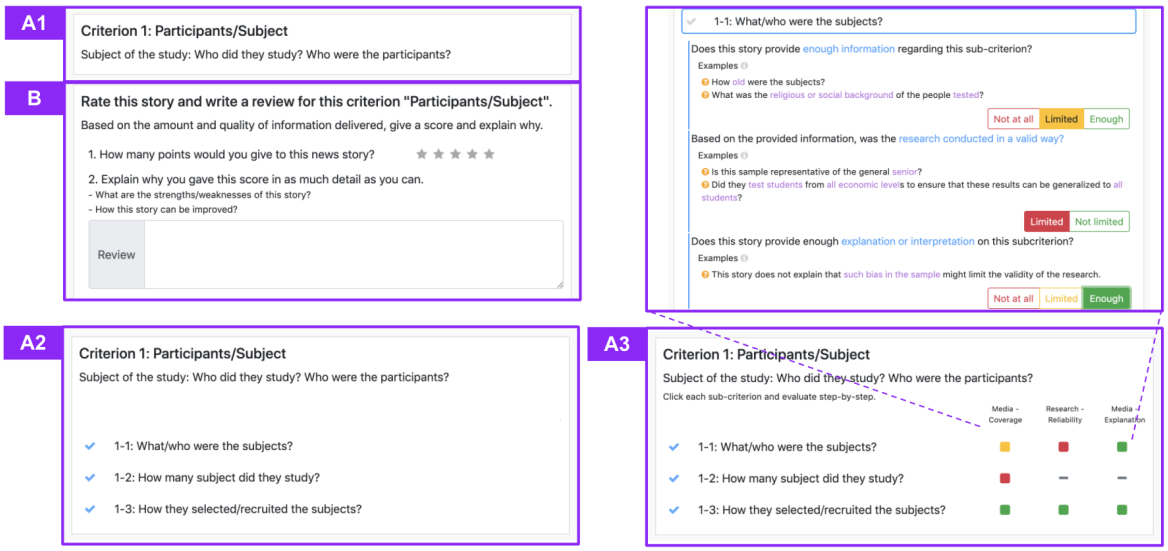


Figure 3.2: Reviewing interfaces for Baseline, SubCriteria, and ReviewAid conditions. (A1) Baseline: High-level criterion is shown. (A2) SubCriteria: High-level criterion and subcriteria are shown. (A3) ReviewAid: High-level crietrion and subcriteria are presented. Participants are guided to conduct a scaffolded evaluation for each subcriteria. (B) Participants in all three conditions are asked to give score and write a review for each criterion.

present the design and result of Study 1 in this section and Study 2 in the next section (Section 7).

In Study 1, we compared the Baseline (high-level criteria only, as usual guidelines), SubCriteria (subcriteria for each high-level criterion provided), and ReviewAid (SubCriteria with the scaffolded evaluation process) conditions. We hypothesized that readers will write higher quality reviews and perceive a higher level of self-efficacy with ReviewAid. Specifically,

H1. Participants provide more rationales for their evaluation with ReviewAid.

H2. Participants provide a more concrete and specific rationale with ReviewAid.

H3. Participants write a more appropriate review with ReviewAid.

H4. Participants hold a higher level of self-efficacy in ReviewAid.

H5. Participants hold a more critical view of the story with ReviewAid.

3.6.1 Method

Study Design and Conditions

We conducted a between-subjects study with three conditions. In the Baseline condition, only the high-level criteria were given, and in the SubCriteria condition, 2-4 subcriteria for each high-level criterion were provided. In the ReviewAid condition, participants were asked to follow the scaffolded evaluation process before giving a score to the story and writing a review text. In all three conditions, participants were given seven high-level criteria and asked to rate the health news story and write a review text for each criterion. Figure 3.2 shows interface for each condition.

We added the SubCriteria condition, in addition to the Baseline and ReviewAid condition, in order to separately understand the effect of providing detailed subcriteria and the effect of providing a scaffolded evaluation process framework.

Participants

We recruited 66 participants from Amazon Mechanical Turk (Approval Rate greater than 95%, located in the US). We randomly assigned the participants to one of the three conditions and we had 22, 21, 23 participants for the Baseline, SubCriteria, and ReviewAid conditions, respectively. For their 40-60 minutes-long participation, participants were paid $8. The average age was 35.9 (SD: 9.00) with min. 24 and max. 65. Five participants said they received post-graduate training (two in the Baseline and three in the ReviewAid condition). Participants' educational background was as follows: High school graduate (11), Technical, trade, or vocational school (4), Some college with no 4-year degree (19), College graduate (27), and post-graduate training (5).

Tasks and Procedure

In the main task, each participant read a health news story and reviewed the story with seven criteria. For each criterion, participants scored the story (1-5) and wrote a review (minimum 20 characters) including rationale, pros/cons, and suggestions for improvement (Figure 3.1-(D)). We used "Tofu might harm memory in elderly" from The Telegraph[116], previously used in[117]. The story reported a study on soy consumption and memory among elderly Indonesians, though no causal linkage was identified.

Participants answered a pre-survey at the beginning of the experiment. Pre-survey measured education level, research experience, perception of health news, and need for cognition (REI-10[118]). We also measured participants' prior belief on the subject matter (how much they agree with the claim that 'Tofu is good for one's memory in general') and self-efficacy in reviewing (confidence in judging quality, pointing out inadequacies, and suggesting improvements).

After the main task, they answered a post-survey. Post-survey re-measured belief and self-efficacy, assessed workload (NASA-TLX), and asked participants to describe how criteria, subcriteria (SubCriteria and ReviewAid), and scaffolding (ReviewAid only) affected their experience.

Measuring the quality of reviews

We conducted discourse analysis to measure review quality: number of rationales (H1), concreteness and specificity (H2), and appropriateness (H3). We analyzed 132 review texts (66 each for Criterion 1: participants/subject and Criterion 6: theory/related research), covering both content and context of research.

The analysis involved three steps conducted by one author and an external coder (Ph.D. student in biology), blind to study conditions.

The analysis was done in three steps, and one of the authors and one external coder (Ph.D. student in biology) worked together. The study condition of each review text was hidden during the analysis.

In the first step, we split each review text into multiple arguments so that each argument contains a single intention or meaning. To ensure consistency between coders, the two coders first split 10% (7 for each criterion) of review texts together and then coded the next 10% of review texts independently, compared the result, and discussed to reach an agreement. After building the consistency, the remaining

review texts were split in a distributed manner; each coder split half of the remaining review texts. As a result, a total of 132 review texts are divided into 598 arguments.

In the second step, they coded the arguments by their type: general evaluation, rationale, summary/repeat, and others (Table 3.4). As in the first step, 20% of the data were used in consistency building (10% for calibration, 10% for consistency building). The remaining arguments were coded by both coders independently, and the inter-coder agreement was 0.80 (Cohen's κ), which indicates a high level of inter-rater reliability. The two coders finalized the result by discussion. There were 75 (12.5%), 443 (72.4%), 58 (9.7%), and 32 (5.4%) arguments in general, rationale, repeat, and other types, respectively.

Table 3.4: Description for each category and example arguments

Type	**Description**	**Example arguments**
General	Argument indicating stance with simple ground	"I give this 3 stars simply because the study design was just okay" (Crit. 2 – Study Design)
Rationale	Argument with elaborated grounds	"The story should have specified the exact amount beyond which it becomes harmful."
Summary/repeat	Argument that repeats the content of the story with no indication of stance	"There were 719 subjects in the study."
Others	Argument of other kinds, e.g., personal experience or feeling	"I feel in today's day and age, the health craze is all over."

In the last step, we coded concreteness, specificity, and appropriateness (all in binary) of each rationale-type argument. Table 3.5 shows example arguments. Three labels are marked simultaneously for each argument. As in the previous steps, the two coders had a consistency building session with 20% of data and then labeled the remaining data independently. The inter-coder agreements (Cohen's κ) were 0.83, 0.86, and 0.83 for concreteness, specificity, appropriateness.

3.6.2 Result

Overall, participants showed a moderate to high level of confidence in their media literacy (Mean: 5.28, SD: 1.24) and ability (Mean: 4.9, SD: 1.31)to evaluate health news stories. There was no between-group differences in self-efficacy and media-literacy, as well as in perception of health news and need for cognition.

The average of time spent on review was 20.8 (SD: 13.2), 17.3 (SD: 8.1), and 21.0 (SD: 10.7) minutes for the Baseline, SubCriteria, and ReviewAid conditions, respectively. This does not include time spent on the pre/post-survey and instruction for the task.

The average word counts of review text written by each participant (for all seven criteria) was 421.7 (SD: 242.0), 386.1 (SD: 131.3), and 425.5 (SD: 189.6) for the Baseline, SubCriteria, and ReviewAid conditions, respectively.

The average of the number of arguments made in a review text was 4.11 (Median: 4, SD: 2.14), 4.23 (Median: 4, SD: 1.8), and 5.19 (Median: 5, SD: 2.65) for Baseline, SubCriteria, and ReviewAid conditions. Table 3.6 shows average number of each type of arguments per review text in each condition.

Table 3.5: Description for each label and example arguments

Label	Example arguments
Concreteness	(Concrete) "I think more information about how much tofu was consumed to be considered "high consumption" is necessary." (Abstract) "It does not provide enough information to make a reasonable conclusion about the results of the study."
Specificity	(Specific) "The link between the phytoestrogens in tofu and memory is briefly mentioned in the story, but it is not fully explained." (Non-specific) "No theory to explain these results is offered."
Appropriateness	(Appropriate) "There is no information on if the researchers controlled for these variables or not." (Inappropriate, confusing media and research) "The researchers did not control other factors." (Inappropriate, misaligned with criterion) "They studied a huge number of subjects." (for Criterion 3 - Measures)

Table 3.6: Average number of each type of arguments per review text in each condition

	General	Rationale	Summary/Repeat	Other	Total
Baseline	0.70 (Med: 1, SD:0.82)	2.61 (Med: 2, SD:1.85)	0.46 (Med: 0, SD:0.85)	0.34 (Med: 0, SD:0.68)	4.11 (Med: 4, SD:2.14)
SubCriteria	0.38 (Med: 0, SD:0.62)	3 (Med: 2, SD:1.05)	0.60 (Med: 0, SD:1.01)	0.26 (Med: 0, SD:0.59)	4.23 (Med: 3, SD:1.8)
ReviewAid	0.61 (Med: 0.5, SD:0.71)	4.17 (Med: 4, SD:2.58)	0.28 (Med: 0, SD:0.50)	0.13 (Med: 0, SD:0.54)	5.19 (Med: 5, SD:2.65)

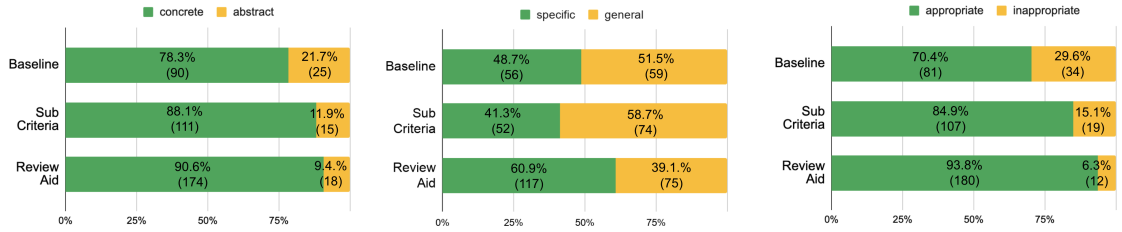
H1. Participants provide more rationales for their evaluation with ReviewAid.

The average number of rationales provided in a review text was 2.61 (Median: 2, SD: 1.85), 3.00 (Median: 3, SD: 1.95), and 4.17 (Median: 4, SD: 2.58) for Baseline, SubCriteria, and ReviewAid condition. The difference between groups was significant (Kruskal-Wallis Test, $p< 0.01$, H=9.42) and the post-hoc test showed that differences between Baseline and ReviewAid, and SubCriteria and ReviewAid were significant (Dunn-Test, $p<0.01$ for Baseline and ReviewAid and $p<0.05$ for SubCriteria and ReviewAid). The difference between Baseline and SubCriteria was not statistically significant.

H2. Participants provide a more concrete and specific rationale with ReviewAid.

Figure 3.3 shows the proportion of concrete (a) and specific (b) rationales in each condition.

Concreteness The average number of concrete rationales provided in a review text was were 2.05 (Median: 2, SD: 1.68), 2.64 (Median: 3, SD: 1.82), and 3.78 (Median: 4, SD: 2.21) for Baseline, SubCriteria, and ReviewAid condition. To exclude the effect of providing more (or less) rationales, we compared the concreteness of each rationale between conditions. Rationales provided in the ReviewAid condition tend to be more concrete than in the SubCriteria and Baseline. The ratios of concrete rationale were 0.78, 0.88, and 0.90 for Baseline, SubCriteria, and ReviewAid conditions, respectively. The pairwise differences between groups were significant for Baseline-Subcriteria and Baseline-ReviewAid (χ^2 Test, $p<0.05$ with χ^2=4.20 and $p<0.05$ with χ^2=9.12, respectively). The difference between SubCriteria and ReviewAid was not significant. "I think that it added to my ability to see piece by piece what was



(a) Concreteness (b) Specificity (c) Appropriateness

Figure 3.3: Ratio of concrete (a), specific (b), and appropriate (c) rationales in each condition

affecting my opinion and better articulate my thoughts in my reviews. It helped me notice what was missing."

Specificity The average number of specific rationales provided in a review text was 1.28 (Median: 1, SD: 1.28), 1.24 (Median: 1, SD: 1.32), and 2.54 (Median: 2, SD: 2.61) for Baseline, SubCriteria, and ReviewAid condition. Rationales were more specific in the ReviewAid than in SubCriteria and Baseline conditions. The ratios of specific rationale were 0.49, 0.41, and 0.61 for Baseline, SubCriteria, and ReviewAid conditions, respectively. The SubCriteria conditions showed the lowest level of specificity. The pairwise differences between Baseline and ReviewAid, and SubCriteria and ReviewAid conditions were statistically significant (χ^2 Test, $p<0.005$ with $\chi^2=4.38$ and $p<0.001$ with $\chi^2=11.82$) while the difference between Baseline and SubCriteria was not. One participant in the ReviewAid condition said "I liked seeing these type of examples. Helped me fully understand what was being asked and how it applied to this specific article. I feel like it kept my reviews very focused."

H3. Participants write a more appropriate review with ReviewAid.

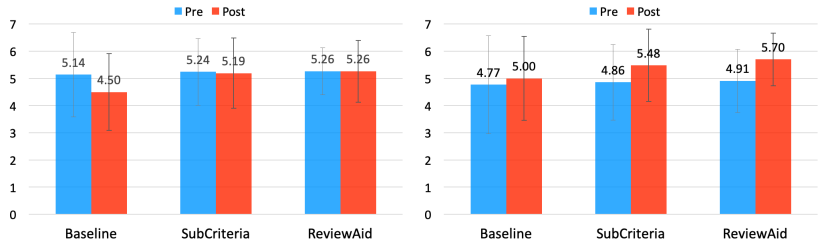
Out of 121 review texts with rationales, 39 (32.2%) contained inappropriate arguments. The number or review texts with inappropriate rationales were 17 (out of 38, 44.74%), 13 (out of 38, 34.0%), and 9 (out of 45, 20.0%) for Baseline, SubCriteria, and ReviewAid conditions respectively.

The average number of inappropriate rationales in a review text was 0.89, 0.50, and 0.27 for Baseline, SubCriteria, and ReviewAid conditions. The difference between groups was significant (Kruskal-Wallis Test, $p<0.05$, $H=7.09$) and the post-hoc test showed that differences between Baseline and ReviewAid to be significant (Dunn-Test, $p<0.05$). The difference between Baseline and SubCriteria, and SubCriteria and ReviewAid were not statistically significant.

The ratios of inappropriate rationale to all rationales were 0.30, 0.15, and 0.06 for Baseline, SubCriteria, and ReviewAid conditions respectively. The pairwise differences between groups were all significant (χ^2 Test, $p<0.05$ with $\chi^2=7.35$, $p<0.05$ with $\chi^2=6.74$, and $p<0.0001$ with $\chi^2=30.69$ for Baseline-SubCriteria, SubCriteria-ReviewAid, Baseline-ReviewAid pairs respectively).

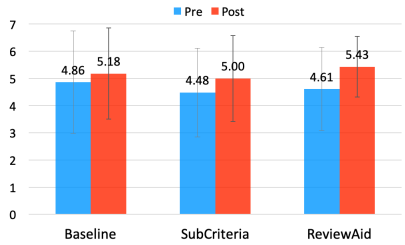
H4. Participants hold a higher level of self-efficacy in ReviewAid.

We measured the participants' self-efficacy in the reviewing task by asking three 7-point Interval-scale questions (1=Not confident at all, 7=Very confident) on how confident they are with 1) judging the quality of a health news story, 2) pointing out inadequacies of a health news story, and 3) suggesting ways to improve a health news story. Figure 3.4 shows the average pre/post self-efficacy scores for each condition.



(a) Judging the quality

(b) Pointing out inadequacies



(c) Suggesting ways to improve

Figure 3.4: Average scores for self-efficacy questions in pre/post-survey for each condition

For the first question on judging the quality, participants showed high confidence in the pre-survey. Average scores in the pre-survey were 5.14 (SD: 1.55), 5.24 (SD: 1.22), and 5.26 (SD: 0.86) for Baseline, SubCriteria, and ReviewAid condition. Average scores in the post-survey were 4.50 (SD: 1.41), 5.19 (SD: 1.29), and 5.26 (SD: 1.14). In the Baseline condition, participants decreased the level of confidence in the post-survey (Wilcoxon Signed-Rank Test, $W=26$ and $p<0.05$). There was no significant change between pre and post-survey in SubCriteria and ReviewAid conditions.

For the second question on pointing out inadequacies, average scores were 4.77 (SD: 1.80), 4.86 (SD: 1.39), and 4.91 (SD: 1.16) in the pre-survey and 5.00 (SD: 1.54), 5.48 (SD: 1.33), and 5.70 (SD: 0.97) in the post-survey for Baseline, SubCriteria, and ReviewAid condition. While the average score has increased in all three conditions, the pre-post difference was significant only in the SubCriteria and ReviewAid condition (Wilcoxon Signed-Rank Test, $W=11$ ($p<0.05$) and $W=20$ ($p<0.05$), respectively).

For the third question on suggesting ways to improve, average scores were 4.86 (SD: 1.88), 4.48 (SD: 1.63), and 4.61 (SD: 1.53) in the pre-survey and 5.18 (SD: 1.68), 5.00 (SD: 1.58), and 5.43 (SD: 1.12) in the post-survey for Baseline, SubCriteria, and ReviewAid condition. While the average score has increased in all three conditions, the pre-post difference was significant only in the ReviewAid condition (Wilcoxon Signed-Rank Test, $W=26$ and $p<0.05$).

H5. Participants hold a more critical view of the story with ReviewAid.

We asked participants how much they agree with the claim 'Tofu is good for one's memory in general' before and after the reviewing task. By comparing the score before and after the task, we can measure how much the participants updated their belief on the subject matter. As the news story was about a negative correlation between tofu consumption and memory functionality, the decrease in the score indicates that participants are more persuaded by the news story.

Average scores were 3.68 (SD: 1.25), 3.90 (SD: 1.00), and 3.96 (SD: 0.93) in the pre-survey and 2.77 (SD: 1.38), 3.14 (SD: 1.11), and 3.43 (SD: 1.04) in the post-survey for Baseline, SubCriteria, and ReviewAid condition. The scores have decreased in the post-survey in all three conditions, however, the

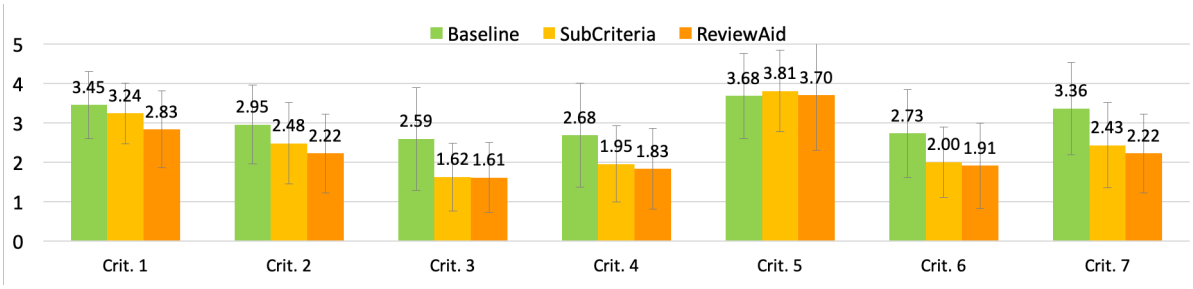


Figure 3.5: Average scores given to the news story for each criterion, for each condition

change was significant only in the Baseline condition (Wilcoxon Signed-Rank Test, W=21.5 and $p<0.05$).

In addition, participants in ReviewAid condition gave lower scores for the news story. Fig. 3.5 shows average scores given to the story for each criterion. In most criteria, the Baseline condition had the highest average score, and the ReviewAid condition had the lowest average score. While the Baseline and the other two conditions showed a notable difference in average scores, the difference between the SubCriteria and ReviewAid was not noticeable in most cases. Despite such tendency observed in this study, however, we do not claim that this result is generalizable as scores given to a news story and their differences can depend on the news story being evaluated.

3.6.3 Discussion

Role of the high-level criteria

Having evaluation criteria itself was welcomed by participants as they provide explicit points to consider. However, the self-efficacy score in judging the quality of news story was decreased in the Baseline condition. Some participants in the Baseline noted that they were not sure about their review and needed more guidance. P6 in the Baseline said *"I felt a little lost in a sense. I knew what the question was asking, however, I didn't know if what I was answering was correct."* However, there was no such decrease in the other two self-efficacy measures. This aligns with our formative study result that participants said that having the detailed criteria helps them learn what to consider.

Role of the subcriteria

Having subcriteria, not necessarily with the scaffolded evaluation process, had a positive effect on the concreteness and appropriateness of review and self-efficacy in pointing out inadequacies. This suggests that the subcriteria were helpful in providing a basis for each high-level criterion and shaping thoughts. P19 in the SubCriteria condition said, *"The subcriteria were very helpful in that it made it simple to know which direction to go with the reviews and what was most important to include specifically within each category (high-level criteria)."* Supporting this, participants in the SubCriteria condition held a more critical perspective on the story.

Role of the scaffolded evaluation process

Compared to the SubCriteria group, the ReviewAid group provided more rationales that were more specific to the news story and appropriate. Participants said distinguishing the media and the research helped them think what aspects the story did and didn't address, and write more accurate review. P21 said, *"I think that it added to my ability to see piece by piece what was affecting my opinion and better*

articulate my thoughts in my reviews." In addition, many participants said that the examples increased their understanding and helped them to write a review specific to the story. P14 said, *" [Examples] helped me fully understand what was being asked and how it applied to this specific article. I feel like it kept my reviews very focused."*

Concern on readers being overly critical

In our study, participants in SubCriteria and ReviewAid hold a more critical view than those in the Baseline. While holding a critical view is important in reading and understanding health news stories, there can be a concern on readers being overcritical and blaming the news story and research more than they deserve. This problem of novice reviewer being overcritical was raised and discussed in other domains such as academic inquiry [119]. Although such problem may disappear as the reviewer gets more experienced, calibrating the level of criticism by comparing their review with expert's one can help[119, 120].

Participants' suggestion on the options in the scaffolded evaluation

While most participants valued the idea of evaluating by aspects (coverage, validity, and explanation), some participants said they want to have more options to evaluate each aspect. For the validity aspect, one participant said that he wanted to have an option such as 'limited validity.' and two participants noted that there were some cases that they were not sure about the validity. For the media aspects, four participants said that there were some cases that some subcriteria are essential in the story and want to ignore how the story addressed it.

These suggestions are closely related to the accuracy of review and score, and self-efficacy that we aim to improve in ReviewAid. In addition, these comments show participants carefully engaged in the scaffolded evaluation process, not by rote. We reflected these comments and improved the scaffolded evaluation process in ReviewAid by adding a 'not necessary' option in the media coverage and explanation aspects, and an 'I don't know' option in the research validity aspect.

3.7 Study 2: Understanding the Use of ReviewAid in a Collaborative Setting

In Study 1, we asked participants to review a given news story for all of the seven evaluation criteria. In our study, participants spent more than 20 minutes on average to review the story. This task design was introduced to have participants make the most of the scaffolded evaluation process. However, it limits the generalizability and impact of study findings as there are only limited scenarios (e.g., paid crowdsourcing or class activity) that ReviewAid is used as in the study setting. In this section, we explore a more practical use of ReviewAid by introducing collaboration in the reviewing task. Specifically, we lower the required effort by asking each user to review for some of the criteria, not necessarily all, and share the reviews among users and form a collective evaluation. In addition, we give users more choices in news stories so that each user can review news stories that they are interested in.

3.7.1 Extending ReviewAid to Support Collaborative and Distributed Review

In this subsection, we present how ReviewAid can support collaborative and distributed review among readers. In the collaborative version of the interface, user can see other's review and score, as well as number of evaluations made on each news stories. Individual evaluation interactions remain the same as in the individual user scenario.

List of stories with evaluation overview

When the user enters the system, a list of health news stories is shown as in Figure 3.6. For each story, a preview (title, image, and leads) of the story and an overview of readers' evaluation (purple boxes in Figure 3.6) are shown. The evaluation overview shows the evaluation status (complete/incomplete), average score, and a list of readers who reviewed the story. When the evaluation is incomplete, it shows the number of criteria reviewed so far, and the average score in low opacity.

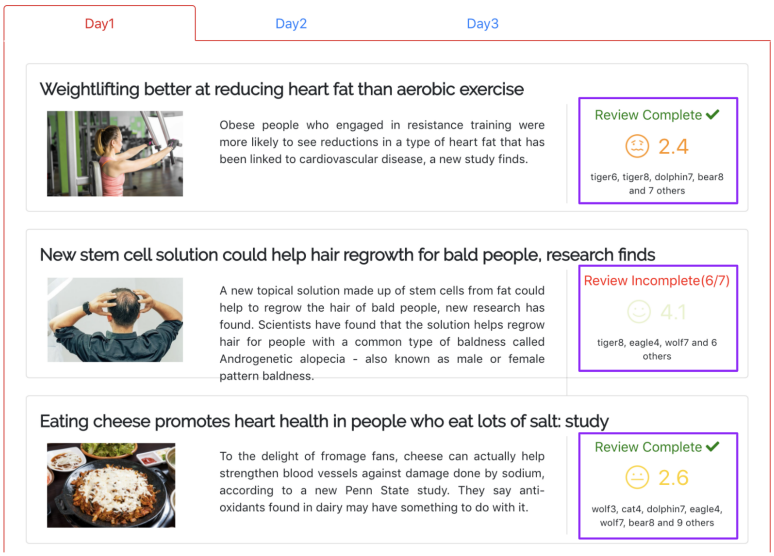


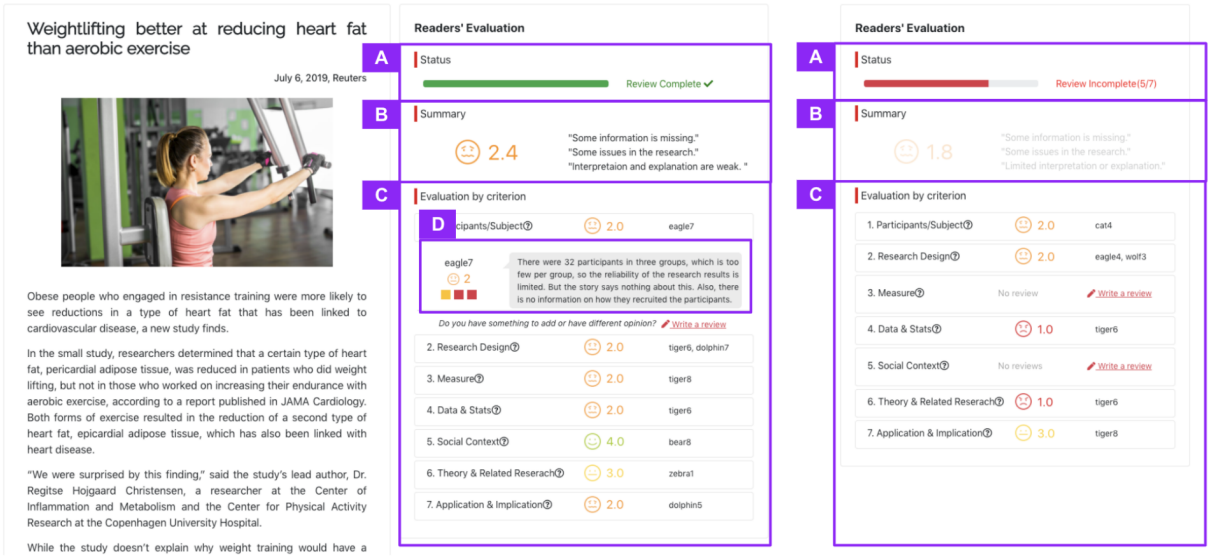
Figure 3.6: List of news stories are shown with a preview of each news story and an overview of readers' evaluation (purple boxes).

Evaluation dashboard

Upon selecting a story from the list, the user can read the news story and look through existing reviews as in Figure 3.7. The evaluation status (Figure 3.7a-A) and the summary of evaluation (Figure 3.7a-B) are shown on the top. The summary contains the average score and summary of the assessment on media coverage, research validity, and explanation aspects.

A description and readers' evaluations on each criterion are shown in a list (Figure 3.7a-C). Each list item contains the average score for the criterion and the nickname of reviewers (if any). Existing reviews are shown when a user clicks on each criterion (Figure 3.7a-D). It shows the score given by the reviewer, summary flags for each aspect, and the review text.

In Study 2, we explore a collaborative use of ReviewAid. First, we study how distributed work happens when each user can choose news stories and criteria to review. Specifically, we aim to understand



(a) Story and dashboard (complete)

(b) Dashboard (incomplete)

Figure 3.7: While reading a news story, users can see other readers' evaluation on the right side. (A) Reviewing status is shown at the top. (B) Summary of readers' evaluation with average scores and summary for each aspect. (C) Readers' evaluation by each criterion. (D) When the user selects a criterion, the evaluation scores and reviews of each reviewer are shown.

participant's consideration in their choice of story and criteria and how they result in the generation of collective outcome. As each participant can have different considerations in choosing stories and criteria to review, it is important to understand their behavior and how individual choices result in the construction of collective reviews.

Second, we study how seeing *readersourced reviews* affects readers' reviewing experience. Seeing others' work can have both positive (e.g., peer learning [121]) and negative (e.g., bias[122, 123]) effects. We take a closer look at what kinds of positive and negative experience the collaboration gives in reviewing health news stories.

Lastly, we evaluate the effect of the scaffolded evaluation process in a collaborative setting. Increased flexibility in reviewing tasks and access to others' reviews can affect the difficulties that readers face when reviewing health news stories. For example, readers can choose stories and criteria that they think are easy to evaluate or learn from other readers' reviews. Therefore, the important question in this regard is whether the scaffolded evaluation process positively affect the quality of reviews and self-efficacy in a collaborative setting.

To summarize, our guiding research questions were:

RQ1. What do readers consider when choosing news stories and criteria to review?

RQ2. How does having readersourced reviews affect readers' reviewing experience?

RQ3. Will the benefits of the scaffolded evaluation process be preserved in the collaborative and practical use of ReviewAid?

For RQ3, we compared SubCriteria and ReviewAid conditions, testing whether ReviewAid leads to higher quality reviews (H1), increased self-efficacy (H2), and higher perceived reliability of readersourced reviews (H3).

3.7.2 Method

Study Design and Conditions

We conducted a five-day-long, between-subjects study with two conditions: SubCriteria and ReviewAid. In both conditions, participants chose news stories and criteria to review. SubCriteria participants viewed subcriteria for 2 minutes before scoring and writing reviews. ReviewAid participants followed the scaffolded evaluation process before providing scores and reviews.

Participants

We recruited 24 participants by posting a call for participation in an online community of a technical university in Korea. We conducted the study in Korean so that participants can read news stories and review in their native language. Participation was limited to people who had no prior research experience. Twelve participants were undergrads, and the other twelve participants had jobs that are not related to the research. We randomly assigned 12 participants to each condition. The average age was 29.7 (SD: 9.70) with min. 20 and max. 57. Each participant received KRW 50,000 (∼USD $45) for their five-day-long participation.

Tasks and Procedure

The study was conducted over five days. On Days 1-3, six health news stories were uploaded daily. Participants selected at least two stories to review for at least one criterion each. On Days 4-5, participants reviewed at least one story (choosing from all 18) for at least one criterion. Stories covered topics relevant to participants (nutrition, exercise, obesity, COVID-19), published within six months. Each day included two controlled studies and four observational studies.

Participants answered a pre-survey at the beginning of the experiment. As in Study 1, we asked questions on factors that may affect their experience and outcome of the main task and participants' self-efficacy in reviewing a health news story. We also measured their expectations on the other's ability to review. Post-survey assessed self-efficacy, perceptions of others' ability, experiences with criteria and scaffolding (ReviewAid), and strategies for selecting stories and criteria to review.

Measuring the quality of reviews

We analyzed a subset of review texts constructed by randomly sampling one review text per participant per day (24 x 5 = 120). Two external coders analyzed the review texts as in Study 1. In the first step, with the first author's guidance, the two coders coded 21 (3 for each of 7 criteria) review texts together and then coded the next 21 review texts independently, compared the result, and discussed to reach an agreement. After building consistency, the remaining review texts were coded in a distributed manner: each coder coded half of the remaining review texts. The second and third steps (categorization and labeling) were done as in Study 1 but in a distributed manner. After building consistency (coding together, and independently and compare), they categorized or labeled each half of the remaining data.

A total of 120 review texts contained 371 arguments. There were 27 (7.3%), 267 (72.0%), 55 (14.82%), and 22 (5.9%) arguments in general, rationale, repeat, and other types, respectively. Out of 267 rationale type arguments, 239 (90%) were concrete, 159 (60%) were specific, and 47 (18%) were inappropriate.

Table 3.7: Number of stories reviewed for at least 1, 4, and 7 (all) criteria for up to each day, for each group

	SubCriteria			ReviewAid		
Coverage	Day 1-3	Day 1-3, 4	Day 1-3, 4, 5	Day 1-3	Day 1-3, 4	Day 1-3, 4, 5
1/7	18	18	18	18	18	18
4/7	18	18	18	10	14	16
7/7	10	13	15	2	3	7

Table 3.8: The number of reviews generated and the number of reviewers (unique) for each criterion

	Crit. 1	Crit. 2	Crit. 3	Crit. 4	Crit. 5	Crit. 6	Crit. 7
# of reviews	54 (19.5%)	41 (14.8%)	37 (13.4%)	35 (12.6%)	31 (11.2%)	34 (12.3%)	45 (16.2%)
# of unique reviewers	18	18	17	16	13	17	19

3.7.3 Result

Overall, participants showed a modest level of confidence in their ability to evaluate health news stories. The average self-efficacy score was 3.9 (SD: 1.09), and the average perceived media literacy score was 4.45 (SD: 0.94). There were no between-group differences in self-efficacy and media literacy, as well as in the perception of health news and the need for cognition.

Participants wrote 277 reviews in total. There were 168 and 109 reviews made in the SubCriteria and ReviewAid conditions, respectively. Out of total 18 news stories, reviews for 15 and 7 news stories are completed in the SubCriteria and ReviewAid group. Table 3.7 shows the number of stories that are reviewed for at least 1, 4, and 7 (all) criteria over time. On the first day, one participant in the SubCriteria condition wrote 13 reviews by misunderstanding the required task. Participants who wrote the most in each condition wrote 43 reviews in SubCriteria and 15 reviews in ReviewAid condition.

The most reviewed criterion was Criterion 1 (Participants & Subject) while the least reviewed one is Criterion 5 (Social Context). Table 3.8 shows the number of reviews generated and the number of unique reviewers (who selected the criterion at least once) for each criterion.

The average time spent per review was 362.5 (SD: 158.0) seconds in SubCriteria and 297.5 (SD: 171.7) seconds in ReviewAid. Excluding time spent on waiting times in SubCriteria (2 minutes) and the scaffolded evaluation process in ReviewAid, the average time spent on the writing review text was 242 (SD: 158.0) seconds for SubCriteria and 150.0 (SD: 114.4) seconds for ReviewAid.

The average word count of review text written by each participant for each criterion was 26.2 (SD: 15.4) and 26.3 (SD: 13.3) for the SubCriteria and ReviewAid conditions, respectively. The average word count for each day was 27.8 (Day 1), 28.0 (Day 2), 22.9 (Day 3), 27.3 (Day 4), 24.9 (Day 5), and there was no time trend. (Note that review texts were in Korean, thereby comparing these numbers with those in study 1 is inappropriate.)

Below we present results correspond to each of the RQs. As the two conditions (SubCriteria and ReviewAid) were introduced mainly for RQ3, the result and discussion for RQ1 and RQ2 will focus on the general use of collaborative reviewing system, not necessarily distinguishing and comparing two conditions.

RQ1. What do readers consider when choosing news stories and criteria to review?

To better understand the participants' consideration in their selection of news stories, we categorized participants' response on how they chose news stories to review in Day 1-3, and Day 4-5. For the question on Day 1-3, 16 participants said they chose news stories related to their everyday life (personal relevance). Four participants said they chose news stories that seem absurd or want to question (curiosity on topic). Two participants said they picked news stories that are easy to review. On the other hand, the other two participants said that they cared more about the groups' work and tried to review stories with incomplete review.

While most participants chose news stories following their personal interest (personal relevance and curiosity on topic) on Day 1-3, only six participants had that consideration on Day 4-5. Rather, 10 participants said that they tried to review stories with the least number of reviews or whose reviews are incomplete. Four participants noted that they chose stories with high and low scores as they became curious about the quality of the story and wanted to review those stories by themselves. Three participants said they tried to review stories with the most number of reviews they can refer to and compare with his evaluation. Table 3.9 summarizes this result.

Table 3.9: Categorized considerations for choosing news stories, with representative responses and frequencies

Category	Representative Response	Day 1-3	Day 4-5
Personal Relevance	P18: *"I reviewed stories that have a lot to do with my life. I chose stories on coffee as I drank coffee a lot these days. ... "*	16	1
Curiosity on Topic	P9: *"After reading the titles, I picked a news story that seemed absurd or interesting. ... "*	4	5
Level of Difficulty	P7: *"I chose news stories that are easy to understand and evaluate."*	2	1
Completness	P23: *" My prior consideration was whether the review was complete. I thought each story is properly assessed only if there is at least one evaluation for each item. ... "*	2	10
Curiosity on Score	P2: *I was interested in stories with low or high review scores."*	0	4
Reference	P8: *" I chose stories with many reviews from others that I could refer to."*	0	3

For the question on how they selected criteria to review, 11 participants said that they chose criteria that seem to capture weak points of the news story. P7 said that *"... If a story does not have information on the participants, I choose that criterion. If I think not enough data is presented, I choose the criterion on data. ..."* On the other hand, eight participants said that they cared more about the completeness of collective review and reviewed criteria that are not yet reviewed by others. There were five participants who said that they had a certain criterion they prefer to review. However, the preferred criterion was different for each participant.

When we take a look at how individual reviewers choose criteria, four participants wrote reviews for all of seven criteria. On the other hand, there was one participant (P22) who chose Crit. 1 in all of his reviews. Table 3.10 shows the number of participants (frequency) by the unique number of criteria reviewed. Most of the participants reviewed news stories for more than 4 (out of 7) criteria. This result

Table 3.10: The unique number of criteria reviewed by each participant and the number of participants (frequency)

# of unique criteria reviewed (out of 7)	1	2	3	4	5	6	7
# of participants	1	0	0	8	9	2	4

aligns with the participants' responses that they chose criteria that the evaluation is most needed.

RQ2. How does having readersourced reviews affect

readers' revieweing experience? In the post-survey, we asked the participants to describe their experience of seeing others' reviews and how it affected their review. To analyze the responses, we 1) split each response into multiple phrases so that each has one perspective, 2) conducted open coding with the phrases, 3) and categorized the codes.

Benefits of having diverse perspectives Seventeen participants said they enjoyed seeing reviews made for each criterion as they broadened their perspective. Also, five participants said that their understanding of the news story has improved by seeing others' reviews.

Learning from others' reviews Seeing readersourced reviews also served as a learning opportunity. Two participants said that they could understand how each news story is limited in certain criteria by reading others' reviews. Three participants said that they could better understand how to review for a specific criterion. Among them, two participants explicitly noted that they referred to others' reviews when they want to evaluate a story for a specific criterion not familiar to them.

Comparing own evaluation with others Four participants said that having reviews that do not align with theirs makes them reflect on their own thoughts. P4 in ReviewAid noted, *"· · · I reread the story and re-evaluated the story more intensively."* Three participants noted that reviews that align with theirs reinforce their thoughts. P20 said that it made him think he did a good job in reviewing and feel more confident about the reviewing task.

Seeing others' evaluation before reading a news story Three participants noted that other readers' evaluation makes preconceptions about the quality of the news story. P24 said, *"On the 4th and 5th day, when judging a story that I hadn't read, the final score in the dashboard gave me a preconception on the news story. Now I understand why, in the real world, comments or ratings in news stories or posts matter."*

Conforming to others' evaluation Four participants said that they tried to match their evaluation to others' evaluation. P19 noted, *"· · · I tried to give scores that are close to other participants'. "*Two participants said they hesitated to write a review that does not align with existing ones.

RQ3. Will the benefits of the scaffolded evaluation process be preserved in the collaborative and practical use of ReviewAid?

H1. Participants write a higher quality review with ReviewAid.

To measure the quality of reviews, we use the measures that we used in Study 1 – number of rationales, concreteness and specificity, and appropriateness.

The average number of arguments made in a review text was 3.09 (Median: 3, SD: 1.50). The average for each condition were 3.23 (Median: 3, SD: 1.59) and 2.95 (Median: 3, SD: 1.45) for SubCriteria and

ReviewAid, respectively. Table 3.11 shows the average number of each type of arguments per review text in each condition.

Table 3.11: Average number of each type of arguments per review text in each condition

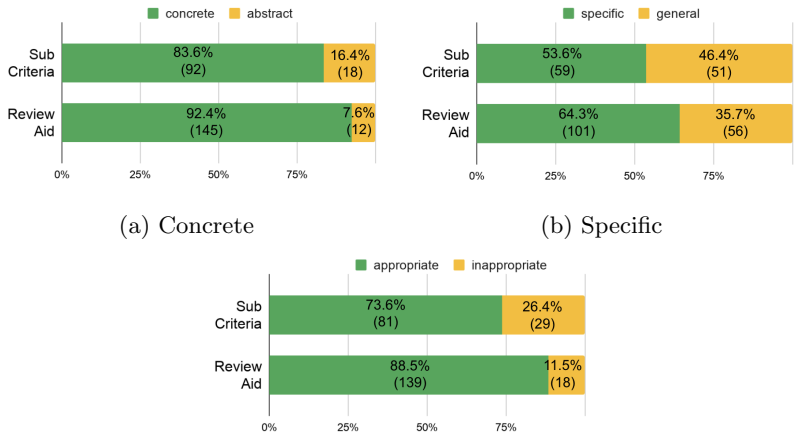
	General	Rationale	Summary/Repeat	Other	Total
SubCriteria	0.33 (Med: 0, SD:0.60)	1.83 (Med: 2.62, SD:1.40)	0.77 (Med: 0, SD:1.07)	0.30 (Med: 0, SD:0.64)	3.23 (Med: 3, SD:1.59)
ReviewAid	0.12 (Med: 0, SD:0.35)	2.62 (Med: 2, SD:1.44)	0.15 (Med: 0, SD:0.42)	0.07 (Med: 0, SD:0.24)	2.95 (Med: 3, SD:1.45)

Number of rationales The average numbers of rationales were 1.83 (Median:2, SD:1.40) and 2.62 (Median:2, SD:1.44) for SubCriteria and ReviewAid. The difference was statistically significant (Mann-Whitney U-Test, p< 0.01, Z=2.79)

Concreteness The number of concrete rationale were 1.53(Median:1, SD:1.25) and 2.42 (Median:2, SD:1.56) for SubCriteria and ReviewAid. To exclude the effect of providing more (or less) rationales, we compared the concreteness of each rationale between conditions. Rationales provided in the ReviewAid condition tend to be more concrete than in the SubCriteria. The ratios of concrete rationale were 0.84 and 0.92 for SubCriteria and ReviewAid. The difference was statistically significant (χ^2 test, p<0.05 with χ^2=4.93).

Specificity The number of specific rationale were 0.98(Median:1, SD:1.04) and 1.68 (Median:2, SD:1.39) for SubCriteria and ReviewAid. Rationales provided in the ReviewAid condition tend to be more specific than in the SubCriteria. The ratios of specific rationale were 0.54 and 0.64 for SubCriteria and ReviewAid. However, the difference was not significant at p<0.05 level. (χ^2 test, p=0.080 with χ^2=3.08).

Appropriateness Out of 108 review texts with rationale, 29 (26.9%) contained inappropriate arguments. The number of review text with inappropriate rationales were 17 (out of 48, 35.4%) and 12 (out of 60, 20.0%) for SubCriteria and ReviewAid. The average number of inappropriate rationales was 0.48 and 0.30 per review for SubCriteria and ReviewAid. The ratios of inappropriate rationale were 0.26 and 0.11 for SubCriteria and ReviewAid. The difference was statistically significant (χ^2 text, p<<0.001 with χ^2=18.4).



(a) Concrete

(b) Specific

(c) Appropriate

Figure 3.8: Ratio of concrete (a), specific (b), and appropriate (c) rationales in each condition

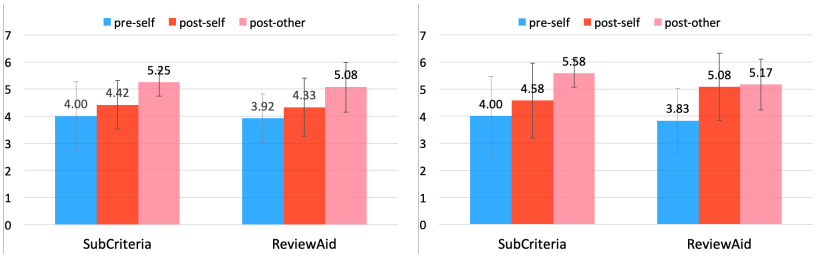
H2. Participants hold a higher level of self-efficacy with ReviewAid.

We measured the participants' self-efficacy in the reviewing task by asking three 7-point Interval-scale questions on how confident they are about 1) judging the quality of a health news story, 2) pointing out inadequacies of a health news story, and 3) suggesting ways to improve a health news story. Average scores for pre- and post-survey are presented in Figure 3.8 (blue: pre, red: post).

Judging the quality of a health news story Average scores for the first question were 4.00 (SD: 1.28) and 3.92 (SD: 0.90) in the pre-survey and 4.42 (SD: 0.90) and 4.33 (SD: 1.07) in the post-survey, for SubCriteria and ReviewAid conditions. There was no significant change between pre and post-survey in both conditions.

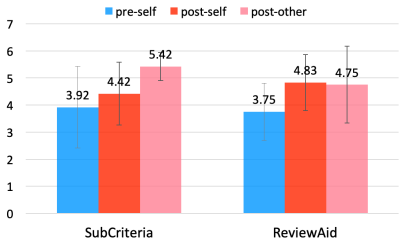
Pointing out inadequacies in a health news story Average scores for the first question were 4.00 (SD: 1.48) and 3.83 (SD: 1.19) in the pre-survey and 4.42 (SD: 1.16) and 4.83 (SD: 1.03) in the post-survey, for SubCriteria and ReviewAid conditions. Only ReviewAid condition showed a statistically significant change in pre-post survey (Wilcoxon Signed-Rank Test, $p<0.05$ with $W=0$, which means that no one lowered the score in the post-survey).

Suggesting ways to improve Average scores for the first question were 3.92 (SD: 1.51) and 3.75 (SD: 1.06) in the pre-survey and 4.58 (SD: 1.38) and 5.08 (SD: 1.24) in the post-survey, for SubCriteria and ReviewAid conditions. Only ReviewAid condition showed a statistically significant change in the pre-post survey (Wilcoxon Signed-Rank Test, $p<0.05$ with $W=0$, which means that no one lowered the score in the post-survey).



(a) Judging the quality

(b) Pointing out inadequacies



(c) Suggesting ways to improve

Figure 3.9: Average scores for efficacy questions in pre-survey (self) and post-survey (self and others) for each condition

H3. Participants hold a higher level of perceived reliability of readersourced reviews in ReviewAid.

In the post-survey, we asked how well they think other people in the group reviewed news stories. We asked the same three questions that we used to measure self-efficacy. Participants in the SubCriteria gave higher scores to others than themselves for all three questions ($p<0.05$ for all three questions, Wilcoxon Signed-Rank Test), while it was not true in the ReviewAid condition. The average scores for

each condition were 5.25 (SD: 0.45) and 5.08 (SD: 0.90), 5.58 (SD: 0.51) and 5.17 (SD: 0.94), and 5.42 (SD: 0.51) and 4.75 (SD: 1.42) (pink in Figure 3.8).

Although participants in the ReviewAid condition showed an increase in self-efficacy in pointing out inadequacies and suggesting ways to improve, it did not make them have a higher evaluation of others. Rather, the average scores were higher in the SubCriteria group, but the differences were not statistically significant.

3.7.4 Discussion

In Study 2, participants chose 1-2 news stories per day from the list of stories we provided. The most popular considerations were personal relevance and curiosity on topic. Given that these are what general people consider when choosing news stories to read[56], it gives more generalizability to the effect of the scaffolded evaluation process.

Seeing others' reviews had mixed effects. Some participants were motivated to contribute their own perspectives, while others expressed concern that reading existing reviews might bias their review. Features to limit the effect of such preconceptions can be introduced while still encouraging positive social elements. For example, presenting the review only to the reader who wants or who already reviewed it, or only partially showing the social signals before a user submits a review could be possible.

In the collaborative setting, users are not guided to go over the criteria in full. Previous research in crowdsourcing has shown that distributing the work can make the workers miss the high-level understanding of the task [124, 125, 126, 127]. Although comparing Study 1 and Study 2 might not be fair, the ratio of inappropriate rationales was higher in the collaborative setting. Given that inappropriate rationales in ReviewAid were mostly from being under a wrong category, reducing the task load might have lowered the participants' understanding of each criterion and high-level understanding of the evaluation criteria.

3.8 Discussion

In the following subsections, we discuss the implication and limitations of ReviewAid project. Then we propose directions for future research.

3.8.1 Implication

ReviewAid can benefit readers in two ways. First, following the evaluation process can benefit individual readers. Readers can hold a more critical perspective on the news story and better elaborate their thoughts. Also, it can serve as a learning opportunity for readers as they can practice their critical thinking skills and increase their awareness of miscommunication problems in health news.

Second, the review generated with ReviewAid can benefit the readers and the community. Reader-sourced reviews can guide other readers in understanding and evaluating health news stories by increasing their awareness of miscommunication problems and broadening their perspectives. Also, more broadly, the collected reviews can serve as feedback to scientists and science journalists by giving structured information [102] about which part or aspect of the story the readers are satisfied or dissatisfied with.

However, these benefits depend on the context of use and user motivation. ReviewAid requires more effort than simply reading headlines or skimming articles, which means it is most appropriate when users are motivated to engage deeply with health information. Our participants were asked to use the system

as part of a study, but in reality, not all readers will have the time or motivation to conduct detailed evaluations. The collaborative version of ReviewAid showed that seeing what others thought and did can increase engagement, suggesting that social features may help sustain use. Additionally, users may engage with ReviewAid not while initially reading an article, but afterward, similar to how people read comments to understand public reactions. In such cases, ReviewAid could serve as a tool for gauging how others have interpreted and evaluated a health news story, rather than as a primary reading aid. This suggests that ReviewAid is most valuable in contexts where users have a specific intention to understand health claims more thoroughly or to explore how others have assessed the information.

The identified design guidelines and the designed scaffolded evaluation process can have a greater impact when introduced to or integrated with online news media platforms or fact-checking websites. When applied to real-world platforms, the task should be carefully designed considering the trade-offs between the task load, participation rate, and the quality of the review. Also, the quality control measure should be considered against the potentially deceitful and harmful behaviors of online readers.

3.8.2 Limitations

While results from the two studies show promising effects of the scaffolded evaluation process, there are limitations of the current system and study that should be considered.

First, our studies relied on short-term measures and self-reported assessments. Both Study 1 and Study 2 captured participants' immediate perceptions and engagement with the system, but did not track whether these effects persisted over time or translated into different information-seeking behaviors in real-world contexts. Measures such as self-efficacy and perceived difficulty reflect how participants felt about the task, but do not necessarily indicate actual improvement in evaluation quality or long-term engagement with health news. Longitudinal studies would be needed to understand whether the benefits of ReviewAid extend beyond initial use.

Second, the proposed evaluation criteria have not been separately validated. Our design guidelines include both the custom evaluation criteria and the scaffolded evaluation process. Our studies measure the effectiveness of both components as a whole, and it remains unanswered if different evaluation criteria could be integrated into the current scaffolded process, and vice versa.

Third, our analysis of review text was done on subsets of the collected data. In our study, participants generated a large number of reviews as they were guided to review all of the seven criteria in Study 1, and with more freedom in the selection of stories and criteria over 5 days in Study 2. We picked two criteria (out of seven) to review in Study 1 and we randomly sampled review texts in Study 2. With the extensiveness of review text, we believe that it was a sensible decision, but the validity of the findings could be limited.

Fourth, our studies had limited control of participants' prior knowledge and beliefs on the subject matter. One's prior knowledge and belief on the subject matter can significantly affect the perceived level of difficulty of reviewing tasks. Therefore, separating their effect from the effect of the scaffolded evaluation process can be important. However, we measured the pre and post self-efficacy in general, not for each topic or news story, and this makes it hard to tease out the effect of knowledge and beliefs on each subject from the observed pre-post changes.

In addition, the reliability of readersourced evaluation can be further validated. In this work, we evaluated the reliability of reviews by analyzing the appropriateness of each argument made in the review text. In the ReviewAid condition, 93.8% and 99.5% of arguments were evaluated as appropriate in Study

1 and 2, respectively. However, this validation does not verify if readersourced reviews have the quality that is comparable to experts'.

3.8.3 Future work

We identify several directions for future work that might further expand our knowledge of supporting non-experts' collaborative review of scientific research. The evaluation criteria used in ReviewAid were adjusted to non-experts, and the scope of information the readers in ReviewAid investigate is limited compared to what experts use (e.g., [61]). In the future, we plan to explore how experts and non-experts can collaboratively generate quality reviews in a complementary manner.

In this research, we propose ReviewAid as a scalable solution for improving public communication of health-related research. A live deployment of the system could answer important remaining questions in this research, including incentive design, quality control, and long-term effects of reviewing. To collect more realistic usage data, we plan to build a web browser extension or a mobile app to add support for any news article people read.

While we identified the design goals specific to the domain of health news, the guidelines may be applicable to other domains that need non-experts' evaluation of information. For example, the guideline about helping readers to disentangle media and research could be transferred to other domains such as news in other scientific domains as well as political news. With ReviewAid, we envision a future where non-experts can actively engage in a structured review process to make meaningful contributions in evaluating various types of information.

3.9 Conclusion

In this project, we investigated challenges that non-experts face while evaluating health news stories and identified design guidelines for a system supporting them. Following the guidelines, we implemented ReviewAid, a system that supports readers' collaborative review of health news stories. ReviewAid provides evaluation criteria designed for non-expert readers and guides the evaluation with the scaffolded evaluation process. We evaluated the effect of the scaffolded evaluation process by conducting two studies in an individual and collaborative setting.

Chapter 4. PRISM: Capturing Diverse and Precise Reactions to a Comment with User-Generated Labels

This chapter presents PRISM, an interactive system that helps users express and explore nuanced reactions to online discussions by capturing the heterogeneity of viewpoints before they are flattened into aggregate metrics. This chapter focuses on heterogeneity by examining how interpretations and reactions differ across perspectives and how interface design can preserve this diversity rather than collapse it into simplified signals. Through a user study, I demonstrate how PRISM's user-generative labels reaction mechanism enables richer expression of viewpoints and supports better understanding of diverse perspectives in online discourse. This chapter is based on a paper published at WWW 2022 [128]. All uses of "we", "our", and "us" in this chapter refer to coauthors of the paper.

4.1 Introduction

Online comment sections open up public discourse and facilitate interactions among users, including original commenters, reactors, and viewers. As comment sections play a vital role of encouraging engagement from users, one can easily find them across the web in many social media sites. Within these comment sections, users' comments often invite subsequent comments or reactions like up/downvotes. Whereas reply comments allow informationally rich input, reactions provide easier ways to express one's opinion. Once accumulated, such reactions influence the way users perceive and interact with comments [129, 130], their perception of the public opinion [131, 132, 133, 134, 135], and their views of the relevant topic [136, 137, 135, 138].

Whereas up/down votes are the most widely used form of reaction, it is questionable whether they provide sufficient context to readers on why or how people are responding to comments. In our formative study, which aimed to understand how people use and interpret up/downvotes, we found that people's considerations for diverse aspects of a comment were projected onto up/downvotes. Our interviewees noted that they were unable to precisely express what they felt, especially when they have mixed or moderate opinions, and this discouraged them from leaving reactions to comments. However, when they were asked to share their interpretation of aggregated votes, most interviewees simplistically interpreted up/downvotes as users' (dis)agreement to a comment.

In this work, we explore the potential of having users cooperatively generate labels that capture different aspects of a comment that people focus on when they react to a comment. We suggest user-generated labels (UGLs) (Figure 4.1) as an alternative reaction design that captures rich context of user reactions. For each comment, users can create their own UGLs or add votes to UGLs generated by other users to efficiently indicate what aspect of the comment deserves an up/downvote.

We conducted a between-subjects study with 218 participants to explore the effect of UGLs in capturing diverse and precise reactions and users' understanding of multifacetedness of reactions to a comment. We compared the baseline condition, in which participants can leave up/downvotes and reply to a comment, and the UGL condition, in which participants can generate UGLs, vote on UGLs, and reply to a comment. In the UGL condition, participants generated 234 unique labels regarding the degree of agreement, the strength of the argument, the style of the comment, judgments on the commenter, and feelings or beliefs related to the topic. With UGLs, participants felt that their reactions were more precise

and unique, and left more reactions. Moreover, participants better understood the multifacetedness of others' reason for reacting to a comment with UGLs compared to the baseline condition.

The contributions of the project are as follows:

- User-generated labels (UGLs), a reaction design that captures diverse and precise reactions to a comment.
- Observations from a formative study and a user study that provide insights into dimensions users consider when they react to comments.
- Findings from a user study that show (1) UGLs enhance users' ability to express and interpret others' reactions, and (2) users leave more reactions and better understand the multifacetedness of opinion towards a comment.

4.2 Background and Related Work

In this section, we review previous work that explored the effect of diverse reaction designs. Then we discuss previous work on leveraging the intrinsic motivation of users in social systems. Lastly, we discuss the affective polarization and its relation to the perceived diversity in public opinion.

4.2.1 Choice of Reaction Buttons

While up/downvotes are most commonly used, there are different types of reactions available across various platforms. These include like/dislike, emojis, recommend, and the heart button. However, binary buttons like up/downvotes may vary in their meanings across contexts, sites, and users [139]. Inconsistency across users' intention for choosing from pre-set reactions causes lack of clarity. Emojis are often used ironically; individuals may also use them to signal different levels of agreement [135]. As such, current reaction modalities do not effectively present these multidimensional messages, limiting users' abilities to infer information behind the reactions.

Prior work suggested different preset categories. Sumner et al. [2018] suggested Interesting, Amused, and Love buttons as possible additions to a like button for users to be more specific about their positive connection with a comment's content [140]. Nevertheless, these categories still cannot cover evaluation of a comment with mixed sentiments. Based on the Stereotype Content Model, Stroud et al. [2017] proposed the use of "Respect" over "Like" and "Recommend" as an alternative reaction button to reduce users' hostility towards comments with opposite political attitudes [141]. However, "Respect" only covers a relational connection with the comment; it forgoes the evaluative functionality of an upvote.

Evidently, reactors' ability to express is hindered in current social media comment sections as available reactions fail to coherently communicate to other users as to why reactors upvoted a comment. Without placing too much cognitive load on their users, UGLs aim to relay crowds' underlying sentiment and evaluation of a comment in more detail across users.

4.2.2 Designing Incentives for Users to Participate

User comments and reactions are valuable information that people read and interpret to understand public opinion and develop their thoughts. However, only a small portion of users leave comments or reactions and most users prefer to lurk, observing other people's thoughts but not sharing their own [142, 143]. Although lurking is a natural and valuable activity that makes the shared opinions

heard [144], user participation in adding their own comments and reactions is essential to maximize the condition for this public good [145].

Previous studies have explored diverse social-psychological incentives to promote users' participation. Researchers found a positive relationship between individual engagement online and social factors such as expectation of social approval [145, 146], social transparency [147], and reciprocity [148]. Other researchers have shown that intrinsic factors such as enjoyment [149, 150], self-expressiveness [151], and perceived impact [152] and uniqueness [153, 154, 155] of individual contribution increase user participation.

In this work, we explore an alternative reaction design that can capture users' detailed opinions on a comment. As describing one's thoughts in detail requires more cognitive effort from users, it should be designed in a way that triggers users' intrinsic motivation to engage. Reacting through UGLs enables users to create labels in their own words and positively affects the level of and perceived uniqueness of reaction and self-expressiveness.

4.2.3 Affective Polarization and Understanding the Diversity in Public Opinion

Prior work has shown that people overestimate the difference between "the other party" and themselves [156] and underestimate the level of agreement with those who take the opposite stance on an issue [157, 158, 159]. By overestimating disagreement with members from the opposing group, people often become less tolerant toward the outgroup, which lowers their willingness to socialize and have a constructive discussion with the opposite party [157, 158].

The idea of increasing exposure to different points of view is a common intervention suggested by previous studies to decrease polarization. While exposure to diversity is not sufficient by itself to eliminate polarization, many researchers have seen its effect on increasing understanding and decreasing dislike between inter-group members [160, 161]. Our work revisits the effect of exposure to diversity. Whereas most work [162, 45, 30, 160] focuses on the diversity of opinions towards a topic, we focus specifically on the diversity of reactions towards a comment and its effect on reducing affective polarization among users. User-generated descriptive labels have two aims: improve users' understanding of which aspect a comment is valuable and organically set up a space for inter-group members to share their diverse judgment about a comment. We explore whether UGLs affect users' hostility and naive realism, the tendency to consider others' views as homogeneous and incorrect [163]. Specifically, we investigate how UGLs affect users' understanding of multifaceted public evaluation of a comment and tolerance to the opinions that do not align with theirs.

4.3 Formative Study

To understand how users leave and interpret reactions to comments through dichotomous reaction buttons in online discussion, we conducted a series of observations and semi-structured interviews. We recruited 10 participants (6 undergrads, 4 grad students) from an online community of a technical university in South Korea.

4.3.1 Task and Procedure

We implemented a toy comment system where participants can read a news story and user comments, reply and react to user comments, or leave their own comments. We chose three news stories on controversial issues in Korea at the time of study (legalizing abortion, distributing subsidies for COVID-19, and location-tracking to combat COVID-19) from Naver[1]. For each story, we chose ten comments (5 supporting and 5 opposing) from a pool of comments with the most number of votes (upvotes and downvotes).

In each session, participants were asked to think aloud as they read each news story and relevant user comments. During this process, they were able to add comments, reply to comments, or leave up and downvotes. To observe participants' original reaction to a comment without social influence, we hid previously accumulated up/down votes on each comment. After the think-aloud session, we conducted a semi-structured interview and asked participants to describe their positions on the issue and give reasons for their reactions or inactions to each of the comments. At the end of each interview, we revealed the aggregated reactions to the same set of comments to our participants who were then asked to share their interpretation of the reaction statistics along with their perception of the general public opinion.

4.3.2 Observations

We observed that current reaction buttons do not fully capture nor represent users' evaluative judgments about the comment. We present the main findings in more detail below.

A simple up/downvote does not capture nuances and diverse dimensions of people's actual reactions to comments.

Interviewees had different rationales for using the same reaction button. For example, P3 said he downvotes to signal disagreement with the claim made in a comment. P2 said he downvotes when a rationale provided in a comment is not valid. P1 gave an upvote for effort, while P5 upvoted a comment for humor. All of these diverse evaluation aspects were projected onto either an upvote or downvote.

Also, there were individual differences in people's thresholds in up/downvoting a comment. When interviewees partially agreed or disagreed with a comment, some people projected that opinion into an up/downvote while others did not. Some interviewees did not leave reactions when they have a mixed evaluation of a comment (e.g., disagree but well articulated). It is possible to express these mixed evaluations through a reply comment, but participants said that they reply to comments only when they feel the strong urge to say something. Compared to reactions, we can assume that replying to comments requires much more effort from the user and is only utilized when users have a high willingness to express themselves.

Willingness to react to a comment decreased when participants cannot precisely express their opinions or make a meaningful contribution.

We noticed a number of factors that decreased people's willingness to react to a comment. For some interviewees, moderate or mixed evaluation of a comment did not pass their thresholds to up/downvote a comment. An interviewee explicitly said he would not react if it is difficult to accurately articulate his opinion.

[1]Naver is the largest news platform in South Korea with an active comment section. https://news.naver.com

We also observed that people's willingness to react decreased when their perceived magnitude of contribution through participation is low. This was especially true when up/downvotes accumulated to sufficient amounts. Some interviewees said that the marginal impact their inputs could have on others mattered in their decision to react to a comment. This was especially true for comments with many votes, more so when votes were skewed to one side. They felt that even if they downvoted a comment to express their disagreement, it would not change the majority's positive perception of a comment that already has many upvotes.

Users assume that up/downvotes represent reactors' stance about the topic in relation to the comment.

Interviewees placed importance on different aspects of a comment, yet such variance across people's value judgments was indistinguishable when aggregated into up/downvotes. When interviewees were asked to share their thoughts on why some comments got up/downvotes, 7 out of 10 participants said that people who agreed or disagreed with the comment left up/downvotes. Moreover, despite the complexity of people's reasons for reacting to comments, none of our interviewees considered the possibility of mixed evaluation to a comment. These observations imply that up/downvotes fail to deliver reactors' nuanced opinions to readers. We also saw some users expressing greater contempt towards the outgroup as a result. When an interviewee saw many likes attached to contentious comments contradicting his opinion, he immediately called out, "Anyone who liked this comment is probably a misogynist."

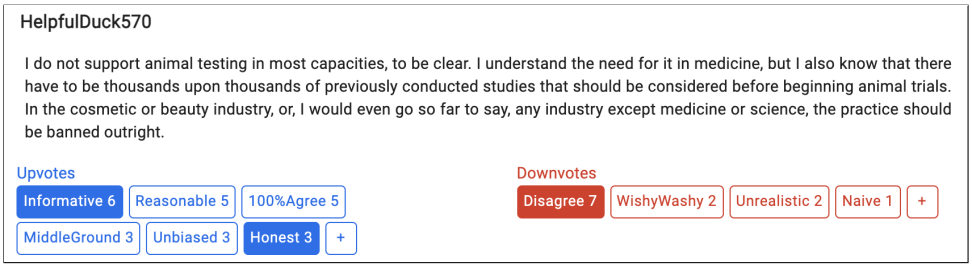


Figure 4.1: Design of User-generated labels (UGL). User-generated labels (UGLs) allow users to express their reactions to a comment by either creating their own text-based UGL or by voting on UGLs generated by other users.

4.4 User-Generated Labels

We introduce the concept of user-generated labels (UGLs) that enable users to create text-based reactions. Users can use UGLs to describe their thoughts on a comment in short text and vote on UGLs created by others. We suggest UGLs as an alternative reaction design that captures diverse and precise reactions to a comment in a light-weighted interaction. In this section, we explain how users can interact with UGLs in an online discussion setting and then discuss how UGLs solve problems identified in our formative study.

4.4.1 Reaction through UGLs

To each comment, a user can add descriptive UGLs or click on existing ones. As in Figure 4.1, UGLs are generated and displayed right below the comment.

Creating UGLs Users can add their own UGL(s) by clicking an add button (button with + mark in Figure 4.1) and writing a short text. Users self-classify each of their UGL as either a positive or negative reaction. To guide users to make short and descriptive labels, we impose character limits (20 characters) and do not allow spaces (CamelCase) for each UGL. Generated UGLs are shared with other users and also serve as a voting button. To prevent users from making duplicated labels, users' input is matched with existing UGLs through autocomplete when applicable.

Reading and voting on UGLs Below each comment, readers can read positive evaluations of a comment from the blue-colored pane on the left and negative evaluations from the red-colored pane on the right. We included this dichotomy to organize generated UGLs such that it is easier for users to read and vote on UGLs with specific sentiment (positive or negative). Indicated sentiment also helps other users to better understand UGLs that are not self-explanatory (e.g., SlipperySlope). Users can click on one or more UGLs to add their votes. For example, it is possible for the user to vote for two positive UGLs and one negative UGL on a comment as in Figure 4.1. Positive and negative evaluations of a comment from reactors are organized and ordered by the number of votes.

Managing toxic or irrelevant UGLs In the pilot study we conducted during the iterative design process, we found that some users generate irrelevant or offensive UGLs. Since UGLs serve as voting options for users and are highly visible right below the comments, such inappropriate UGLs can harm other users' experience. To limit the abuse of UGLs and control for incivility, users can right-click and flag a UGL if they find it inappropriate. When a user flags UGLs, the system asks the user to choose the reason for flagging among irrelevant, insulting, or other (explain) options. UGLs that receive more than three flags are hidden from the system, whereas the threshold can be adaptively determined for specific discussion platforms considering the number of users and the level of civility.

4.4.2 Rationale for the Design of UGLs

UGLs capture and present diverse and dynamic reactions to a comment. In our formative study, we found that people consider different dimensions of a comment and have their own ideas of what decreases or increases the value of a comment. UGLs enable dynamism and diversity that reflects such variety of people's viewpoints about a comment as well as their nuanced differences. Using text-based UGLs, users can precisely express and present their thoughts on a comment to other users with brevity. When a user has mixed evaluation on a comment, they can add multiple UGLs on each pane. Users can also specify their degree of (dis)agreement with the comment like 'StronglyAgree' or 'Acceptable'.

UGLs increase the perceived contribution of each reaction to a comment. The perceived level of contribution plays a key role when users decide to engage with comments and leave their opinions or evaluations. Enabling reactors to generate their own labels to comments augments their abilities to efficiently contribute, share, and exchange evaluative judgments about others' comments in an online discussion space. Users' intrinsic interests in receiving any degree of social approval can potentially be fulfilled by allowing their UGLs to be upvoted by other users. The more expressive nature of UGLs can increase users' ability to "do something about" the perceived media effect [164, 165]; as users observe more people expressing their evaluative opinions about a comment and exerting greater influence on others through UGLs, users' level of engagement can also increase.

UGLs save the effort required to express one's reaction to a comment. While the trade-off between the ability to capture diverse and precise reactions and the ease of engaging is inevitable, UGLs mitigate this trade off by allowing lightweight interactions to express one's opinion. While creating new

UGLs requires a little more effort, voting on UGLs provides an easy way to express one's precise reaction to comments.

4.5 Evaluation

To understand how users collectively utilize UGLs and how having UGLs affects users' experience, we conducted a between-subjects study with 218 participants. To better understand the effect of UGLs, we compared UGLs with up/downvotes. We investigate the effectiveness of UGLs with the following research questions:

RQ1*: How well do UGLs capture opinions towards comments?*

RQ2*: How does having UGLs affect user's experience in evaluating comments?*

Then we expand our investigation by exploring the secondary effect of UGLs on users' perception of the public opinion and outgroups:

RQ3*: Do UGLs allow users to better understand the multifacetedness of public evaluation of a comment?*

RQ4*: How do UGLs affect users' tolerance to the opinions that do not align with theirs?*

4.5.1 Study Design

We conducted a between-subjects study with two conditions: Binary and UGL. In both conditions, participants were given a topic statement and six initial comments (3 supporting, 3 opposing the topic statement). Participants in the Binary condition were able to react to the comments through up/downvotes. In the UGL condition, participants were able to react to the comments through UGLs. In both conditions, participants were able to add a comment or leave a reply to existing comments. Participants could also see others' reactions and reply comments.

We ran a study with four discussion topics: 1) Banning capital punishment 2) Banning affirmative action in hiring practice 3) Banning animal testing 4) Regulating tech companies' use of consumer data. These topics were chosen based on the pre-study survey that we ran on Amazon Mechanical Turk (AMT). We presented 13 different discussion topics and asked participants to share their stance, relevance, and importance (in a 7-point Likert scale) and their opinions (in a short paragraph) on each topic. We collected responses from 30 respondents and chose four topics with evenly distributed opinions. For each topic, we selected 6 open-ended responses (3 supporting, 3 opposing) and used them as initial comments so that users who joined the system in the early stage can still have some comments to read and react to.

Participants

We recruited 218 participants from AMT and randomly assigned them into one of eight groups (4 topics, 2 experimental conditions). The number of participants in each group varied from 24 to 31. We had 109 participants on the Binary and UGL condition each. Participants were paid $4 for their participation in a one-time, 30-minutes long session. The average age of participants was 40.7 (SD: 11.2). We had 139 participants with university education, and 21 of them had post-graduate education.

Tasks and Procedure

Before the main activity, participants were asked to answer the pre-survey that asked how often they participate in an online discussion (reading, writing, or reacting to a comment). In addition to their stance, relevance, importance, and willingness to express on the given topic issue, we also asked questions on participants' level of tolerance for people with opposite stance in 7-point Likert scale questions.

At the start of the main activity, we explained to participants that they can comment, reply, or react to a comment as in a common discussion platform. Participants then entered a system where they were presented with six initial comments. They were also able to see earlier participants' comments, replies, and reactions. To replicate real-life comment sections and examine how users' behavior changes as reactions accumulate over time, participants' reactions were saved and were shown to subsequent participants. While participants were told that they could read and participate in the online discussion (comment, reply, or react) as long as they wanted, they had to wait at least two minutes until they could move on to the post-survey link. We designed the time constraint to get participants engaged in the discussion, not necessarily forcing them leave comments, replies, or reactions.

In the post-survey, we asked questions on participants' experience in reacting to a comment or seeing other participants' reactions. Then participants were asked to explain what other people's reasons for up/downvoting a comment would be. We also asked questions on usability, perceived multifacetedness of users' opinion on a comment, and tolerance to people with opposite stance.

4.5.2 Measures

We used multiple measures that operationalize variables of interest with regards to each RQ. Below we summarize measures that we used to answer each RQ.

To answer **RQ1** we analyzed (1) the diversity in UGLs, (2) the number of reactions, and (3) the number of comments that each participant reacted to. To measure the diversity in UGLs, we first generated the list of unique UGLs by manually merging labels with the same meaning (e.g., 'Misinformation' and 'WrongInformation') into one. Then we categorized unique UGLs based on the aspects each UGL focuses on. To establish the set of categories, one researcher conducted an open coding on half of UGLs and another researcher reviewed the categories. Table 4.1 shows the established set of categories. Then the two researchers independently coded every UGLs, and they compared and discussed to resolve conflicts. The inter-rater reliability was 0.71 (Cohen's κ, with SE: 0.04). To see how UGLs affect the number of reactions, we compared the number of reactions left on the six initial comments, which were presented to all participants in both conditions. Lastly, we counted the number of comments that each participant reacted to among the six initial comments.

Table 4.1: The categories of UGLs with descriptions, examples, the numbers of UGLs generated, and the number of votes.

Category	Description	Examples	# UGLs generated	# Votes (% of total)
General	General labels with positive or negative sentiment	Superb, Excellent, Nah	10	52 (3.4%)
Agreement	Level of agreement or acceptance on a comment	Agree, 50%Agreed, IAcceptThis, KindaDisagee	111	517 (34.2%)
Argument	Strength of weakness of argument made in a comment	Logical, Pragmatic, Misinformation, RashAssuption	172	693 (45.8%)
Style	Writing styles or tone of a comment	TooEmotional, Vague, Forthright, HardtoRead	29	68 (4.5%)
Commenter	Judgement or impression on a commenter	WillingToListenm, Honest, Dogmatic, Compassionate	28	70 (4.6%)
Feeling	What reactors emotionally feel from a comment	Hopeful, Confused, Obnoxious	7	8 (0.5%)
Belief	Topic-specific belief or opinion	MutualBenefit, Equality,TooMuchTaxMoney	37	105 (6.9%)

Regarding **RQ2**, we distributed 7-point Likert scale questions on their experience in leaving reaction and seeing others' reactions. For both conditions, we asked about the perceived precision of reaction and perceived uniqueness of contribution that they make through reaction. In the UGL condition, we distinguished the experience of generating a UGL from voting on others' UGLs. We also asked the perceived level of understanding of others' reactions and whether their evaluation of comments was affected by reactions from others. We then asked how mentally and physically demanding it was to react with UGLs and up/downvotes.

To see if UGLs help users better understand the multifacetedness of public evaluation (**RQ3**), we analyzed the open-ended response in which participants explained other people's reasons for up/downvoting. In our analysis, we measured the number of reasons that participants can think of as reasons behind up/downvotes and the proportion of participants who mentioned reasons other than a simple agreement. Two external raters counted the number of reasons together for the first 5% of the data, and then each external rater analyzed each half of the data separately. The same two external raters coded each response individually (first 5% of the data was coded together to build a consensus among raters), and the inter-rater reliability was 0.76 (Cohen's κ, with SE: 0.03). We also measured how they were able to see users' diverse reasons for up/downvoting a comment on a 7-point Likert scale.

Regarding **RQ4**, we asked 7-point Likert scale questions on participants' tolerance towards opinions that do not align with theirs in the pre- and post-survey. Specifically, we asked their willingness to listen to, learn from, accept the opinion of commenters with opposing stances, and join a face-to-face conversation with them [166, 167]. We also asked participants' tolerance for reactors who have sentiment towards comments that conflict with their own. We asked how much they are willing to listen others' reason for up/down voting a comment and how likely they would find these reasons justifiable. We also measured the number of positive reactions each participant left to comments with opposite stance.

4.5.3 Result

Overall, participants showed a moderate level of topic relevance (M: 4.55/7.00, SD: 1.94), importance (M: 4.69, SD: 1.79), and willingness to express their opinions in an online space (M: 4.05, SD: 2.06). Participants had the highest topic relevance for the consumer data topic (M: 5.50, SD:1.50) and lowest for the capital punishment topic (M: 3.60, SD: 2.20). However, there was no difference between Binary and UGL conditions for all four topics.

There were 109 participants for each condition. Participants generated a total of 54, 56 comments and 224, 265 replies in Binary and UGL conditions, respectively. Participants in the Binary condition made 1,027 up/down votes while UGL participants generated 394 labels and made 1,630 votes (including those votes on one's own UGLs). Numbers for each condition are reported in Table 4.2. Figure 4.2 and Figure 4.3 shows subset of generated UGLs for discussion on affirmative action and animal testing.

There were 13 flags in total on the 11 UGLs. Eight flags were on the irrelevance of UGLS (e.g., TooMuchTaxMoney), and four were marked as insulting (e.g., Dumb). There were no UGLs hidden for getting three or more flags.

RQ1: How well do UGLs capture opinions towards comments?

Diversity in generated UGLs UGL participants generated 394 UGLs and there were 234 unique UGLs (109 positive, 125 negative). Among 394 UGLs, 'Agree' was the most frequently generated UGLs (34 times, 170 votes), followed by 'Disagree' (29 times, 96 votes) and 'Logical' (24 times, 92 votes).

Table 4.2: Number of participants, comments, replies, UGLs, and votes generated by topic and condition.

Topic	Condition	# participants	# comments	# replies	# UGLs	# unique UGLs	# votes
Capital punishment	Binary	28	16	65	-	-	247
	UGL	29	17	58	85	49	369
Affirmative action	Binary	24	13	53	-	-	241
	UGL	25	10	40	69	50	299
Animal testing	Binary	26	16	47	-	-	249
	UGL	28	15	82	142	92	488
Consumer data	Binary	31	9	59	-	-	290
	UGL	27	14	85	98	72	474
Total	Binary	109	54	224	-	-	1027
	UGL	109	56	265	394	234	1630

Participants generated UGLs on their degree of agreement, strength of the argument, style of the comment, judgments on the commenter, and feelings or beliefs related to the topic. Table 4.1 shows the established categories of UGLs with description, examples, and the number of each case, and its vote counts.

Number of reactions For the six initial comments, on average, UGL participants generated 1.44 (SD: 1.90) UGLs per person and clicked on 5.51 (SD: 3.05) UGLs made by others. On the other hand, participants in the Binary condition made 4.38 (SD:1.57) up/downvotes. The difference between the number of up/down votes and the number of votes on UGLs was significant (Mann-Whitney (MW) test, Z=3.05 with p<0.005). Figure 4.4 illustrates the average number of up/downvotes (Binary) in comparison with UGLs (UGL) on six initial comments for each topic. There was no interaction effect between condition and topic.

However, the number of reactions in UGL condition highly depends on the number of UGLs generated at the moment of reaction. Figure 4.5 shows how the cumulative average number of up/downvotes (Binary), UGLs generated, and votes on UGLs (UGL) (on six initial comments) change with the accumulated number of participants for almost all topics. In the early stage when there are only a small number of UGLs that participants can vote on, the number of reactions in the Binary condition is higher than the UGL condition. Once a certain number of UGLs accumulated, UGL participants begins to leave more reactions than Binary participants. The exception was the consumer data topic in which the first UGL participants created 10 UGLs, leaving subsequent participants with more options of UGLs to vote on even in the early stage.

Number of comments that each participant reacted to UGL participants reacted to more comments (among six initial comments) than Binary participants. UGL participants reacted to 5.12 (SD: 1.40) comments while Binary participants reacted to 4.39 (SD: 1.57) comments on average. The difference was statistically significant (MW test, Z=3.81 with p<0.0005).

Mixed evaluation captured in UGLs Out of 109 participants in the UGL condition, 14 participants left both positive and negative reactions to a comment. For example, in response to the comment "Affirmative action is necessary for a country as racist as ours is," one user labeled "Agree" and "NotThorough", which are two labels of the opposite sentiment. These 14 participants left mixed

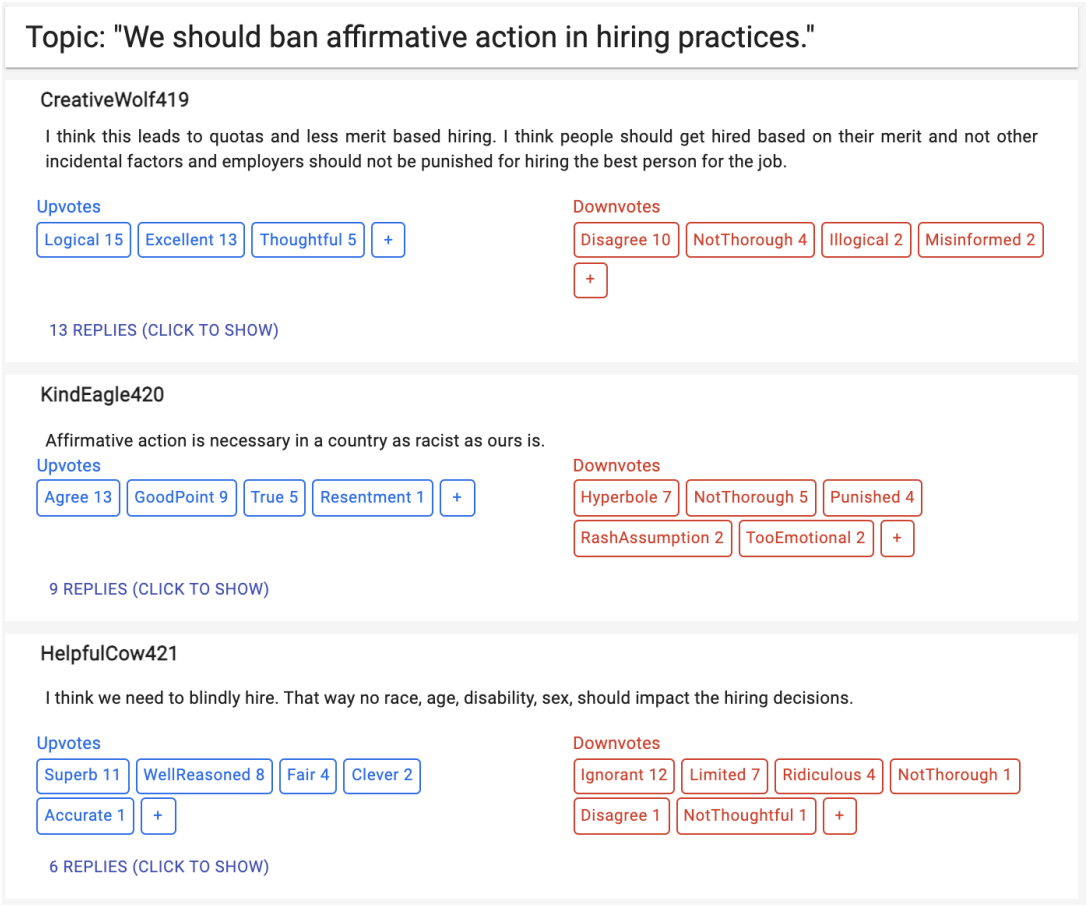


Figure 4.2: Interface with UGLs for the affirmative action discussion.

evaluation to 21 comments total.

RQ2: How does having UGLs affect users' experience in evaluating comments?

In terms of the perceived accuracy and uniqueness, participants found UGLs more satisfactory. When asked how accurately they could express their thoughts (i.e., perceived accuracy), UGL participants scored 5.24 (SD:1.35) for generating and 5.02 (SD:1.48) for voting on UGLs, while Binary participants scored 4.56 (SD:1.60). The pairwise difference between the three types of reaction was all statistically significant (MW test, Z=3.27 with p<0.001 for generating UGL vs. Binary, Z=2.01 with p<0.05 for voting on UGLs vs. Binary. Wilcoxon signed-rank (WS) test for generating UGLs vs. voting on UGLs with W=622 with p<0.05).

Participants felt that they made unique contributions when generating UGLs (M:5.24, SD:1.40) than voting for UGLs (M:4.13, SD:1.72) or casting up/downvotes (M:4.00, SD:1.77). The difference between generating UGLs and voting on UGLs or up/downvoting was statistically significant (MW test with Z=5.36 with p<0.0001 for generating UGL vs. Binary, WS test with W=278.5 with p<0.0001 for generating UGLs vs. voting on UGLs). One UGL participant noted, *"I like that text based reactions allow me to completely show my opinion rather than just going along or against others."*. There was no difference between voting on UGLs and up/downvoting.

When asked how well they could interpret other users' opinions about a comment, UGL participants scored higher (M:4.7, SD:1.5) than Binary participants (M:3.8, SD:1.70) (MW test, Z=3.75 with

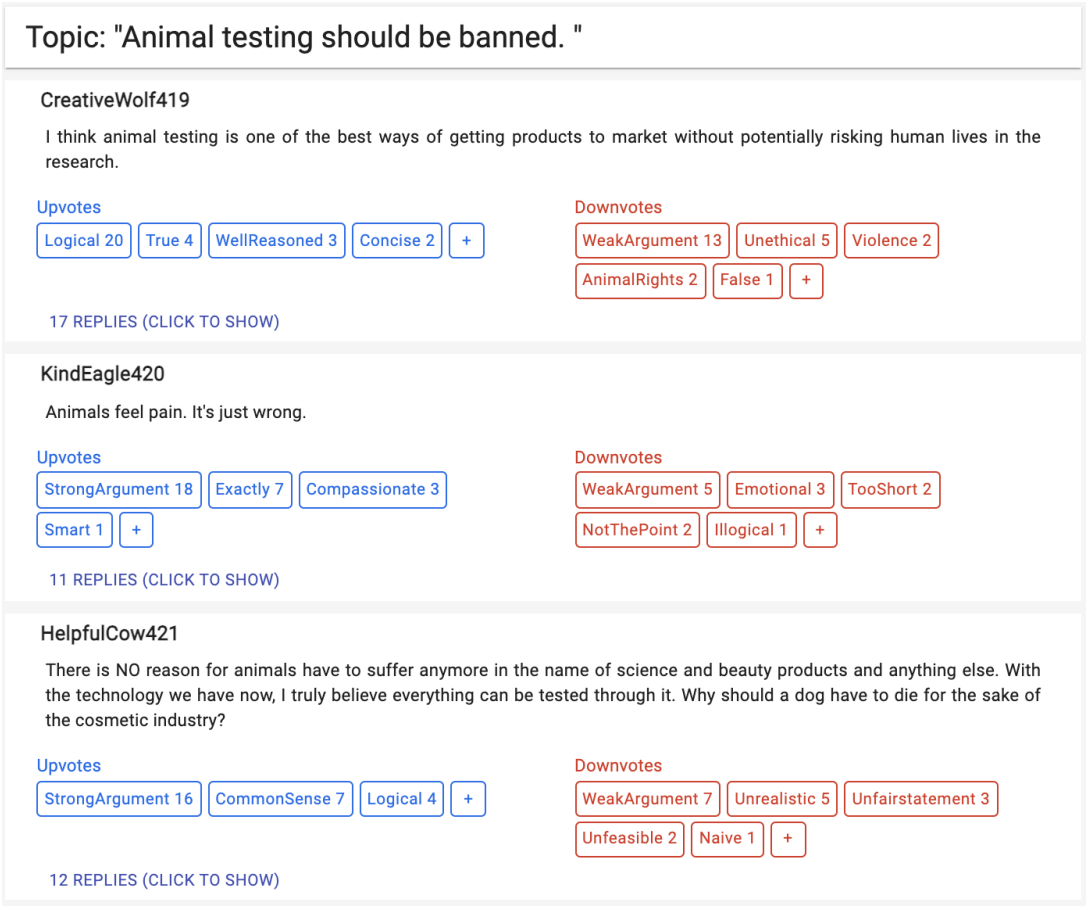


Figure 4.3: Interface with UGLs for the animal testing discussion.

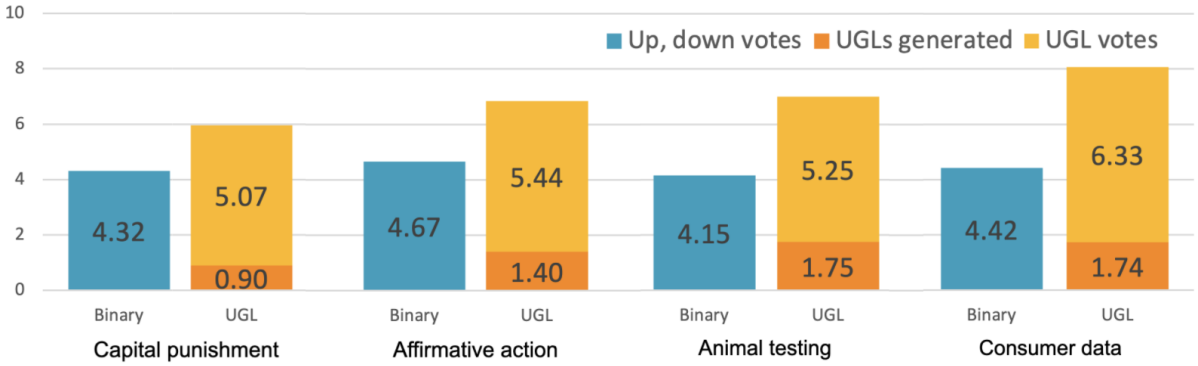


Figure 4.4: The average number of up/downvotes (Binary) and generated/voted UGLs (UGL) on the six initial comments.

p<0.0005). Moreover, when asked whether others' reactions influenced their evaluations of a comment, UGL participants reported a higher score (M:4.56, SD:1.61) than Binary participants (M:3.72, SD:1.70) (MW test, Z=3.58 with p<0.0005). *"I like seeing other's perspective and thinking about them to see if I am wrong in my thinking."* said one participant in UGL condition.

Participants reported higher mental and physical demand for generating UGLs than voting on UGLs or up/downvoting. When asked about mental and physical load needed for each type of reaction, UGL participants rated 3.42 (SD:1.73) and 2.46 (SD: 1.74) for generating UGLs, and 2.10 (SD:1.58) and

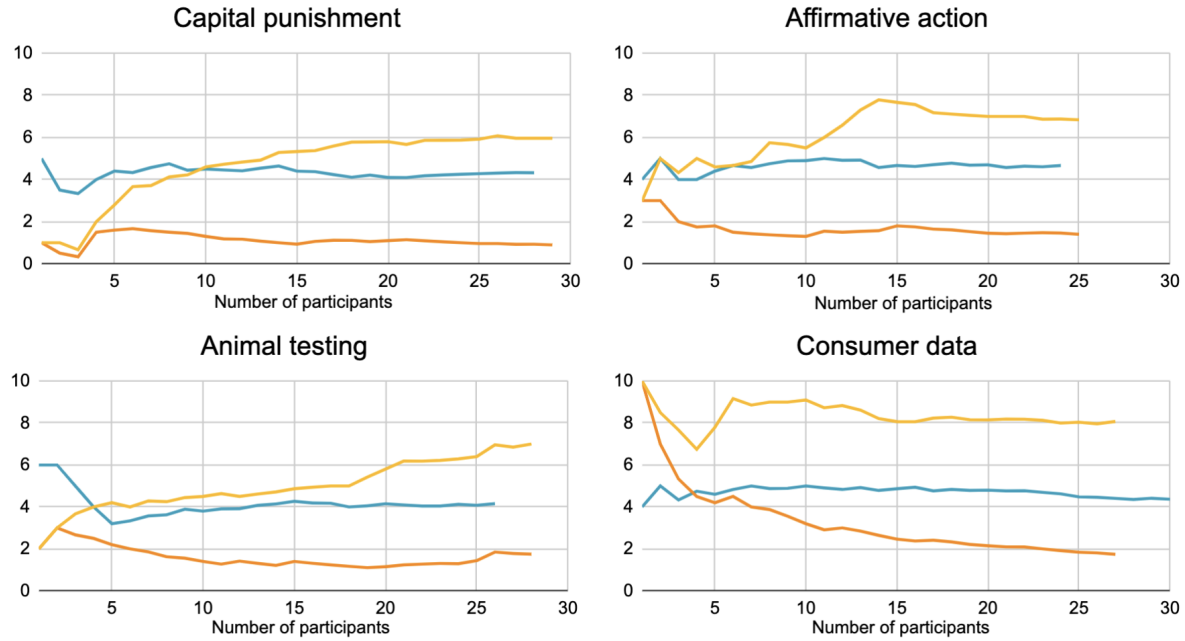


Figure 4.5: The cumulative average number of up/down votes (Binary), created UGLs (UGL), and votes on UGLs (UGL) by the accumulated number of participants for each topic.

1.91 (SD:1.66) for voting on UGLs. Binary participants rated 2.09 (SD:1.64) and 1.74 (SD:1.49) for mental and physical demand, respectively. For both questions, the difference between generating UGLs and voting on UGLs (WS test, W=97 and W=187 with p<0.0001, for mental and physical demands) or up/downvoting (MW test, Z=6.10 and Z=4.07 with p<0.0001, for mental and physical demands) was statistically significant while there was no difference between voting on UGLs and up/downvoting. However, many participants described UGLs as an easy way to precisely express one's thought. *"Text based reactions seem like a very simple way to still get what you want to say about the comment. ⋯ I love it, a one worded comment that can still be taken as a vote."*

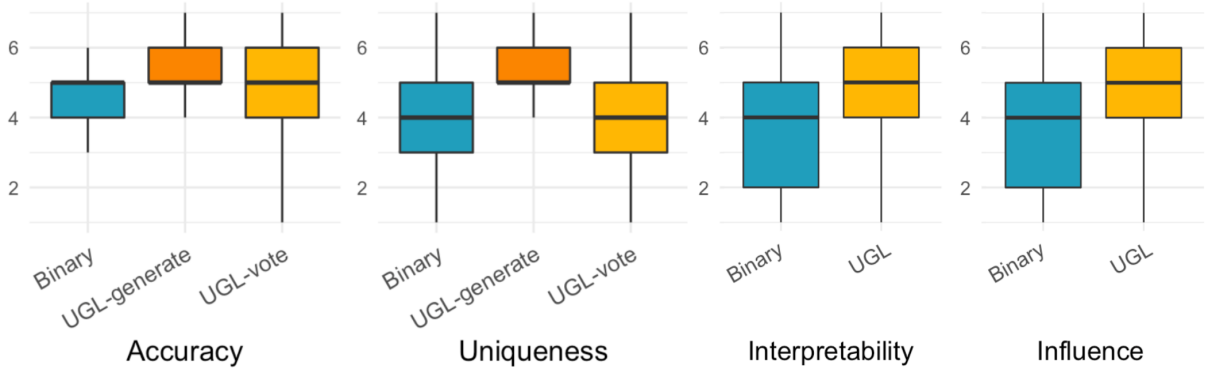


Figure 4.6: Perceived accuracy and uniqueness of own reaction and interpretability and influence of others' reactions, for each reaction type

RQ3: Do UGLs allow users to better understand the multifacetedness of public evaluation of a comment?

When asked to list possible reasons behind up/downvotes, UGL participants listed 1.51 (SD:0.71) reasons for upvotes and 1.71 (SD:1.00) reasons for downvotes. Conversely, Binary participants mentioned 1.30 (SD:0.58) reasons for upvotes and 1.35 (SD:0.64) reasons for downvotes. The differences was significant (MW test, Z=-2.04 with p<0.01 for upvotes and Z=-2.34 with p<0.05 for downvotes).

UGL participants better understood the multifaceted aspects of up/downvoting, as 48.6% and 51.3% of them mentioned reasons behind upvotes and downvotes other than simple agreement or disagreement with a parent comment. On the other hand, only 29.4% and 33.9% of Binary participants came up with reasons behind up or downvotes other than agreement or disagreement. The differences were significant (χ^2=8.50 with p<0.005 for upvotes and χ^2=6.77 with p<0.01 for downvotes).

UGL participants were better able to find diverse rationale for others' upvotes and downvotes (M:5.82, SD:1.05 for upvotes and M:5.62, SD:1.22 for downvotes) than Binary participants (M:4.99, SD:1.61 for upvotes and M:4.72, SD:1.67 for downvotes). The differences were significant (MW test, Z=4.08 for upvotes and Z=4.08 for downvotes, both with p<0.0001). UGL participants liked how diverse reasons for up/down voting a comment are expressed through UGLs. One UGL participants said, *"It's nice to see why another person reacted to a comment. For instance, was it because the commenter was rude or was it that they were misleading or was there just a disagreement on the entire subject premise."*

RQ4: How do UGLs affect participants' tolerance to the opinions that do not align with theirs?

In both pre- and post-surveys, participants expressed moderate tolerance towards people with opposite stance on each topic. For both UGL (M: 4.54, SD: 1.44 for pre-survey and M:4.53, SD: 1.48 for post-survey) and Binary (M: 4.76, SD: 1.30 for pre-survey and M:4.74, SD: 1.45 for post-survey) conditions, participants' level of tolerance did not significantly change after the main activity. Likewise, participants in both conditions showed moderate tolerance towards others' reactions that do not align with theirs in both pre-and post-survey. For both UGL (M: 4.70, SD: 1.22 and M:4.64, SD: 1.21 for pre- and post-survey) and Binary (M: 4.84, SD: 1.15 and M:4.77, SD: 1.22 for pre- and post-survey) conditions, there was no significant difference between participants' tolerance level before and after the main activity.

Although there was no significant change in the tolerance towards people with opposite stances on the topic, UGL participants left more positive reactions to comments with opposite stances from theirs. Among three initial comments with opposite stances from the participant, UGL participants left positive reactions to 1.23 (SD: 1.55) comments on average. On the other hand, Binary participants left positive reactions to 0.58 (SD: 0.87) out of three comments that had opposite stances on the topic. The difference was statistically significant (MW test, Z=2.74 with p<0.05). One participant in UGL condition also noted that *"I like that there is reasoning behind the upvoting and downvoting process rather than just a simple 'I like it or I don't'. There are things that you can disagree with but are very well written and logical, and it's great to be able to convey that. ⋯"*

4.6 Discussion

Our evaluation showed that UGLs help users better express their thoughts and understand the multifacetedness of people's reaction to a comment. In this section, we revisit our findings and refine the design implications of our current study on 1) the information rich-reach trade-off of reactions in the comment section and 2) the effect of capturing diverse reactions through UGLs on tolerance towards the outgroup.

4.6.1 Information Rich-Reach Trade-off.

While generating new UGLs requires more mental and physical effort, UGLs mitigate the information-rich-reach trade-off in a comment section. UGLs are information-rich, reflecting nuanced differences across people's viewpoints about a comment. Replacing up/downvotes with UGLs did not reduce the overall level of clicks. In fact, the number of votes with UGLs eventually surpassed the number of up/downvotes across all four topic conditions.

Our findings imply that facilitating open evaluation is key to active participation in the UGL condition. UGLs enabled participants with moderate or mixed opinions to express their opinions without projecting them onto up/downvotes. Moreover, participants leaving reactions to comments felt that they were able to contribute more through UGLs. This aligns with previous studies that report how the perceived uniqueness of individual contributions may increase user participation [168, 154, 155].

As we observed in the study, however, the success of UGLs depends on how early users engage with the system. The benefits of having UGLs become pronounced when there exists a certain number of UGLs that users can engage with. This observation is in line with previous work on how initial contributions (e.g., first comment [169] or first review [170]) govern the success of an online community. Systemic support that can reduce early users' burden of generating UGLs, e.g., suggesting a list of potential UGLs that users can refer to, can be introduced to accelerate the generation of UGLs.

4.6.2 The Effect of Showing Diverse Reactions through UGLs.

We found that presenting the rich context of user reactions to comments affects how readers understand and evaluate each comment. In response to the open-ended question about their experience with UGLs, participants noted that UGLs helped them evaluate a comment more thoroughly. For example, one user mentioned that she re-considered a comment that she had originally thought was good because she found a UGL pointing its inaccuracy. We found that this mechanism had a positive effect in encouraging more deliberate participation across users.

The effect of diversity on reducing affective polarization has gained scholarly attention [171, 45, 161, 30, 172, 160, 173]. While aggregated UGLs improved participants' understanding of others' multifaceted evaluation of a comment, our findings show that simply increasing exposure to diversity does not increase the tolerance towards the outgroup. Participants in our study noted that seeing other's UGL that aligns with their opinion increased their confidence. It could be that, others' endorsement of labels they agree with 1) reinforced their attitudes and beliefs and 2) offset the positive effect of exposure to diverse opinions from the other side. Recent research reports similar findings from exploring the effect of seeing diverse opinions on social issues [45]. Nevertheless, participants' qualitative responses suggest that UGLs could still provide a starting point for a more productive conversation in the future. One participant pointed out that UGLs helped them understand how others' evaluations of a comment are different from

their own. This implies that UGLs can mitigate misunderstanding and misinterpretation over aggregated up/downvotes.

4.7 Limitations and Future Work

Our current study has several limitations, which leave room for future work. In our study, we only look at interactions between readers and reactors. The complete mechanism of the comment section involves commenters, readers, and reactors, though one user can take multiple roles. Future studies could further explore whether UGLs have positive effects on subsequent comment writing behaviors of reactors and parent commenters. Increased richness in feedback from users could motivate commenters to be more deliberate and leave higher quality comments, which can lead to higher quality replies [174].

Furthermore, future work should conduct a longitudinal deployment study to re-examine the effect of UGLs. Our study design and setting (having crowd workers as participants or assigning a topic they discuss) limit the generalizability of our study findings. Also, there could have been a novelty effect of UGLs, likely leading to more active usage. A longer-term observation of how UGLs would work in real-world settings could help us validate our results.

Last but not least, we could explore the application of evaluative UGLs in a different setting other than commenting context on serious or controversial current events. For example, one could investigate how UGLs could improve user experience on community platforms focused on asking and answering questions.

4.8 Conclusion

In this project, we identified limitations of up/downvotes and proposed user-generated labels (UGLs) as an alternative reaction mechanism that captures diverse and precise reactions to a comment. UGLs leverage social-psychological incentives and enables the production of a more nuanced, rich picture of a comment's value to other users. Our evaluation results demonstrated that users were more expressive and left more reactions with UGLs than with up/downvotes. We also show that UGLs help users interpret others' reactions and better understand the multifacetedness of people's reactions to comments. We anticipate that our design of UGL and study findings can guide and inspire the future design of reactions to better capture and deliver users' thoughts on comments.

Chapter 5. PolicyScope: Supporting Citizens' Exploration of Diverse Policy Effects

This chapter presents PolicyScope, an interactive system that helps citizens understand how policy impacts vary across different stakeholders and conditions by making contingency explicit and explorable. This chapter focuses on contingency—how policy outcomes differ depending on specific conditions and contexts. Through a user study, I demonstrate how PolicyScope's interactive canvas and structured condition exploration enable users to grasp the conditional nature of policy effects, moving beyond simplified universal claims toward nuanced understanding of how policies affect different groups under different circumstances. All uses of "we", "our", and "us" in this chapter refer to collaborators who contributed to this work.

5.1 Introduction

Policies are decisions and actions made by governments or organizations to address public issues, which are often implemented through laws, regulations, or programs. In democratic societies, it is crucial for citizens to understand public policy in order to make informed decisions and effectively engage in civic actions such as voting, public discourse, or signing petitions. Policies, however, are not very straightforward. They affect different stakeholders in varied ways, and their actual impacts often depend on how they are implemented, how people and organizations behave, and various external factors (e.g., economic shifts or technological advancement). This inherent complexity makes it challenging for citizens to fully grasp policy outcomes, and as a result, their understanding often remains oversimplified.

What often limits policy understanding is not just the complexity of policies, but the way people approach and reason about them. People often seek information that confirms what they already believe, flatten uncertain or condition-dependent policy outcomes into definitive expectations, and struggle to consider how others with different experiences might be affected. These tendencies make it difficult to understand who is affected by a policy, how, and under what conditions, leading to shallow or fragmented public understanding.

Beyond these cognitive tendencies, the information environments through which citizens usually encounter policies contribute to oversimplification. News articles, campaign materials, and online commentaries often emphasize a small set of familiar storylines while leaving out many of the underlying mechanisms and conditional factors. As a result, people may form impressions based on surface-level cues or prior beliefs rather than through a structured examination of who is affected and under what circumstances. Even when citizens try to explore further, the effort required to filter, interpret, and connect scattered information can make deeper reasoning difficult.

Existing tools for public policy communication provide limited support for this type of reasoning. Deliberation platforms and data dashboards allow users to browse arguments or visualize selected outcomes, but they usually encode predefined framings such as fixed stakeholder groups, specific indicators, or expert-written scenarios. These tools offer ways to navigate what has been curated for them, but give users little flexibility to define which aspects of a policy matter for their own questions, concerns, and values.

In this work, we explore how interactive systems can support a more constructive and user-driven

understanding of policy impacts. Rather than presenting a single account of a policy, we examine how a system can help people identify relevant stakeholders, consider conditional factors, and explore possible impacts in ways that reflect what they find meaningful. To do this, we introduce PolicyScope, an interactive system that uses large language models to generate candidate stakeholders, conditions, and impact scenarios. Users can reorganize, refine, or ignore these suggestions as they explore a policy. The goal is to reduce the cognitive burden of identifying relevant elements while preserving the user's agency in constructing their own interpretation.

This work makes the following contributions:

- Design of PolicyScope that supports user-driven exploration of policy impacts
- A technical pipeline that identifies candidate stakeholders, infers policy-relevant conditions, and produces user friendly impact descriptions for different stakeholder–condition combinations
- A comparative evaluation result that illustrates how PolicyScope helps users consider a wider range of stakeholders and conditions, form more grounded arguments, and experience lower cognitive burden during exploration.

5.2 Background and Related Work

5.2.1 Challenges in Understanding Policy Impacts

Public policies shape many aspects of citizens' everyday lives and decisions, making it important for people to form a grounded understanding of what policies propose and whom they affect. However, understanding a policy is not as simple as reading its written plans and terms. Policies often affect diverse groups of people in different ways, and their meaning can shift depending on one's values, priorities, and lived experiences [175, 17, 176]. Furthermore, policy impacts are highly conditional. The same policy can lead to very different outcomes depending on external conditions (e.g., macroeconomic trends, demographic change, environmental shocks) and the behaviors of internal actors such as implementing agencies or local governments [177, 178, 179].

Yet widely used communication formats—such as press releases, campaign platforms, and policy briefs—often fail to convey this multifaceted structure. These materials commonly present simplified narratives or headline figures that obscure underlying mechanisms, indirect effects, or distributional consequences. Prior work in public policy and political communication shows that citizens often rely on such simplified representations when forming opinions, even though these representations may omit key contingencies or mask who benefits and who bears the costs [180, 181]. As a result, people may form impressions of policies that do not fully reflect the actual trade-offs or the complexity of real-world implementation. In this work, we design a system that helps citizens understand diverse policy outcomes by examining who is affected by a policy and how these effects may change under different conditions.

5.2.2 Interactive Tools for Policy Communication and Participation

A growing body of work in HCI and civic technology has explored interactive tools that support public engagement with policy issues. Online deliberation platforms such as ConsiderIt [182] and Pol.is [183][1] enable citizens to express positions and examine how others cluster around different arguments, revealing areas of consensus and disagreement. Other systems focus on civic participation and co-creation,

[1] https://pol.is/

supporting activities such as participatory budgeting, collaborative drafting of policy proposals, or civic crowdfunding of local projects, often with interfaces that visualize options, trade-offs, or geographic distributions of benefits [184].

Recent work also investigates systems that help laypeople and policy professionals make sense of complex policy materials. Several studies use generative AI to make policy texts more accessible. Safaei et al. [185] generated policy briefings from short prompts using a fine-tuned GPT-2 model, and Yun et al. [186] used large language models to translate legislative language into plain explanations and customized summaries. Other efforts integrate AI into policy-analysis workflows, such as Wang et al. [187], who synthesized citizens' lived experiences from online discussions to support memo writing. Complementing these text-oriented tools, visualization systems like LegiScout [188] and scenario dashboards [189] help users explore relationships among policies or examine potential impacts. Together, these approaches show a growing interest in combining AI and visualization to support sensemaking in complex policy domains.

While these systems broaden access to policy information, they usually encode a fixed framing of what matters, such as predefined outcome indicators, argument categories, or stakeholder groups. Users can navigate and respond to these representations but have limited agency to define what counts as a relevant stakeholder or meaningful scenario for their own priorities. In this work, we explore a different direction by designing a system that helps people actively examine policy impacts through their own exploratory process, constructing interpretations that reflect what they find meaningful rather than relying only on expert-defined framings.

5.2.3 Systems that Support Public Sensemaking of Complex Information

Beyond the policy domain, HCI researchers have developed systems that help non-experts make sense of complex information in areas such as science, AI, and online discourse. Dialogue-based intelligent tutoring systems like AutoTutor guide learners through conceptual domains such as physics and computer literacy by engaging them in natural-language conversations and prompting them to articulate explanations, rather than passively consuming content [190]. Similarly, interactive simulations such as PhET enable students to explore scientific phenomena through scientist-like experimentation, making invisible processes visible through dynamic visual representations [191]. Learning analytics and simulation work has explored how to surface rich data in forms that learners can interpret. For instance, Martinez-Maldonado et al. proposed layered storytelling techniques that present multimodal traces from nursing simulations in increasingly interpretable layers, helping students reflect on teamwork, communication, and patient care in complex scenarios [192].

More general sensemaking support systems provide interactive workspaces and visual structures to help users externalize, organize, and revisit evolving interpretations over time [193, 194]. Recent systems also support sensemaking about AI and algorithmic behavior for lay users. Bhat and Long introduced interactive explainable AI tools that allow adults to explore how classification systems behave in concrete scenarios, with the goal of improving AI literacy and encouraging reflection on ethical implications [195]. Other work has examined interfaces that help people reason about contentious public issues by exposing multiple perspectives and encouraging reflection through structured prompts, visual overviews, or guided comparison of arguments [196]. For example, ConsiderIt frames online deliberation around pro/con points that participants create and share, surfacing salient arguments while enabling users to explore key perspectives from different groups [44].

Across these systems, a common pattern is to move beyond static explanations toward interactive

structures that invite users to probe, compare, and reorganize information. Rather than treating the system as an oracle that provides the answer, these designs aim to scaffold users' own sensemaking processes. However, most focus on specific domains (e.g., STEM learning, AI literacy, or online discussions) and do not directly address the particular challenges of policy impacts, such as multi-stakeholder distribution and strong dependence on contextual conditions.

5.2.4 Constructivist Foundations of Interactive Sensemaking Systems

Many of the systems described above draw, implicitly or explicitly, on constructivist learning theory, which views learning as an active process in which people construct knowledge and meaning based on their prior experiences, beliefs, and social context [197, 198]. Rather than treating learners as passive recipients of information, constructivist perspectives position them as active agents who build understanding through exploration, reflection, and social interaction.

Constructivist principles have informed the design of interactive systems across diverse domains beyond those discussed earlier. In programming education, environments like Scratch embody Papert's constructionism by enabling learners to construct knowledge through creating personally meaningful artifacts—games, animations, and interactive stories—rather than following prescriptive tutorials [199]. These block-based programming environments provide manipulable building blocks that users combine in different ways, supporting iterative experimentation and learning from failure. In data analysis, interactive machine learning systems apply constructivist principles by allowing users to build their own understanding of model behavior through hands-on manipulation and immediate visual feedback, making the model-building process itself a site for learning [200].

Beyond formal learning contexts, constructivist ideas have inspired systems that support meaning-making in personal and creative domains. Augmented reality tools like Virtual Chemist allow learners to explore molecular structures through interactive 3D visualization, helping them construct spatial mental models of abstract chemical concepts [201]. Other systems support personal reflection and creative expression by providing scaffolds for users to externalize, annotate, and reframe their own experiences, positioning the system as a tool for active meaning-making rather than a source of predetermined interpretations.

These projects share several design principles relevant for policy understanding. They provide manipulable building blocks that learners can combine in different ways; they encourage perspective-taking and reflection; they support iterative refinement rather than demanding correctness upfront; and they position the system as a scaffold rather than a fixed source of truth. PolicyScope builds on this tradition by applying a constructivist orientation to the domain of public policy. Instead of offering a single expert-curated explanation of a policy, it provides stakeholders, conditions, and impact scenarios as configurable elements that users can explore according to their own questions and values. This connection suggests broader opportunities for constructivist, AI-supported systems that help citizens actively construct their understanding of complex policies, rather than only consuming pre-packaged narratives.

5.3 Formative Study

To understand how people understand policy in early stage and difficulties in that, we conducted a formative study. Specifically, we took a look at how citizens interpret election pledges, which represent

the most common and intuitive form of early-stage policy information that people encounter.

5.3.1 Method

Participants We conducted a semi-structured interview with 13 participants. We collected the campaign platforms from the most recent national assembly election. These materials typically presented pledges in short, bullet-point formats, offering only brief descriptions of intended policy directions. Because pledges are often the public's first exposure to emerging policy ideas, they served as an appropriate setting to examine early-stage policy sense-making.

Procedure Participants were first asked how they usually encounter, interpret, and use information about election pledges. We also asked about difficulties they face when trying to understand a specific pledge or comparing alternatives. Then participants were guided to review pledges of the candidates and select two or three pledges they found personally relevant or interesting. For each selected pledge, they were asked to explain their perspective on the pledge, including what they saw as its strengths and weaknesses. Finally, we asked participants which candidate's pledges they considered more favorable overall and why.

5.3.2 Findings: Challenges in Understanding and Using Early-Stage Policy Information

Across the interviews and exploration activities, participants described several difficulties in understanding the pledges. From these accounts, we identified three core challenges that characterize how people struggle with simplified policy information.

Difficulty Understanding the Effects of Pledges Participants often had trouble understanding the basic meaning or starting point of a pledge. Even when a pledge mentioned a specific action, such as relaxing building height restrictions or expanding childcare support, participants were unsure what that action involved, how it would be implemented, or what kinds of changes it would bring. Many were unable to articulate even a rough mental model of how a pledge could influence real situations. They frequently asked questions such as What exactly does this do?", Would this change anything in my daily life?", and Who is impacted by this?". These reactions showed that people lacked clarity not only about effects but also about the underlying mechanisms, responsible actors, and primary beneficiaries. The absence of this foundational understanding made it difficult for participants to build further reasoning. As P9 explained, *For areas I wasn't familiar with, I couldn't even understand what kind of impact it would have on me—it didn't feel tangible."* Participants often did not know which perspectives were relevant when interpreting a pledge. They were unsure who might be affected, how impacts might differ across groups, or under what conditions the policy would matter. Many described pledges as all sounding similar" because they lacked visibility into stakeholder differences, contextual dependencies, or potential trade-offs. P6 expressed this frustration: *It sounds good, but what about the downsides? There will be some side effects that I don't know."* Without such perspectives, participants struggled to form substantive interpretations of a policy.

Desire to understand factors that might affect the policy outcome and feasibility Most participants tried to assess the feasibility of election pledges as well as the outcome of the pledges. Especially for pledges that are closely related to their lives, they wanted to form anticipated understanding of what will happen with the policy. However, they struggled to identify the conditions that would determine success or failure.

"(reconstruction pledge) I don't think one lawmaker can make it happen. What is needed to this happen? I want to know that to evaluate if this is good pledge or not." (P9)

"If long-term rentals increase, I honestly can't predict the outcome. It will depend on who moves in - the dynamics would be completely different depending on the residents' economic situations." (P12) Participants recognized that policy outcomes are contingent on circumstances: whether other stakeholders cooperate, whether economic conditions change, whether implementation follows through. But they lacked structured ways to explore these dependencies. P13 noted that feasibility varies dramatically by political context: *"Maybe 30-40 out of 100 will be implemented, but the rest depends on whether they're in the majority or opposition party, how much support they can secure."* Yet this conditional reasoning remained vague and speculative, as participants had no systematic method to examine how different conditions would shape outcomes.

Low Perceived Value in Exploring Pledges Participants frequently expressed that reviewing pledges felt uninformative or unhelpful. Some described the content as generic or overly promotional, and noted that it lacked the detail necessary to form meaningful opinions. One participant commented, "These lists of pledges are so uninformative that I gain nothing from looking at them." The absence of explicit impacts, conditions, and trade-offs reduced their motivation to engage more deeply.

Low Confidence in Their Own Understanding Many participants expressed uncertainty and low confidence in their interpretations. They worried that they were "missing something important," "not understanding the full picture," or "not qualified to judge" the implications of a pledge. This lack of confidence often led them to withdraw from further exploration or rely on superficial cues rather than analytical reasoning. This finding indicates the importance of designing systems that provide gentle scaffolding, clarify reasoning structures, and help users feel more capable of evaluating policy information.

While our formative study centered on pledges, the sensemaking challenges participants described, such as difficulty identifying affected stakeholders, understanding the conditions that shape effects, and comparing alternatives, extend beyond the election context. Because pledges often activated political identity and reduced participants' willingness to engage deeply, we shift our focus to policies. Policies involve similar cognitive demands but allow users to explore impacts with less partisan pressure. For this reason, the formative findings provide a useful foundation for designing our policy exploration system.

5.4 Design Goals

Our formative study showed that participants struggled with early-stage policy information because it was presented without clear structure or cues about how to interpret it. They did not know which perspectives to consider, how to break down a policy into meaningful parts, or how to reason about possible outcomes. Based on these observations, we defined the following design goals.

DG1. Enable Exploration Across Multiple Perspectives

Participants had difficulty identifying which perspectives were meaningful when trying to understand a pledge. Without clarity about who is affected, under what conditions the policy might work, or what trade-offs may arise, pledges often felt generic or indistinguishable. The system should help users explore a policy from multiple angles—such as different stakeholder groups or contextual situations—so they can see how impacts vary across people and circumstances. Making these perspectives explicit enables users to move beyond surface impressions and build richer interpretations.

DG2. Provide Structured and Separable Building Blocks

A recurring challenge for participants was not knowing where to begin. They were unsure who

was involved, what mechanisms were implied, or what types of outcomes to expect. To address this, the system should present clear, separable building blocks—stakeholders, conditions, and impacts—that users can assemble as they explore. This structure reduces ambiguity in the early stages of understanding and provides a concrete starting point for reasoning about potential effects.

DG3. Encourage Self-Directed Exploration

Participants often approached pledge information passively because they did not see clear entry points for exploration. When information felt generic or overly promotional, they quickly lost interest. The system should encourage users to pursue their own questions, follow lines of reasoning that matter to them, and choose which stakeholders or scenarios to examine. The goal is to support exploration that is guided by users' own motivations and concerns, rather than pushing them toward a predefined interpretive path.

DG4. Provide Exploratory Support

Even motivated users sometimes struggled with how to continue their exploration—what to examine next or how pieces of information might relate. Unlike DG3, which focuses on agency and motivation, this goal focuses on interaction-level support that helps maintain momentum. The system should offer small, non-directive cues such as previews, gentle prompts, or contextual hints that highlight possible next steps without steering users toward a single interpretation. These touches help users stay oriented in the exploration process while retaining full control over their reasoning.

5.5 PolicyScope

Based on the design goals presented in the previous section, we developed PolicyScope, a web-based interface that supports users in exploring the diverse effects of a policy. In this section, we first describe the overall interface and its core components, then walk through an example user flow, and finally discuss the design considerations that guided the interface.

5.6 System Design

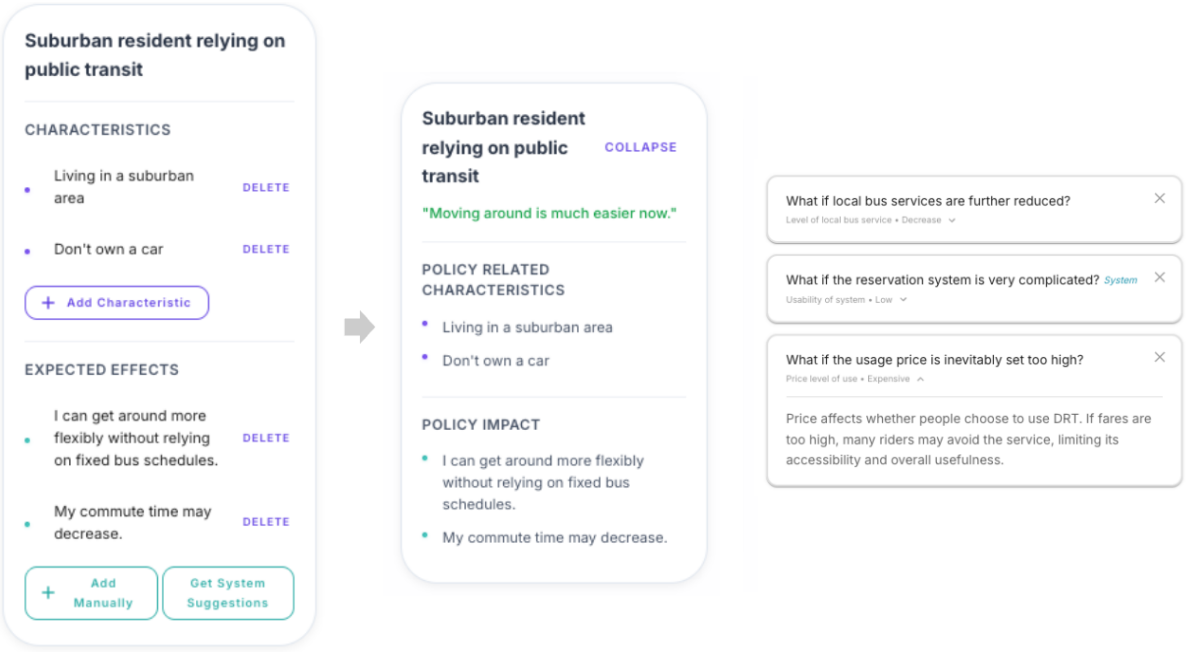
Based on the design goals described earlier, we developed PolicyScope, a web-based interface that helps users explore how a policy may affect different people under varying conditions. In this section, we describe the core components of the system—Stakeholder Cards, Condition Cards, and the Impact Canvas—and explain how users interact with them through the three-stage workflow.

5.6.1 Interface Components

Stakeholder Card

A Stakeholder Card represents a perspective relevant to the policy. Each card includes a title, a brief perspective summary, a list of personal characteristics, and expected impacts. The title names the stakeholder group or individual (e.g., "Person in their 30s living with an elderly parent"). Characteristics capture attributes that shape how the policy might affect them, such as job type, caregiving responsibilities, or living situation. The impacts section lists anticipated effects from that stakeholder's viewpoint.

Users can create Stakeholder Cards manually by entering characteristics and writing expected impacts. They may also request the system to generate more impacts through "AI suggestion" button.



(a) Stakeholder Card (b) Condition Card

Figure 5.1: Examples of the two card components in PolicyScope. (a) A Stakeholder Card with a title, characteristics, and expected impacts in its edit and view mode. (b) A Condition Card that expresses a situational factor as a what-if question with a short description.

Each request adds one system-generated impact produced through a structured prompting pipeline. If no further meaningful effects can be inferred, the system displays a fallback message. Users can freely edit or delete system-generated content. This on-demand and incremental support helps users expand their thinking without losing control of the construction process.

Condition Card

A Condition Card represents a situational factor that may influence how a policy unfolds. Each card consists of a what-if question (e.g., "What if the local government lacks administrative capacity?") and a short description of how this condition could meaningfully affect policy outcomes. Users can write their own what-if questions or select from system-generated questions.

Impact Canvas

The Impact Canvas is the main workspace for exploring how stakeholder impacts differ across conditions. It consists of the Impact Map in the center, the Stakeholder Panel at the top, and the Condition Panel on the left.

Impact Map The Impact Map positions Stakeholder Cards on a two-axis layout. The horizontal axis represents direction and scale of impact (negative to positive), and the vertical axis represents relevance to the policy. Cards can be positioned by the user or adjusted automatically by the system. The map updates dynamically as users add stakeholders, edit impacts, or apply conditions.

Stakeholder Panel The Stakeholder Panel allows users to select which stakeholders appear on the map. Users may create cards manually or get system suggestions. On clicking the system sugges-

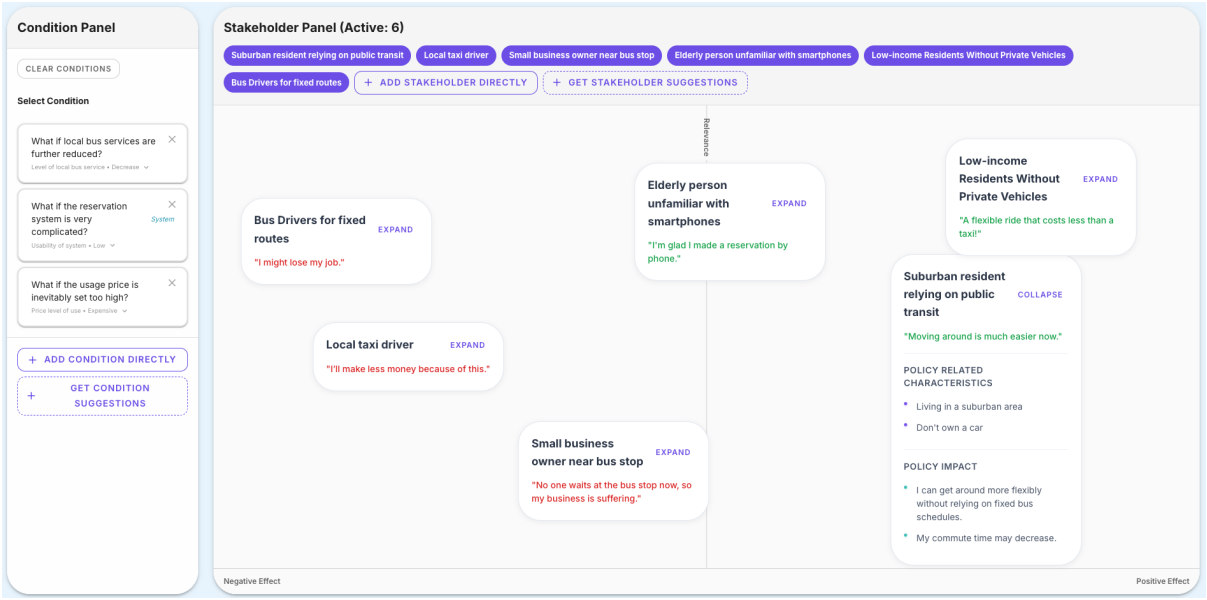


Figure 5.2: Overview of the Impact Canvas, which consists of the Impact Map, Stakeholder Panel, and Condition Panel. Selecting conditions updates the suggested impacts and positions of stakeholder cards, allowing users to examine how policy outcomes vary across conditions.

tion button, a modal with system-proposed stakeholders will be shown. Users can also request further suggestions. Created or selected stakeholders appear immediately on the Impact Map.

Condition Panel The Condition Panel displays the set of what-if questions the user has created or selected. Users may create cards manually or get system suggestions. On clicking the system suggestion button, a modal with system-proposed condition cards are shown. Users can also ask for additional suggestions in real time. This process allows users to build a set of conditions they find meaningful or worth exploring further. When a condition is selected, the system generates updates to each Stakeholder Card's impacts and suggests new positions on the Impact Map. Users can edit or override these suggestions at any point.

5.6.2 User Flow

In this subsection, we explain interaction flow that is designed to guide user to progressively expand the range of exploration, from one's own perspective to multiple perspectives and conditions. Once entering the system, users are given with a list of policies and user can start their interaction by selecting one policy they would like to explore.

Reflecting on One's Own Perspective After selecting a policy from the list, users start exploration by considering how the policy may affect them personally. The left side of the interface shows a short overview of the policy (background, method, expected effects) and the right side prompts the user to describe their personal characteristics relevant to the policy and possible impacts to them. Users enter characteristics (e.g., "My parents live in a rural area") and write expected impacts (e.g., "It may become easier to visit my parents, and they may move around more comfortably"). Based on this, the system generates a structured Stakeholder Card summarizing their input, which users can review, revise, or expand. Users may also request AI support to generate additional impacts for themselves.

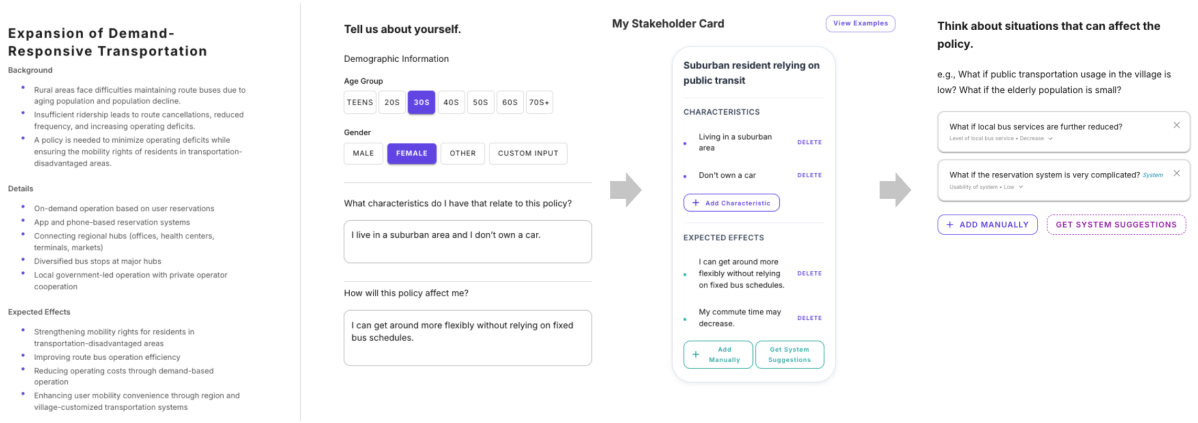


Figure 5.3: Personal framing step in which users construct their own Stakeholder Card and identify condition what-if questions prior to using the Impact Canvas.

Making What-if Questions After making their own Stakeholder Card, users are guided to consider situations that may change the policy's outcomes. The policy overview remains on the left, and the right side allows users to create Condition Cards. Users may write their own what-if questions or request system suggestions.

Users can then select from the questions they created or the ones suggested by the system, keeping only those they want to explore. Through this process, they construct a personal list of conditions that they consider relevant or worth examining further.

Exploring Impacts on the Impact Canvas Then users are guided to explore policy through the Impact Canvas. At first, Stakeholder Panel shows their own stakeholder name and the Condition Panel shows what-if questions for the conditions that the user generated in the previous step. From here, users can freely add or remove stakeholders to consider, position the Stakeholder Cards, select or add conditions, and compare the Impact Map between conditions.

As conditions change, the system generates updated impact descriptions and suggested positions for each stakeholder, but users retain full control to edit or override any suggestion. The Impact Canvas becomes a workspace for organizing perspectives, analyzing how outcomes change under different conditions, and gradually building a richer interpretation of the policy.

5.6.3 Design Considerations in PolicyScope

Card-Based Representations for Flexible and Structured Exploration

A core design intention behind PolicyScope is to make complex policy structures both visible and manipulable. Stakeholder Cards and Condition Cards provide modular building blocks that users can compose, reorganize, and reinterpret as they explore. This card-based representation serves several purposes. First, cards act as concrete anchors for reasoning. By externalizing key elements such as characteristics, conditions, and impacts, the system reduces users' cognitive load and allows them to focus on how these elements relate to one another. Second, the manipulability of cards enables flexible exploration. Users can freely add, remove, or reposition cards, treating them as representations of hypotheses rather than fixed truths. This supports iterative sensemaking as users test and refine their understanding.

In addition, the spatial layout of the Impact Map provides additional cues for interpretation. The horizontal axis (positive–negative impact) and vertical axis (relevance) allow users to see how impacts shift as conditions change. These spatial changes function as intuitive signals that highlight differences across scenarios, helping users grasp conditional variations without requiring extensive textual comparison.

Human–AI Collaborative Workflows for Policy Impact Exploration

PolicyScope is built around a division of labor between the user and the system. In the construction stage, users create the foundational components of the analysis by authoring stakeholders, defining their characteristics, outlining possible effects, and specifying conditions. The system supports this process through suggestions that identify additional stakeholders, impacts to them, or important conditions. In the system, there suggestion features are only presented as optional, ensuring that users retain conceptual control while benefiting from the system's broader generative capacity.

In the exploration stage, users and the system play complementary roles. Users shape the *intention of exploration* by selecting or creating stakeholders and conditions that reflect the aspects they wish to examine. Based on this user-defined intention, the system *instantiates the exploration* by generating anticipated impacts for each stakeholder–condition combination. Beyond producing textual impact descriptions, the system also generates concise summaries and updates the suggested positions of Stakeholder Cards on the Impact Map, showing how each stakeholder's situation may shift under the selected condition. These visual and positional cues help users quickly see how their choices translate into concrete differences in impact. This workflow preserves user agency while reducing the cognitive effort required to enumerate and organize possibilities.

5.7 Technical Pipeline Behind PolicyScope

PolicyScope is based on a multi-step pipeline that analyzes a policy and generates three key components for exploration: the stakeholder groups affected by the policy, the conditions under which the policy may perform differently, and the impacts each group may experience under specific conditions. These elements together form the basis of the interactive experience, allowing users to examine how a policy plays out across different scenarios and for different people. The following subsections describe how the pipeline generates conditions, stakeholders, and impacts, and how the system supports additional on-demand refinement during user interaction. Figure 5.4 provides an overview of the generation pipeline. In this section, we describe the detailed generation methods and then present the results of our expert evaluation of the generated stakeholders, conditions, and impacts.

5.7.1 Generation Methods

Generating Stakeholders

Our system identify policy-relevant stakeholders through a multi-stage LLM pipeline that iteratively expands and refines candidate groups. This pipeline integrates information from both the policy text and external evidence, refines broad categories into meaningful subgroups, and consolidates the results into a coherent and non-redundant set. The overall process consists of four stages: initial generation, evidence-based expansion and refinement, intra-stakeholder differentiation, and final consolidation.

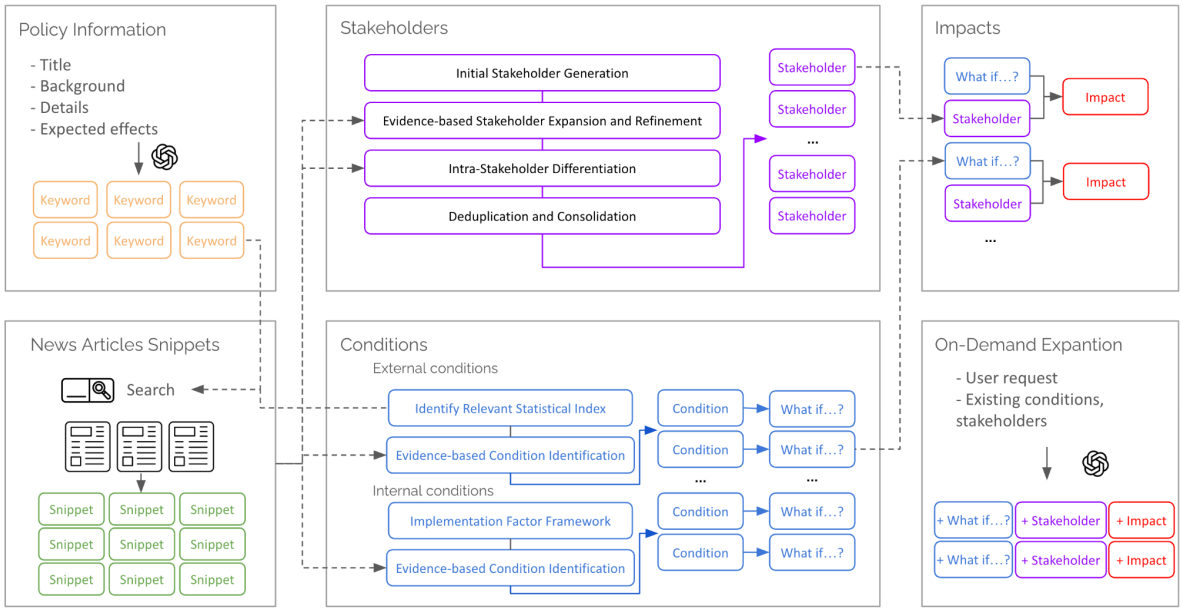


Figure 5.4: Overview of the PolicyScope pipeline. The system analyzes a policy together with evidence from news snippets to generate (1) conditions expressed as user-facing what-if questions, (2) stakeholder groups refined through a multi-step LLM process, and (3) impacts for each stakeholder–condition pair. The system also supports on-demand expansion, allowing users to request additional conditions, stakeholders, or impacts during exploration.

Initial Stakeholder Generation The pipeline begins by constructing an initial set of stakeholders using the policy description. The system prompt LLM to identify both positively and negatively affected stakeholders and consider economic, social, and environmental impact of the policy[202, 203] when generating this initial set. To guide this process, the prompt includes concrete examples aligned with these dimensions, as well as illustrative stakeholder groups drawn from two different policy contexts. This initial stakeholder set serves as the starting point for the subsequent evidence-based expansion and refinement stage.

Evidence-based Stakeholder Expansion and Refinement Based on the initial set of stakeholders, the system expand the stakeholders set by incorporating real-world discussions around the policy. It constructs search queries by combining the title of the policy and LLM-generated keywords, and retrieves articles through the Naver News[2] API. The retrieved articles are segmented into sentence-level snippets and served as input to LLM prompt later. The system then prompt the LLM to expand the set to include stakeholders mentioned in or implied by the scraped news snippets. This step allows the pipeline to incorporate real-world discussions to identify groups that may not be immediately apparent from the policy description alone (e.g., taxi drivers for DRT policy), but clearly affected by the policy. After expanding the stakeholder set, we instruct the LLM to refine stakeholder names and clarify broad or ambiguous categories, using both the policy description and the evidence from news snippets as reference points. This ensures that each stakeholder is labeled in a way that reflects how the group is discussed or positioned in real-world contexts. For example, a broad stakeholder such as 'students' can be refined into a more policy-relevant category like 'students who rely on public or shared transportation for commuting

[2]https://news.naver.com/

to school', or Similarly, 'citizen groups' can be refined into 'civic organizations focused on transportation issues or mobility rights'.

Intra-Stakeholder Differentiation In this step, the pipeline re-examines each stakeholder group to see if there are any factors or conditions under which members might be affected by the policy differently from one another. Such divergence may arise from differences in resources, capabilities, or living circumstances of individuals. For example, under a demand-responsive transport (DRT) policy, older adults who are familiar with mobile apps may benefit from the convenient booking process, while older adults with limited digital access may experience significant barriers to using the service. We prompt the LLM to identify the specific factors within each stakeholder group that could create meaningful differences in how members experience the policy.

Deduplication and Consolidation After differentiation, the expanded stakeholder list may include overlapping or semantically similar groups that emerged across stages. To produce a coherent final set, the pipeline performs a consolidation step in which the LLM is prompted to identify near-duplicate stakeholders and merge them under the clearer or more representative label. During this process, the model is instructed to preserve meaningful distinctions introduced in earlier steps (e.g., condition-dependent subgroups), while reconciling stylistic or redundant variations in naming.

Generating Conditions

To identify conditions that may affect the outcomes of the policy, our pipeline employs a multi-step, LLM-assisted condition generation process. We generate two types of conditions, external and internal, each representing a different source of variability in policy outcomes.

External Conditions System constructs set of external conditions that may affect the policy's performance (e.g., birth rate, healthcare accessibility, crime rate). Our pipeline identifies relevant external conditions by (1) inferring potentially relevant statistical indicators associated with the policy and (2) gathering recent news articles related to the policy and each indicator. We leverage a curated list of statistical indicators provided through the e-Nara Indicator system[3], an official web-based national statistics platform that offers government-selected indicators used for policy formulation, monitoring, and performance evaluation. The system comprises 747 indicators across 16 policy domains and 61 subdomains. Given policy information, our pipeline first infers up to two relevant policy domains. All indicators associated with these inferred domains are then treated as potentially relevant indicators for analyzing external conditions that may alter the policy's effects.

For each indicator, the system constructs search queries by combining the title of the policy, LLM-generated keywords, and the indicator name, and retrieves articles. The retrieved articles are segmented into sentence-level snippets and served as input to an LLM prompt that generates a list of external conditions. Lists of conditions generated from each indicator were then aggregated and deduplicated to obtain the final set of external conditions.

Internal Conditions In addition, our pipeline constructs a set of internal conditions that reflect implementation-related factors within the policy itself. It is related to how the policy is designed, communicated, executed, and received or adopted by the citizens. Based on prior discussion on factors of policy performance ([204, 205, 206], we prompted the LLM to identify potential internal factors by considering administrative capacity, beneficiary uptake and compliance, information and awareness, inter-agency coordination, and procedural or design-related complexity. Following the same procedure

[3]https://www.index.go.kr/

as for external conditions, the system retrieves relevant articles and uses them to construct final set of internal conditions.

Converting Conditions into What-If Questions To make these conditions more accessible to end users, system translate each condition into a short what-if question using LLM. This question frames the condition as a plausible scenario the user can quickly grasp. For example, an external condition such as 'high elderly population rate' is presented as 'What if the region has a much larger share of older residents¿, and an internal condition such as 'low beneficiary uptake due to digital barriers' becomes 'What if people struggle to use the mobile application?'.

Generating Impact

After constructing the final set of stakeholders and conditions, the pipeline generates the impacts associated with each stakeholder–condition pair. In this step, the LLM is prompted to describe how the policy affects a given stakeholder under a specific condition, considering both positive and negative consequences. The prompt also guides the LLM to describe each impact in terms of relevance and impact magnitude, using a 1–5 scale. Relevance captures how directly the stakeholder relates to the policy, and magnitude reflects how significant the effect might be.

In addition to producing the full impact description, the pipeline also generates a short impact highlight that shows stakeholder's point of view (e.g., "I don't know how to request a ride."). This highlight is designed to help users to easily get the idea without reading the full explanation.

On-Demand Expansion of Condition, Stakeholder, and Impact

In addition to the outputs generated through the full pipeline, the system also supports on-demand expansion during user interaction. When users wish to explore additional conditions or stakeholder groups beyond those initially produced, the system issues lightweight LLM prompts that take the existing list, the policy description, and evidence snippets as context, and generate new candidates that were not included in the initial set. This hybrid approach was chosen to minimize latency during interaction. Pre-generating all impacts for every stakeholder–condition combination would be computationally expensive and slow to load, whereas on-demand generation ensures that only relevant content is produced at the moment a user requests it, keeping the interface responsive.

5.7.2 Evaluation of Generated Stakeholders, Conditions, and Impacts

To assess the quality and plausibility of the system-generated stakeholders, conditions, and impacts, we conducted an expert evaluation with professionals who have substantial experience in policy analysis and public administration. We recruited four experts specializing in transportation policy, public health, and labor policy, and emergency management. Two of them were government officials directly involved in policy development and implementation, and the other two were PhD-level researchers who study the corresponding policy areas. Experts were recruited via cold outreach using publicly available contact information from policy schools, government agencies, and academic publications. We prepared two to three policies closely aligned with each expert's domain and asked them to select one policy for the evaluation. The policies used in the study were:

- **Demand-Responsive Transport (DRT).** A transportation policy that provides flexible, on-demand mobility services in areas where conventional public transit is difficult to operate efficiently.

- **Long-Term Care: Expansion of Monthly Allowance for Intensive Home-Care Services.** A policy proposal to expand the monthly usage limits for high-intensity home-care services and strengthen alternatives such as short-term respite care and home nutrition services.
- **Prohibition of Comprehensive Wage System.** A policy that ban comprehensive wage system, which is a pay structure where overtime, nightshift, and holiday pay are pre-included in the base salary, regardless of actual hours worked.
- **Advanced Equipment and AI-Based 119 System Infrastructure Expansion** A policy to expand advanced firefighting equipment and AI-based emergency response (119) system infrastructure.

Method

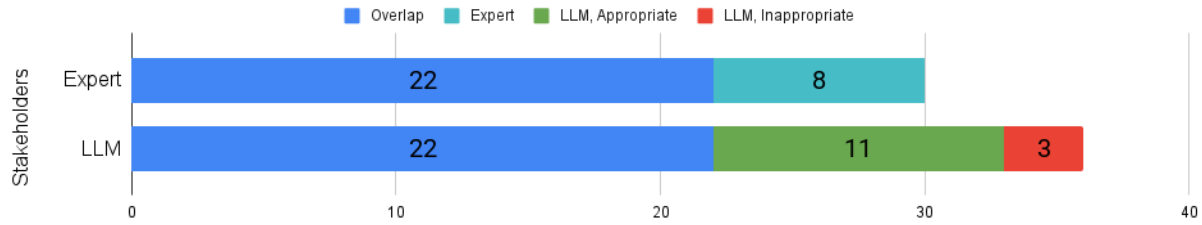
We evaluated the quality of the system-generated output by comparing it against expert-generated stakeholders, conditions, and impacts. Before beginning the evaluation, each expert received a brief description of the target policy and was informed that they could look up any additional information or reference materials as needed. However, none of the experts consulted external sources, as they were already highly familiar with the policy domain and felt confident proceeding based on their existing knowledge.

Stage 1: Identifying Stakeholders and Conditions In the first stage, experts were asked to identify a set of stakeholders and conditions. They were asked to list all stakeholder groups they considered relevant and describe the potential impacts the policy may have on each group. For every stakeholder, experts also rated (1) how relevant the group is to the policy (1 = peripheral, 3 = indirectly but meaningfully affected, 5 = primary target) and (2) the magnitude of the expected impact (1 = minimal, 3 = moderate or localized, 5 = substantial change to livelihood, rights, or roles). Experts then identified internal and external conditions that could meaningfully shape policy outcomes.Each condition was rated for its importance (1 = minimal influence, 3 = partial or localized influence, 5 = potentially decisive for policy success). All expert-generated items were collected in a shared spreadsheet and used as the reference set for the next stage.

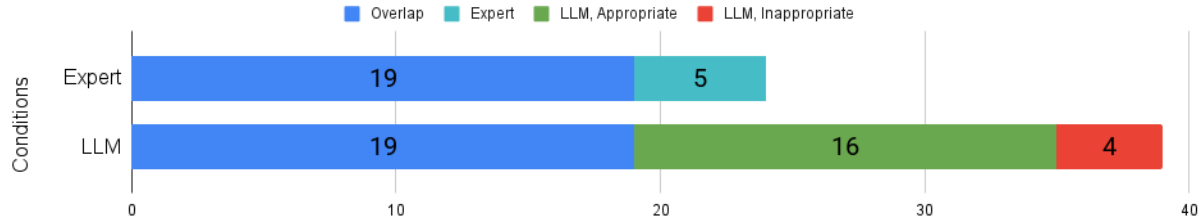
Stage 2: Evaluation of System-Generated Stakeholders, Conditions, and Impacts In the second stage, experts were guided to evaluate system-generated stakeholders, condition, and impacts. Before starting the evaluation, experts were given an explanation of how the system pipeline works. Then, they evaluated the outputs produced by the system using the same structure as in Stage 1. For each system-generated stakeholder, experts indicated whether it appeared in their own set. If not, they judged whether the stakeholder was still appropriate for the policy. Appropriate stakeholders were then rated for relevance and impact magnitude using the same 1–5 scales, allowing direct comparison with expert-generated assessments.

Experts also evaluated the system-generated internal and external conditions. For each condition, they marked whether it overlapped with their own list. When a condition was not part of their baseline, they assessed whether it was nevertheless reasonable and, if so, assigned an importance score.

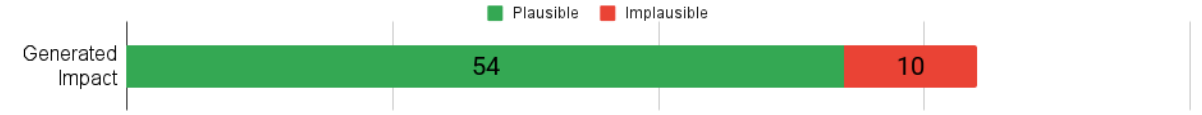
Finally, experts were asked to select at most 4 stakeholders and 4 conditions. Based on their choice, the researcher retrieved system-generated impact for those stakeholder and condition pairs. Then, for each of the retrieved impact, experts were asked to evaluated the plausibility of the stakeholder–condition scenarios generated by the system.



(a) Stakeholders



(b) Conditions



(c) Impacts

Figure 5.5: Comparison and Evaluation of Pipeline-Generated Policy Elements. Subfigures (a) and (b) compare the number of stakeholders and conditions identified by the pipeline and by experts. Subfigure (c) shows experts' judgments of the appropriateness of impacts generated by the pipeline.

Post-Evaluation Interview After the evaluation, we conducted a brief interview with each expert to understand how they made their judgments. We asked what criteria they used, which system-generated items they found appropriate or inappropriate, and whether any outputs revealed perspectives or conditions they had not initially considered.

Results

Overall, experts found the system-generated outputs largely satisfactory. Although the recall of expert-identified stakeholders and conditions was moderate, the system successfully captured most stakeholders and conditions that experts rated as highly relevant or important. Moreover, most of the system-generated items that experts had not initially identified were judged as appropriate and even valuable additions. At the same time, experts pointed out several issues that reflected the system's limited understanding of policy implementation processes or the technical use of domain-specific terminology.

Coverage and Accuracy of Stakeholders In total, experts identified 30 stakeholders (min:6, max:8 per policy) and our system pipeline identified 36 stakeholders (min:7, max:10 per policy). Out of 30 expert-identified stakeholders, 22 were also identified by the system. When examining stakeholder relevance, the system covered almost all stakeholders (16 out of 19) that experts rated as highly relevant (relevance higher than 3). In reviewing the generated list, experts identified several outputs that exceeded their initial expectations. For instance, regarding the system identified stakeholder 'local small business owners' for the demand-responsive transport (DRT) policy, the expert said *"This is better than I expected.*

Local small business owners are on my list as well, but they are a group that most people would not typically think of.". Likewise, regarding the 'family caregiver', expert said *"It is quite notable. In Korea's caregiving system, if a family member holds a certified care-worker qualification, the system allows that family member to receive the care-worker allowance. This is a distinctive feature of the Korean system. [...] Although many people assume that care workers are external third-party providers, in reality, caregiving is frequently carried out by family members."*

Out of 14 system-generated stakeholder that are not identified by experts, 11 were viewed as appropriate by the experts and half of them were viewed as of high relevance. One example was the stakeholder 'local taxi drivers' generated for the DRT policy. The experts noted that, although they had not initially considered this group, taxi drivers operating only within a local area could experience substantial losses in customers and income when DRT services expand. They also pointed out that similar dynamics had already appeared in real-world deployments of the policy.

However, there were also cases where system-generated stakeholders were not considered plausible from the experts' perspective. One example was 'local ICT startups' identified as a stakeholder in the DRT policy. Experts noted that, while such firms could exist in principle, the actual implementation of DRT services typically involves established companies with reliable operational capacity as local governments tend to partner with large mobility platforms or major automotive companies. Experts commented that this type of output reflects situations where the system lacks awareness of how policies are operationalized on the ground, making some inferences less realistic.

Coverage and Accuracy of Conditions Experts identified a total of 24 conditions (min: 5, max: 7 per policy), while the system generated 39 conditions (min: 8, max: 11 per policy). Out of the expert-identified conditions, 19 were also identified by the system. Among the remaining 20 system-generated conditions not included in the expert lists, 16 were viewed as appropriate, and all of these were evaluated as highly important (importance score higher than 3). Experts noted that the system was generally effective at identifying conditions and highlighted that several system-generated conditions directly matched factors known to affect policy outcomes in real deployments. For example, regarding the system-generated factor "population density of the region," the expert noted that it substantially affects policy performance by shaping service pricing, operational efficiency, and ultimately levels of user demand in DRT policy.

Still, there were cases where system-generated conditions were not plausible. As with stakeholders, these typically occurred when the system assumed implementation mechanisms that do not reflect how the policy operates in practice. For instance, the system identified "local governments' capacity to develop and operate ICT systems" for the DRT policy, but experts commented that this factor is not especially important because local governments generally partner with external operators. Such conditions were viewed as stemming from the system's limited understanding of institutional processes and operational constraints. Likewise, the system identified "What happens if the long-term care insurance fund becomes financially unstable?" as a condition; however, the expert noted that this scenario is unrealistic because the insurance contribution rate is adjusted annually based on financial projections and is set only within ranges that ensure the fund's long-term stability.

Some system-generated conditions were considered overly broad. This occurred mostly with questions derived from internal factors, such as "What if policy outreach and information accessibility are low?" and "What if inter-agency coordination is poor?" Experts noted that while these are indeed important considerations, they are too general to offer meaningful insight for evaluating a specific policy.

Plausibility of Generated Impacts Experts evaluated a subset of system-generated scenarios

and found that most impacts were not only plausible but of unexpectedly high quality. Out of 64 impact scenarios, 54 were marked as plausible. Experts repeatedly emphasized that the system did "a very good job" in capturing how policies produce downstream effects, often in ways that exceeded their expectations for an automated model. They noted that the system's reasoning was more nuanced than anticipated and that it identified connections and institutional dynamics that non-experts would rarely consider. Experts also highlighted that the system was able to correctly anticipate several real-world mechanisms, for example, identifying relevant institutional actors or recognizing how service expansion could place additional financial pressure on existing systems. Overall, the experts characterized the generated impacts as impressively aligned with real policy behavior.

Still, there were cases in which the generated scenarios were viewed as unrealistic or incorrect. In most instances, these issues arose from misinterpreting the given condition or misunderstanding the roles of certain stakeholder groups. For example, one expert explained:

> *"In transportation policy and related fields, the term 'transportation infrastructure' has a specific technical meaning, referring only to fixed physical assets such as roads, railways, or bridges. In this impact scenario, however, it appears that the system treated taxi drivers as part of the infrastructure, which is not correct. This kind of distinction is rarely discussed in public discourse, so it is not surprising that an LLM—or a non-expert—might conflate the terms."*

There were also cases in which the generated scenario was conceptually reasonable but framed in an overly extreme way. For instance, in response to the condition "What if medical professionals and care workers are in short supply?", the system predicted that the National Health Insurance Service would face "policy failure" and that services would become "unavailable" even if benefit limits were raised. The expert noted that while service shortages are indeed a plausible concern, describing the outcome as outright policy failure or complete unavailability was unnecessarily extreme.

Coverage, Relevance, and the Risk of Noise While the result shows that the system generates many appropriate stakeholders and conditions, including some that experts initially overlooked, this breadth is open to different interpretations. Not all appropriate outputs are equally relevant or important for understanding a given policy. Some system-generated elements, though technically correct, may serve as distractions rather than helpful guidance. For instance, outputs that are appropriate but of low importance can add cognitive load without meaningfully improving users' ability to evaluate the policy. This suggests that simply maximizing coverage is insufficient. Instead, system must also consider how to surface elements that are most likely to support users' exploration and understanding. At the same time, the fact that the system identified stakeholders and conditions that experts missed highlights its potential to broaden the range of perspectives users consider, as long as users have sufficient support to distinguish between what is central and what is peripheral.

5.8 User Evaluation

We evaluate how PolicyScope supports users in understanding complex policy information, we conducted a within-subjects user study with 16 participants. Each participant analyzed two policies across two conditions: a baseline condition and the PolicyScope condition. This study was designed to compare how users explore, interpret, and evaluate policy information when using PolicyScope versus their natural information-seeking methods, such as search engines, news articles, and conversational AI tools.

Our evaluation focuses on three research questions:

RQ1. How do users engage with PolicyScope to explore and make sense of policy information? This question examines the ways users make use of PolicyScope's features during policy exploration. In particular, we analyze how the system supports users in generating and selecting stakeholders, formulating what-if questions, and navigating different aspects of the policy.

RQ2. Does PolicyScope help users consider a broader range of factors when evaluating a policy? Here, we investigate whether users leverage the variety of information they encounter through PolicyScope to construct richer interpretations. We analyze whether their justifications, assessments of strengths and weaknesses, and suggestions for improvements reflect a broader set of factors, multiple perspectives, and recognition of trade-offs.

RQ3. Does PolicyScope lead users to feel more informed in their policy evaluation? In this research question, we examine whether PolicyScope influences users' subjective experience of political decision-making—specifically, whether they feel more informed, more confident, and better supported when evaluating policies.

The following sections describe our study design, experimental conditions, policies, participant recruitment, tasks, and measures.

5.8.1 Method

Study Design

We used a within-subjects design with two conditions (Baseline vs. PolicyScope) and two policy scenarios. Participants were asked to analyze and evaluate one policy in each condition. We employed a within-subjects design to compare differences in exploration patterns, reasoning structure, and perceived informedness within the same individuals. The order of conditions and policies was counterbalanced.

Conditions In the baseline condition, participants were asked to analyze a policy using any information-seeking tools they would naturally use in real life, such as search engines, news articles, or conversational AI tools. This setup is intended to reflect a naturalistic policy sensemaking process. In the PolicyScope condition, participants were asked to analyze a policy using our system and may optionally incorporate other information-seeking tools as needed. In both conditions, participants were provided with a blank Google Docs file where they could take notes or record any information they considered important during their analysis.

Policies We used the following two policies for the main experimental tasks: (1) Construction and expansion of the GTX and metropolitan railway network and (2) Enhancing the feasibility of urban housing supply through redevelopment and reconstruction. These policies were selected from the national agenda announced by the newly elected president in the year the study was conducted. From the broader set of policy items, we chose those that are relatively easy for the general public to understand while involving a diverse set of stakeholders.

Participants We recruited 16 participants (balanced by political orientation) by posting a study call on university community boards and the researcher's social media accounts. Participants completed a 1–1.5 hour session and were compensated 40,000 KRW for their participation.

Task and Procedure In each condition, participants first viewed the title of the assigned policy and indicated how familiar they were with it. They then read a short policy summary and completed a pre-task survey that asked about the policy's relevance, their initial stance, and their expectations about its potential effects. Following this, participants were given up to ten minutes to freely explore the

policy before writing their evaluation. In the baseline condition, participants were encouraged to use any information-seeking methods they would normally rely on—such as search engines, online news articles, or conversational AI systems. In the PolicyScope condition, participants were encouraged to use our system as much as they wished, while still being allowed to engage in additional free-form exploration using any external tools they normally rely on, similar to the baseline condition.

After the exploration period, participants wrote their evaluation of the policy by responding to four open-ended questions about their overall stance, the policy's strengths, its weaknesses or risks, and possible improvements. Participants were allowed to continue gathering information during this writing stage. At the end of each condition, participants completed a post-task survey that measured their perceived understanding of diverse stakeholders and conditions, perceived informedness, confidence in their decision, and cognitive load. Participants also completed a usability questionnaire when using PolicyScope.

After completing both conditions, participants had a short semi-structured interview in which participants reflected on their exploration strategies, the perceived advantages and limitations of PolicyScope, and how the system influenced their reasoning, sense of informedness, and confidence in decision-making.

Measures

Regarding RQ1, we examine participants' exploration patterns. In the PolicyScope condition, we examine interactive behavior within the interface, including the breadth of elements explored (e.g., stakeholders, impacts, conditions), transitions between these elements, and the overall structure of their navigation patterns. We also compare these behaviors to the baseline condition by examining how participants seek and reference external materials to understand how their information-seeking strategies differ when using our system. We additionally assess users' subjective perceptions of engagement and ease of exploration to understand whether PolicyScope supports more flexible and curiosity-driven exploration of complex policy information.

To answer RQ2, we analyze the participants' written responses across three components of the evaluation task: position justification, strengths and weaknesses, and suggestions for improvements. In the justification and strength/weakness responses, we analyze the diversity of factors they reference (e.g., affected stakeholders, contextual conditions, feasibility), how well these factors are connected into a rationale, and whether they acknowledge trade-offs or consider multiple perspectives. In the improvements response, we assess how specific and concrete participants' suggestions are, for example, whether they propose identifiable changes to particular components of the policy rather than broad or generic statements.

Regarding RQ3, we used 7-point Likert scale questions on participants' perceived informedness, confidence in decision, perceived usefulness of exploration, perceived support. We also ask open-ended questions on their experience using baseline and PolicyScope.

Coding Scheme for Policy Evaluation Responses

We analyzed participants' open-ended responses to the policy evaluation questions to assess the quality of their reasoning. First, we counted the number of distinct arguments in the response. Argument is defined as a meaningful idea such as a justification, mechanism, or expected effect. In addition, we counted how many specific stakeholders were explicitly mentioned. We included only clearly identified groups (e.g., "small business owners," "construction firms," "local residents") and excluded broad

and unspecific terms such as "people" or "society." Likewise, we coded whether the response included a conditional statement describing how the policy's impact may change under certain situations. General statements that described problems without an explicit condition were not counted as conditional reasoning.

5.8.2 Results

RQ1. How do users engage with PolicyScope to explore and make sense of policy information?

Overall, participants actively engaged with PolicyScope during policy exploration. Even without any prescribed tasks or instructions, participants created and selected multiple stakeholders and conditions, arranging them into impact scenarios to understand the diverse potential effects of the policy. In general, participants created a median of 2 stakeholder profiles (including their own, min: 1 , max: 7), selected 2 stakeholders (min: 1, max: 5) recommended by the system. Six participants asked the system to generate more stakeholders and the system generated 12 more stakeholders total. Regarding conditions, participants made 4 (min: 0, max: 9) what-if questions, selected 2 (min: 0, max: 4) system-generated questions, and asked the system to generate 8 (min: 2, max: 13) more questions. In the exploration stage, each participant considered 6 stakeholders (min: 4, max: 9) under 5 conditions (min: 3, max: 12). Table 5.1 shows stakeholders and questions that P1 (railway) and P13 (urban housing) selected in using PolicyScope.

Post-task interviews showed that the system's structured exploration helped participants broaden their thinking. P12 noted that *"there were many situations and group of people I hadn't thought of, and it was helpful to consider people and situations together."* Also, P1 said *"I appreciated that the system recommended stakeholders I would have never thought of on my own. Some of them were clearly important, yet I would not have considered them without the system's suggestion."*. At the same time, there were cases that system recommended stakeholders or conditions that looked less related or important to participants. Nevertheless, participants saw the occasional irrelevant recommendations as an acceptable side effect of being offered a more comprehensive set of options. P13 said that it is better to receive a wide range of recommendations and select the meaningful ones than to risk missing potentially important perspectives.

PolicyScope also shaped how participants sought additional information outside the system. In the baseline condition, 11 out of 16 participants used external resources to obtain additional information about the policy. All of them used Google or Naver search, entering the policy title or technical terms from the description as their queries. Five participants also used tools such as ChatGPT or Gemini to request further explanations or analyze the policy. In the PolicyScope condition, 9 out of 16 participants used external resources, with 4 used AI tools, in addition to the system. The overall pattern was similar to that of the baseline condition, but participants in the PolicyScope condition used more specific and targeted search queries, often referencing particular situations or viewpoints, such as redevelopment gentrification (P14), redevelopment with low profitability (P3), or GTX project backlash (P12). Regarding this, P10 noted that he searched using queries he *"would not have come up with without the system."*

In addition to behavioral patterns, we examined whether PolicyScope affected the subjective workload associated with understanding and analyzing policy. After each condition, participants rated perceived cognitive load and level of effort for the task on a 7-point scale. Cognitive load was significantly lower in the PolicyScope condition (median: 3.5) compared to the baseline condition (median: 5, Z =

Table 5.1: Examples of system-supported exploration: (a) railway policy and (b) urban housing policy.

Category	Item	Generated by
Stakeholders	Resident in the outer Seoul metropolitan area	user (own)
	Apartment resident near the GTX construction site	user
	Landowner in the construction area	user
	Environmental organization	system
	Construction and infrastructure firm	system (added)
	Resident in a transport-disadvantaged area	system
Conditions	What if low ridership keeps the line in deficit?	user
	What if Korail or express-bus unions oppose the project?	user
	What if transfer centers are poorly connected to other transit?	system
	What if population density in outer areas is low?	system
	What if the environmental impact assessment is negative?	system (added)

(a) Stakeholders and conditions selected for railway policy exploration (P1).

Category	Item	Generated by
Stakeholders	Woman in her 20s seeking to buy housing in the Seoul metropolitan area	user (own)
	Nearby property owner	user
	Resident living outside the city seeking to move to an urban area	user
	Building owner who can secure temporary housing during reconstruction	system
	Reconstruction-target building owner with limited cash liquidity	system
	Private construction company	system
	Small business owners in the redevelopment area	system (added)
Conditions	How would construction costs change if inflation fluctuates?	user
	What if many apartments remain unsold?	user
	What if approval procedures are overly simplified?	system
	What if old housing is scattered across the region?	system
	What if access to financial support increases?	system (added)

(b) Stakeholders and conditions selected for urban housing policy (P13).

-2.27, $p<0.05$). Effort showed a similar pattern, with lower scores in the PolicyScope condition (median: 4) than in the baseline (median: 5.5, Z = -1.99, $p<0.05$). Interview responses aligned with these findings. P5, who used PolicyScope first and then analyzed another policy without the system, described the contrast clearly: *"When I used the first system, it felt easy to engage with. However, when analyzing the second policy without the system, it felt like cramming. It just wasn't enjoyable, so I didn't feel motivated to seek additional information."* They added that while external sources such as Namuwiki or Google

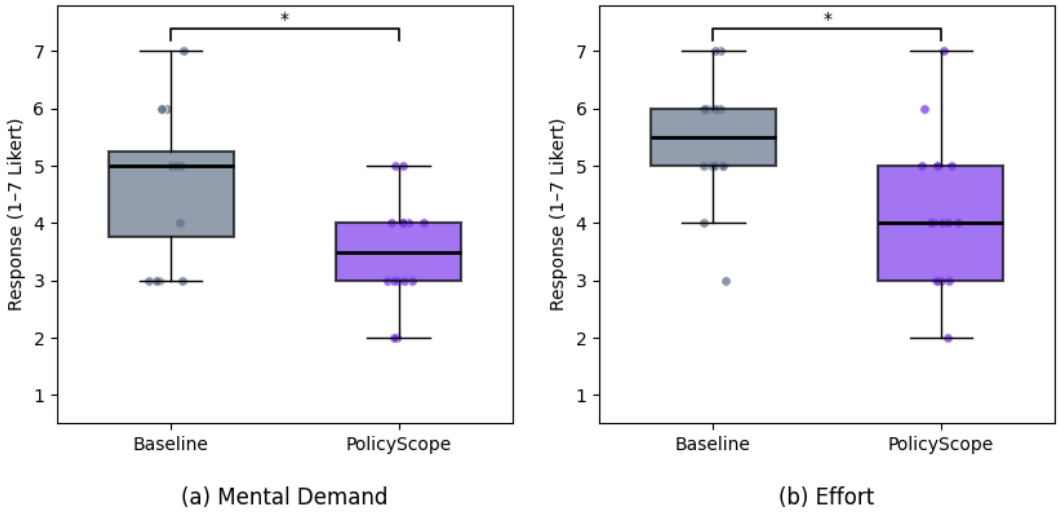


Figure 5.6: Perceived Cognitive Demand and Effort by Condition. Asterisks indicate statistically significant differences ($^*p < 0.05$).

offered more content, *"they include a lot that isn't necessary for me, so I have to sort through everything myself, which takes a lot of effort. And it's not fun. It's not fun at all! It's much more engaging when it's interactive."*

RQ2. Does PolicyScope help users consider a broader range of factors when evaluating a policy?

Participants provided a greater number of distinct arguments, discussed more number of stakeholders and conditional factors that may affect the policy effects. In the PolicyScope condition, participants provided more grounded arguments when explaining the pros and cons of the policy. Their responses included a higher number of arguments (med=7.5, min: 5, max: 9) compared to the baseline condition (med=4.5, min:2, max:8). Participants also drew on a wider range of considerations in their reasoning. Participants mentioned median of 3 stakeholders (min: 0, max: 7) in PolicyScope condition, compared to a median of 1 (min: 0, max: 3) in the baseline condition (Z=-2.79, p<0.05). Likewise, participants referenced more contextual conditions that could influence the policy's outcomes in the PolicyScope condition (med = 2, min: 0, max: 7) than in the baseline condition (med = 0, min: 0, max: 3).

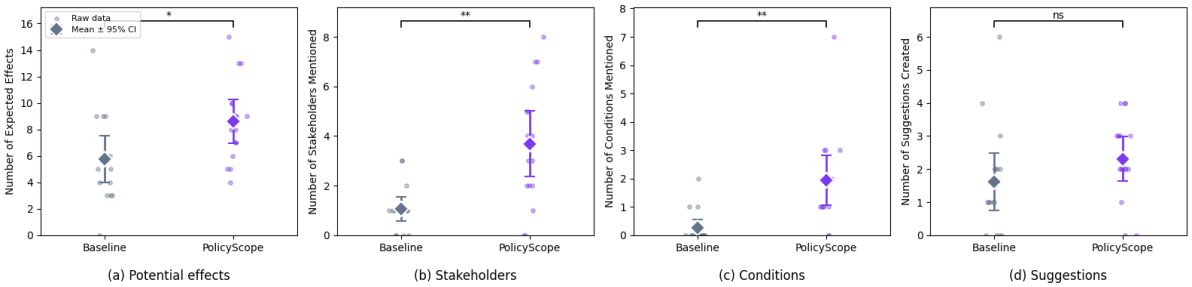


Figure 5.7: Comparison of argument quality metrics: (a) Potential effects, (b) Stakeholders, (c) Conditions, and (d) Suggestions. Error bars show 95% CI. Wilcoxon signed-rank test: * $p < 0.05$, ** $p < 0.01$.

Table 5.2 shows stakeholders and conditions considered in participants' analysis of each policy. This suggests that PolicyScope helped participants consider a wider set of affected groups and identify more concrete situational factors that could influence the policy's impact, leading them to provide more detailed and specific reasoning in their evaluations. Participants also proposed a slightly greater number of improvement suggestions in the PolicyScope condition (med=2, min: 1, max:3) than in the baseline condition (med=1, min: 0, max: 4), though this difference was not statistically significant (Z=-1.91, p = 0.06).

Participants reported that PolicyScope significantly helped them understand a policy from multiple perspectives. They noted that the system-generated stakeholders and policy impacts on them made it easier to grasp the broader landscape of impacts. Several participants said that this information helped them anticipate what might happen next if the policy were implemented, which they found difficult to do with policy descriptions alone. Seven participants also emphasized that they usually encounter policy information through news articles but often perceive such coverage as biased or one-sided. Indeed, P3 mentioned that they routinely check comment sections to gauge how "ordinary people" think about a policy. Four participants explicitly mentioned that having a system like PolicyScope attached directly to news content would be very helpful for the public to understand the policy in diverse, unbiased perspectives.

RQ3. Does PolicyScope lead users to feel more informed in their policy evaluation?

Participants showed a clear increase in their perceived level of knowledge when using PolicyScope with median score of 3.5 and 5 for pre and post task survey, respectively ($Z = -2.50$, $p < .05$). In contrast, in the baseline condition, only 3 participants reported higher post-task scores whereas 13 out of 16 reported increase in the level of knowledge in PolicyScope condition.

Self-reported understanding of policy strengths and weaknesses did not significantly differ across conditions. Although scores were slightly higher in the PolicyScope condition (med: 6) than in the baseline (med: 5), the difference was not statistically significant ($Z = -1.64$, $p = .10$). Still, participants felt more capable of proposing improvements or refinements to the policy in the PolicyScope condition. Reported self-efficacy was significantly higher in PolicyScope (med: 5) than in the baseline (med: 4, $Z = -2.29$, $p<.05$).

These increases in perceived level of knowledge and internal political efficacy were not related with change in stance. In both conditions, participants' position remained almost the same before and after the task, with only 3 participants updating their position score more than 2 points (out of 7). Still, participants expressed higher confidence in their stance when using PolicyScope. Confidence scores were significantly higher in the PolicyScope condition (med: 6) compared to the baseline (med: 5, $Z = -2.98$, $p <.05$). In the post-task interview, participants explained that the structured representation of policy outcomes, especially the visual representation of impact scenarios, helped them articulate and justify their position more clearly. P4 noted, *"rather than simply listing positive and negative effects, placing the impact cards on a coordinate plane helped me organize my stance much more intuitively."* P5 also mentioned that *"I really liked that I could place stakeholders on the map. It helped me organize my thoughts, visualize how my views were leaning to one side, and even made it easier to explain my reasoning."*

Table 5.2: Stakeholders and conditions considered in participants' analysis of each policy.

Policy	Type	Example Responses
Railway	Stakeholder	**Baseline:** People who cannot afford housing in Seoul; residents in outer metropolitan areas; suburban–city commuters; companies; job seekers; homeowners near future stations. **PolicyScope:** Commuters; car owners; residents in provincial regions; financial investors; nearby housing and land owners; environmental groups; existing transit operators; older adults; young taxpayers; workers in industrial complexes; business owners near stations; low-income outer-area residents.
	Condition	**Baseline:** If fares become too high; if the project shifts to a privately financed model. **PolicyScope:** If the route is poorly designed; if expected ridership is inaccurate; if new towns attract or fail to attract population; if fares become overly expensive; if transfer systems are well integrated; if speculative investment becomes excessive.
Urban housing	Stakeholder	**Baseline:** Current residents; local shop owners; construction company; low-income households in old apartments; 2030s; primary-residence buyers. **PolicyScope:** Current residents; people who want to live in urban areas; young entry-level workers; newcomers moving into the city; local property owners; construction company; finance firms; long-time low-income original residents; local shop owners; landlords; multi-property owners; newlyweds; households without a home; residents who do not have the resources to relocate
	Condition	**Baseline:** If reduced construction costs are reflected in consumer prices; if the area include large privately owned plots; if public agencies rush a project **PolicyScope:** If agreements with existing residents cannot be reached smoothly; if the stability of the redevelopment process may not be fully ensured; it it comes with measures to prevent speculation; if this policy draws even more people into the urban core; if private companies are selected to carry out redevelopment

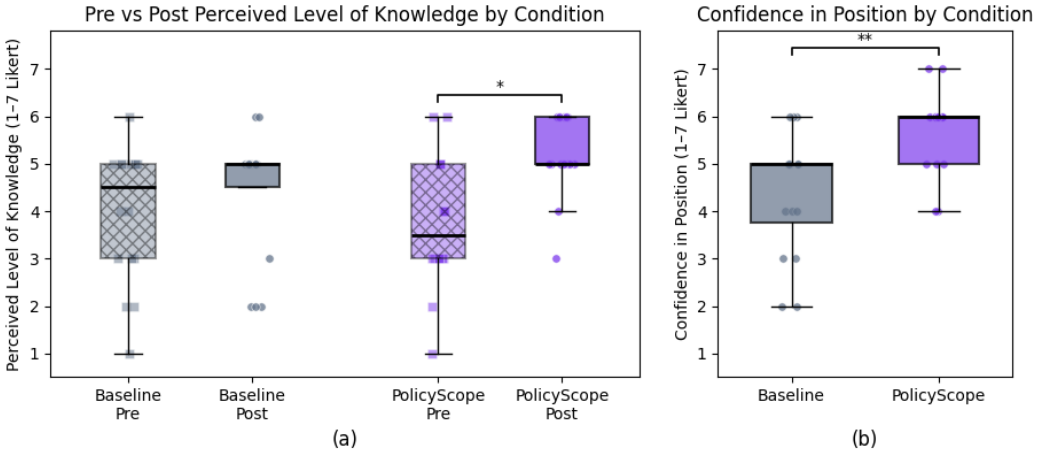


Figure 5.8: Pre–Post Changes in Perceived Knowledge (a) and Confidence Across Conditions (b). Wilcoxon signed-rank test: * $p < 0.05$, ** $p < 0.01$.

5.9 Discussion

Our findings show that PolicyScope meaningfully shaped how participants explored, interpreted, and evaluated complex policy information. The system presents structured representation of factors to consider and engage participants to play with those factors. Participants engaged with these elements to broaden the range of perspectives they considered, connect policy outcomes to concrete situations, and articulate more detailed and grounded explanations in their evaluations. At the same time, PolicyScope reduced cognitive effort and increased participants' sense of informedness and confidence. In this section, we extend our discussion...

5.9.1 Supporting Constructive and User-Driven Understanding of Policy Information

One of the important design choice made in PolicyScope is that users can decide on which aspects of a policy they want to explore. Such design allows users to construct meaning for themselves rather than consuming a fixed, expert-curated explanation of a policy. Prior tools for public policy communication often rely on simplified summaries or expert interpretations that present a predetermined view of what matters. In contrast, PolicyScope adopts a more constructivist orientation by presenting building blocks—stakeholders, conditions, and visualized impact scenarios—but leaves it to users to assemble these elements in ways that make sense for their own questions, concerns, and values. Participants' exploration behaviors and evaluation responses show that many used the system not as a source of answers, but as a scaffold to think through the policy in their own terms.

This constructivist structure is especially important in policy contexts, where what matters is not uniform across people. A condition that experts view as critical (e.g., limited administrative capacity) may feel irrelevant to lay users, and vice versa. In our study, several conditions that expert identified as important were dismissed as being "not important". This gap illustrates that public understanding is less about receiving the "correct" factors, but more about recognizing what feels meaningful to oneself. PolicyScope enabled this by giving users the flexibility to disregard suggestions, generate alternatives, and focus on the elements that fit their line of thinking. Several participants also expressed interest in

seeing how other people conceptualized the same policy within the system. Allowing users to contribute their own stakeholders and conditions, or to browse alternative maps created by others, could further strengthen this constructivist value.

5.9.2 Risks of Algorithmic Supports in Policy Sensemaking

While PolicyScope enables flexible, user-driven exploration, it also raises an important concern about how system-generated suggestions may shape users' interpretations. In fact, participants largely relied on the system's recommended stakeholders and conditions as starting points in our evaluation, which means that the initial framing supplied by the system can subtly influence what users consider relevant or worth examining further. One participant explicitly raised a concern, noting that *"This system will be very useful if we can have it in real-life. However, who will manage the system will matter a lot. If a particular institution operated a system like this, its suggestions might subtly reflect the interests of whoever runs it."* This underscores the need for transparency in how suggestions are generated, as well as design mechanisms that communicate that system-generated elements are possibilities rather than authoritative, suggestive guidance.

At the same time, PolicyScope itself is not free from these concerns. Our evaluation of pipeline was done with only four policies, which means we were not able to fully assess whether the pipeline systematically over- or under-generates certain types of stakeholders or conditions. A more comprehensive, large-scale evaluation would be necessary to identify potential coverage patterns and to refine the pipeline so that it better supports a diverse range of perspectives.

Another related risk is users' limited ability to detect errors in system-generated content. In our evaluation, some participants failed to identify spurious stakeholders or conditions that the LLM pipeline generated, accepting them as plausible without further scrutiny. This is particularly concerning because the outputs are presented in a structured, authoritative format that makes them appear reliable. When users feel uncertain about a policy, they may be even more likely to trust system suggestions, making it difficult to distinguish between appropriate, relevant outputs and plausible but incorrect ones. This highlights the unintended epistemic authority of system-generated suggestions. Especially when users feel uncertain, the breadth of suggestions produced by the system may be interpreted as expert-like guidance, leading users to treat them as more authoritative than intended. Communicating the system's incompleteness, for example by indicating uncertainty or clarifying that suggestions are not expert judgments, will be important to avoid over-reliance.

5.9.3 Reducing Cognitive Burden Through Structured Exploration

In our evaluation, PolicyScope helped make policy exploration feel more manageable while still broadening the range of factors participants considered. The structured representation of stakeholders, conditions, and impact scenarios worked as a form of cognitive offloading. In contrast, the baseline condition required participants to search, filter, and organize information on their own, which made the task feel heavier and more effortful. PolicyScope reduced this overhead by presenting relevant elements in a format that was easy to interpret and work with.

This sense of ease, however, does not imply a superficial engagement. In fact, participants generated more arguments, considered more stakeholders, and identified more conditions in the PolicyScope condition than in the baseline. In designing PolicyScope, we aimed to preserve users' exploratory intent while allowing the system to assist with tasks that are more mechanical, such as gathering, surfacing, or

organizing relevant elements, so that users can focus on the reasoning itself. Our findings suggest that this form of human-AI collaboration can lower cognitive barriers to engagement without narrowing the depth or richness of users' thinking.

5.10 Limitation and Future Work

Our findings should be interpreted in light of several limitations. First, the evaluation focused on a small number of policies, which constrains our ability to assess how well the pipeline generalizes across broader policy domains. A larger-scale expert assessment would be needed to identify systematic coverage patterns in the generated stakeholders and conditions, and to refine the pipeline for more diverse contexts.

Second, the study examined a single-session interaction. While participants showed richer reasoning and higher perceived informedness, it remains unclear how such effects persist over time or influence real decision-making. Future longitudinal studies could examine whether repeated use strengthens users' ability to analyze policies independently or supports more informed political choices.

Third, our evaluation relied on written explanations rather than behavioral outcomes. Participants' stance rarely shifted, suggesting that PolicyScope may influence how people reason without necessarily changing what they believe. Future work could examine how systems like PolicyScope affect collective deliberation, trust formation, or participation in civic processes.

Fourth, although PolicyScope was designed to preserve user autonomy, the system still provides suggestions that carry potential risks of framing and epistemic authority. Future research should explore design strategies that surface uncertainty, disclose generative provenance, or provide users with more control over the level of algorithmic guidance.

Finally, participants expressed interest in understanding how others conceptualize the same policy. Extending PolicyScope to support collaborative exploration—such as sharing maps, browsing others' stakeholder sets, or participating in group scenario building—may further strengthen its role as a tool for public reasoning.

5.11 Conclusion

In this work, we introduced PolicyScope, a system that helps users explore public policies by examining stakeholders, contextual conditions, and possible impact scenarios. PolicyScope supports user-driven sensemaking and provides structured assistance that makes complex policy information easier to engage with. Through our evaluation, we found that users generated more detailed reasoning, considered a wider range of factors, and felt less burdened when using PolicyScope compared to baseline exploration. We also identified risks related to the influence of system-generated suggestions and highlight the need for transparency and safeguards. We expect that the design of PolicyScope and our study findings can inform future tools that support the public in understanding and evaluating policy information.

Chapter 6. Discussion

6.1 Complexity as a Medium for Engagement

Across the three projects, a consistent pattern emerged: people struggled not because information was inherently too complex, but because its structure was hidden. Participants often noted that simple summaries were easy to skim yet unhelpful for forming grounded interpretations. In contrast, when relationships, assumptions, and distinctions were visible, the same content felt more approachable and easier to reason about. This suggests that complexity can support understanding as long as its structure is visible. Revealing reliability, heterogeneity, and contingency helped users feel more confident in their assessments. Rather than asking how to simplify complexity, the central design question becomes how to make its structure understandable and explorable. This highlights the value of interactive scaffolding over surface-level simplification.

This approach becomes more important as generative AI increasingly mediates how people access information. LLMs are widely used to summarize and simplify complex information, but these outputs often present the same structural challenge: complexity is hidden rather than eliminated. The systems in this dissertation offer a contrasting approach, using computational scaffolding to make complexity navigable by users themselves. Where LLMs collapse information into coherent but opaque outputs, interactive systems can surface the relationships and contingencies that would otherwise be smoothed over. In ReviewAid, users could trace claims back to sources and compare divergent interpretations. In PRISM, users could see where reactions clustered and diverged. In PolicyScope, users could manipulate conditions and observe how impact projections shifted.

However, when systems mediate public sensemaking, they introduce risks beyond traditional misinformation concerns. Information can be shaped by how relationships are structured, what comparisons are made salient, and which framings are presented as default. These design choices are often invisible to users, making systematic biases difficult to detect. In ReviewAid, decisions about which sources to surface could privilege certain narratives. In PolicyScope, the LLM pipeline produced spurious outputs that participants often could not identify. The systems in this dissertation do not eliminate these risks, but they preserve the visibility of informational structure, providing scaffolding for users to question and scrutinize what they see. This raises a question fundamentally about governance: who manages these systems, and how can accountability be ensured?

6.2 Reliability, Heterogeneity, and Contingency as a Lens on Oversimplification

The framework introduced in Chapter 2: reliability, heterogeneity, and contingency helps clarify where oversimplification occurs and how interactive systems can address it. ReviewAid underscored the importance of reliability. Making uncertainty and underlying evidence explorable helped participants overcome the illusion of certainty created by simplified headlines, leading to more calibrated and grounded interpretations. By tracing information pathways from scientific findings through media representation, users could assess how claims were supported and what caveats existed.

PRISM highlighted the role of heterogeneity. By enabling users to express nuanced reactions beyond binary signals, the system preserved the diversity of viewpoints that would otherwise be flattened by platform-level aggregation. Flexible labeling revealed that interpretations vary meaningfully, countering the false consensus implied by simple vote counts.

PolicyScope demonstrated the value of both heterogeneity and contingency. By representing different stakeholders and allowing users to manipulate conditions, the system showed that policy impacts vary not only across groups but also depending on specific circumstances. Users came to recognize that policies do not have single effects but rather patterns of effects shaped by who is affected and under what conditions.

Table 6.1: How each system addresses different dimensions of complexity through distinct interaction modes

	ReviewAid	**PRISM**	**PolicyScope**
Dimension	Reliability	Heterogeneity	Contingency
User Task	Evaluation	Expression	Exploration
System Support	Scaffolding	Capturing nuanced expression	Active construction of understanding
Design Approach	Disentangling media and research; Contextual examples	User-generated labels that balance expressiveness and easiness	Manipulable cards; AI-supported instantiation

Taken together, these systems demonstrate how different dimensions of complexity call for different modes of interaction (Table 6.1). Addressing reliability involves scaffolding for evaluation, allowing users to trace and assess evidence. Supporting heterogeneity involves expressive tools that let users articulate nuanced positions. Revealing contingency involves explorable representations that support active construction of understanding. These findings suggest that addressing oversimplification is not just about revealing complexity, but about matching the right interactive affordances to the structure of what needs to be understood.

6.3 A Temporal Perspective: Retrospective, Real-Time, and Prospective Understanding

The three systems can also be understood through a temporal lens (Figure 6.1), focusing on when users engage with information and what their specific needs are at that point. This perspective highlights a dimension often overlooked in prior work on sensemaking tools, which tends to focus on what information is presented rather than the temporal context of user engagement.

ReviewAid supports retrospective understanding by helping users revisit previously consumed information to identify missing context. This mode is critical because initial consumption is often shallow, and users rarely have the opportunity to examine evidence at the moment of reading. PRISM enables real-time expression, preserving nuance at the moment of reaction before it is flattened by aggregation. Individual interpretations are most rich when they are first formed, but platform-level aggregation typically collapses them into simplified metrics. Finally, PolicyScope facilitates prospective reasoning, helping users explore possible futures and consider how outcomes shift under different scenarios. This

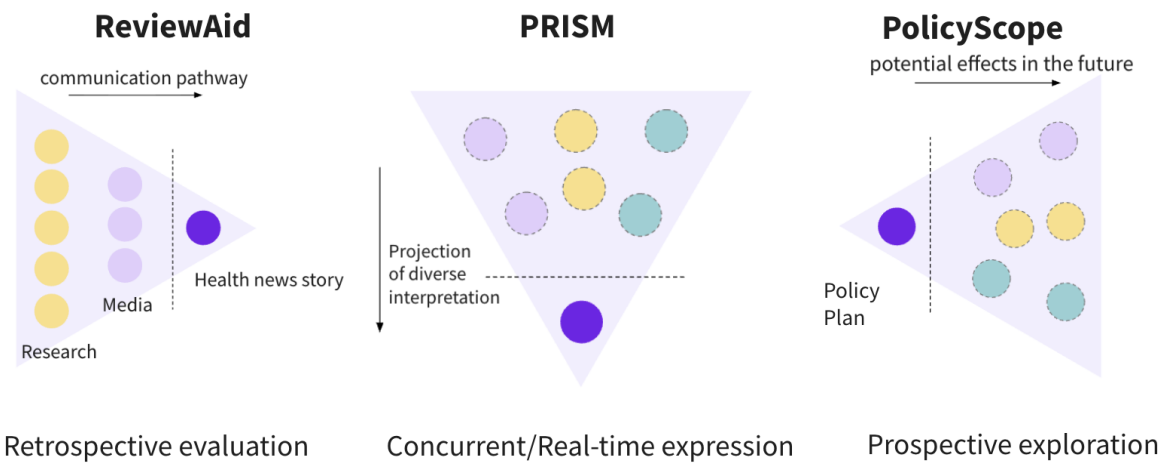


Figure 6.1: ReviewAid supports retrospective evaluation by tracing communication pathways, PRISM enables real-time expression of diverse interpretations, and PolicyScope facilitates prospective exploration of potential policy impacts.

forward-looking mode addresses how people typically engage with policy information, which is often presented as a fixed plan rather than a set of conditional possibilities.

These temporal differences show that misunderstandings can arise at multiple stages, including after reading, during expression, and when projecting implications. Designing for understanding therefore requires attention not only to what people reason about but also when they reason about it.

6.4 Navigating Trade-offs in Designing for Complexity

Designing for complexity involves trade-offs. Revealing reliability, heterogeneity, and contingency can lead to better understanding and greater awareness of nuance, but it requires more cognitive and time effort from users. This can result in indecisiveness or reduced trust when uncertainty is made visible.

The appropriateness of this approach varies by domain, situation, and user. Stakes differ across domains: health information involves personal decisions where uninformed choices can cause direct harm, policy information relates to collective decision-making and civic participation, while public opinion discourse centers on social interaction where misunderstanding has lower immediate consequences. Within domains, situations matter: a patient making treatment decisions has different needs than a general reader skimming health news. User expertise and motivation also influence when structured complexity is helpful versus overwhelming.

Our findings suggest that in high-stakes contexts, revealing complexity supports more grounded judgment. In ReviewAid, making epistemic limitations visible helped users form more calibrated assessments of health claims. In PolicyScope, exposing stakeholder differences and conditionality of effects supported more nuanced policy sensemaking. Participants across these systems reported feeling more confident even as they became more aware of complexity. However, in lower-stakes contexts or time-sensitive tasks, simplified presentation may be more appropriate.

As information systems increasingly shape how people understand science, policy, and public discourse, deciding when to reveal structure versus when to simplify becomes consequential. In domains

where understanding matters for participation and informed decision-making, systems that scaffold complexity offer a promising path forward.

6.5 Rethinking Complexity Dimensions for AI-Generated Information

In this section, we extend the discussion around oversimplification and complexity to consider how the widespread adoption of generative AI reshapes both the nature of oversimplification and the dimensions needed to address it.

Generative AI changes oversimplification in two ways: how information is produced and how it is used. On the production side, AI accelerates the tendency toward oversimplification. While conventional information is shaped by human experts who make deliberate choices about what to include and omit, users now routinely ask Large Language Models (LLMs) to generate one-line summaries of papers, policies, or news articles, creating a new layer of potentially shallow interpretation. On the usage side, however, generative AI offers opportunities for users to initiate exploration and delegate discovery tasks to the system. Nevertheless, this shift requires users to have sufficient expertise to evaluate and guide the process effectively.

This transition raises critical questions about how the reliability, heterogeneity, and contingency framework should evolve. In the current framing, these dimensions address complexity that already exists in the information landscape but has been flattened by simplification. However, with AI-generated information, this reference point may not exist. When an LLM synthesizes information from its training data, there may be no single source document to trace back to, making it difficult to identify which complexities have been omitted or distorted.

Consequently, the focus of each dimension must shift. Reliability becomes not just about tracing evidence pathways but also about assessing the provenance and confidence of AI-generated claims. Heterogeneity extends beyond surfacing existing diverse perspectives to recognizing when AI outputs collapse multiple valid interpretations into a single, homogenized narrative. Contingency must account for how AI systems handle conditional reasoning and whether they make their underlying assumptions explicit or bury them in fluent prose.

Beyond adapting existing dimensions, it remains an open question whether new dimensions are needed to capture the unique qualities of AI-generated information. One candidate is explicability, the degree to which users can understand how information was produced, what inputs shaped it, and what alternatives were considered. Another possibility is provenance, which involves distinguishing between human-authored analysis, AI synthesis of existing sources, and AI generation based on training data patterns. While these dimensions may overlap with reliability, they highlight distinct concerns about the opacity of computational processes.

Ultimately, the value of this framework lies not in its completeness but in its ability to direct attention toward what is lost in the process of oversimplification. As the information landscape evolves, so too must our understanding of which structures matter and how to make them visible.

Chapter 7. Limitations and Future Work

7.1 Limitations

This dissertation has several important limitations that should be acknowledged. First, all three systems were evaluated through single-session studies, which limits our understanding of long-term effects. While participants reported increased confidence and engagement, these findings reflect immediate perceptions rather than sustained behavioral change. It remains unknown whether users would continue to engage with complexity over time or whether initial understanding translates into different decision-making in real-world contexts.

Second, our evaluations relied on participants who were asked to engage with the systems as part of a study. Although these systems are positioned as most valuable in high-stakes contexts where users are motivated to understand complexity, the gap between study participants and actual stakeholders introduces uncertainty about real-world adoption. Future research should examine these tools in situ, where users face real-world consequences for their decisions.

Third, while we measured users' confidence and perceived understanding, we did not directly assess whether their actual comprehension improved. In PolicyScope, for example, some participants failed to identify spurious AI-generated outputs, suggesting that increased confidence does not necessarily correspond to improved accuracy in judgment. This highlights the risk of misplaced confidence, where a well-structured interface might make users feel more informed than they actually are.

Additionally, the design of these systems involves subjective choices about which structural elements to surface. These choices can introduce unintended framing effects, potentially guiding users toward certain interpretations even while attempting to promote critical thinking. In PolicyScope specifically, the reliance on generative AI models introduces additional risks related to algorithmic reliability. The system depends on the underlying model's ability to identify relevant stakeholders, conditions, and impacts. If the AI fails to capture a critical dimension of a problem or introduces hallucinations, the resulting scaffolding may be fundamentally flawed.

7.2 Future Work

The findings point toward several directions for future research. First, recent developments in AI technologies open new possibilities for reducing interaction barriers while preserving user agency. PolicyScope demonstrated how AI-generated stakeholders and conditions can make conditional exploration accessible without overwhelming users. This approach could be extended to ReviewAid and other cases that deal with complicated or difficult information. However, this raises a fundamental tension: AI support makes exploration more accessible, but how do we ensure users retain agency in constructing their own understanding rather than passively accepting AI-generated framings? Future work should explore how to design AI assistance that scaffolds exploration without determining interpretations.

In addition, ensuring accuracy and reliability of user-generated information or AI-generated information will be an important challenge in extending this line of work. PolicyScope relies on AI-generated stakeholder impacts, while PRISM depends on user-generated labels. Both raise validation questions: who verifies that PolicyScope's impacts accurately reflect real policy effects? How do we prevent PRISM

labels from being manipulated by bad actors? Future work should investigate hybrid approaches that combine AI generation with expert review, community validation, or transparent confidence indicators to maintain quality without sacrificing accessibility.

Another promising direction is developing adaptive scaffolding for diverse users. Our systems provide uniform scaffolding, but users vary in expertise, motivation, and cognitive style. Some participants wanted more guidance; others felt constrained. Future systems could adapt the amount and type of support based on users' prior knowledge or reasoning behavior: offering step-by-step prompts for novices while allowing experts to construct their own exploration paths.

Additionally, beyond individual sensemaking, future work could explore how to move from individual exploration to collective artifacts. Current systems support individual understanding, but engaging with complexity often benefits from collective intelligence. Future research could investigate how individual explorations aggregate into shared knowledge: for instance, PolicyScope users collectively building validated maps of stakeholder impacts, or PRISM users developing community interpretation frameworks. The challenge is preserving diverse perspectives while enabling synthesis.

Finally, as these systems move from research prototypes to real-world deployment, questions of governance and accountability become critical. As discussed earlier, when systems mediate public sensemaking about science, policy, and civic issues, decisions about what information to surface and how to structure it carry significant responsibility. Future work should address who manages these systems, how design decisions are made transparent, and what mechanisms ensure accountability when computational processes shape public understanding. This includes exploring models for multi-stakeholder oversight, establishing standards for transparency in AI-assisted sensemaking tools, and developing frameworks for assessing the societal impacts of complexity-oriented design at scale.

Chapter 8. Conclusion

This dissertation challenged a common assumption in information design: that complexity must be eliminated to support understanding. Instead, it proposed that complexity can serve as a medium for deeper engagement when made visible and explorable. As information systems increasingly mediate how people understand science, policy, and public discourse, the question becomes not just what information to present but how to preserve the structures that matter for grounded judgment.

Through three interactive systems—ReviewAid, PRISM, and PolicyScope—I introduced a framework centered on reliability, heterogeneity, and contingency as key dimensions of information complexity. Reliability addresses how claims are supported and where uncertainties exist. Heterogeneity captures how interpretations and impacts differ across perspectives. Contingency reveals how outcomes depend on specific conditions and contexts. Making these dimensions visible through interactive scaffolding helped users develop more grounded understanding, greater awareness of nuance, and increased confidence in their assessments.

The systems demonstrated that different dimensions call for different modes of interaction. ReviewAid scaffolded retrospective evaluation, allowing users to trace health news claims back through media representation to underlying research. PRISM enabled real-time expression, preserving interpretive diversity at the moment reactions are formed. PolicyScope facilitated prospective exploration, helping users investigate how policy impacts vary across stakeholders and conditions. Across these contexts, participants did not report feeling overwhelmed by complexity but rather felt better equipped to engage with it thoughtfully.

However, this work also revealed important tensions and open questions. The appropriateness of complexity-oriented design varies by domain stakes, user motivation, and situational context. Systems that make complexity visible do not eliminate bias or ensure perfect understanding. They provide scaffolding for users to question and scrutinize what they encounter. As these systems move toward real-world deployment, questions of validation, governance, and accountability become critical.

Ultimately, this dissertation offers not a complete solution but a reframing of the problem. Rather than asking how to simplify complexity, it asks how to make complexity understandable and explorable. In domains where understanding matters for informed participation and decision-making, designing for complexity offers a path toward more thoughtful engagement with information in an increasingly mediated landscape.

Bibliography

[1] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[2] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5, pp. 2–4, McLean, VA, USA, 2005.

[3] H. A. Simon, *Models of bounded rationality: Empirically grounded economic reason*, vol. 3. MIT press, 1997.

[4] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[5] G. Schwitzer, "How do us journalists cover treatments, tests, products, and procedures? an evaluation of 500 stories," *PLoS medicine*, vol. 5, no. 5, p. e95, 2008.

[6] N. Oreskes, "Why trust science?," 2021.

[7] B. Fischhoff, "The sciences of science communication," *Proceedings of the National Academy of Sciences*, vol. 110, no. supplement_3, pp. 14033–14039, 2013.

[8] C. R. Sunstein, "Republic: Divided democracy in the age of social media," 2018.

[9] Z. Tufekci, *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.

[10] M. T. Boykoff and J. M. Boykoff, "Balance as bias: Global warming and the us prestige press," *Global environmental change*, vol. 14, no. 2, pp. 125–136, 2004.

[11] C. F. Manski, *Public policy in an uncertain world: analysis and decisions*. Harvard University Press, 2013.

[12] Y. M. Baek, M. Wojcieszak, and M. X. Delli Carpini, "Online versus face-to-face deliberation: Who? why? what? with what effects?," *New media & society*, vol. 14, no. 3, pp. 363–383, 2012.

[13] G. Gigerenzer and D. G. Goldstein, "Reasoning the fast and frugal way: models of bounded rationality.," *Psychological review*, vol. 103, no. 4, p. 650, 1996.

[14] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological science in the public interest*, vol. 13, no. 3, pp. 106–131, 2012.

[15] A. M. Van Der Bles, S. Van Der Linden, A. L. Freeman, J. Mitchell, A. B. Galvao, L. Zaval, and D. J. Spiegelhalter, "Communicating uncertainty about facts, numbers and science," *Royal Society open science*, vol. 6, no. 5, p. 181870, 2019.

[16] B. Fischhoff and A. L. Davis, "Communicating scientific uncertainty," *Proceedings of the National Academy of Sciences*, vol. 111, no. Supplement 4, pp. 13664–13671, 2014.

[17] D. A. Stone, *Policy paradox: The art of political decision making*. WW Norton & company, 2022.

[18] C. Brick, A. L. Freeman, S. Wooding, W. J. Skylark, T. M. Marteau, and D. J. Spiegelhalter, "Winners and losers: communicating the potential impacts of policies," *Palgrave Communications*, vol. 4, no. 1, pp. 1–13, 2018.

[19] M. Vuckovic and J. Schmidt, "On sense making and the generation of knowledge in visual analytics," *Analytics*, vol. 1, no. 2, pp. 98–116, 2022.

[20] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The craft of information visualization*, pp. 364–371, Elsevier, 2003.

[21] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information visualization*, vol. 7, no. 2, pp. 118–132, 2008.

[22] C. A. Ntuen, E. H. Park, and K. Gwang-Myung, "Designing an information visualization tool for sensemaking," *Intl. Journal of Human–Computer Interaction*, vol. 26, no. 2-3, pp. 189–205, 2010.

[23] S. S. Alhadad, "Visualizing data to support judgement, inference, and decision making in learning analytics: Insights from cognitive psychology and visualization science," *Journal of Learning Analytics*, vol. 5, no. 2, pp. 60–85, 2018.

[24] N. Vaupotič, D. Kienhues, and R. Jucks, "Complexity appreciated: How the communication of complexity impacts topic-specific intellectual humility and epistemic trustworthiness," *Public Understanding of Science*, vol. 33, no. 6, pp. 740–756, 2024.

[25] H. A. Simon, "The architecture of complexity," in *The Roots of Logistics*, pp. 335–361, Springer, 2012.

[26] K. Hyland, "Hedging in scientific research articles," 1998.

[27] P. Sumner, S. Vivian-Griffiths, J. Boivin, A. Williams, C. A. Venetis, A. Davies, J. Ogden, L. Whelan, B. Hughes, B. Dalton, F. Boy, and C. D. Chambers, "The association between exaggeration in health related science news and academic press releases: retrospective observational study," *BMJ*, vol. 349, 2014.

[28] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cognitive science*, vol. 26, no. 5, pp. 521–562, 2002.

[29] D. C. Mutz, *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.

[30] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, vol. 115, pp. 9216–9221, Sept. 2018.

[31] H. W. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," *Policy sciences*, vol. 4, no. 2, pp. 155–169, 1973.

[32] D. Spiegelhalter, M. Pearson, and I. Short, "Visualizing uncertainty about the future," *science*, vol. 333, no. 6048, pp. 1393–1400, 2011.

[33] A. Gustafson and R. E. Rice, "The effects of uncertainty frames in three science communication topics," *Science Communication*, vol. 41, no. 6, pp. 679–706, 2019.

[34] J. Zhao, Y. Wang, M. V. Mancenido, E. K. Chiou, and R. Maciejewski, "Evaluating the impact of uncertainty visualization on model reliance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 7, pp. 4093–4107, 2023.

[35] M. Tarvirdians, S. Chandrasegaran, H. Hung, C. M. Jonker, and C. Oertel, "Reflection before action: Designing a framework for quantifying thought patterns for increased self-awareness in personal decision making," *arXiv preprint arXiv:2510.04364*, 2025.

[36] R. Soden, L. Devendorf, R. Wong, Y. Akama, A. Light, *et al.*, "Modes of uncertainty in hci," *Foundations and Trends® in Human–Computer Interaction*, vol. 15, no. 4, pp. 317–426, 2022.

[37] J. Reyes, A. U. Batmaz, and M. Kersten-Oertel, "Trusting ai: does uncertainty visualization affect decision-making?," *Frontiers in Computer Science*, vol. 7, p. 1464348, 2025.

[38] M. Bentvelzen, P. W. Woźniak, P. S. Herbes, E. Stefanidi, and J. Niess, "Revisiting reflection in hci: Four design resources for technologies that support reflection," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–27, 2022.

[39] P. Resnick, "Reputation systems: Facilitating trust in internet interactions," 2000.

[40] J. Z. Wang, A. X. Zhang, and D. R. Karger, "Designing for engaging with news using moral framing towards bridging ideological divides," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–23, 2022.

[41] C. A. Lampe, E. Johnston, and P. Resnick, "Follow the reader: filtering comments on slashdot," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1253–1262, 2007.

[42] S. Ma, Q. Chen, X. Wang, C. Zheng, Z. Peng, M. Yin, and X. Ma, "Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making," in *CHI Conference on Human Factors in Computing Systems*, (Yokohama, Japan), 2025.

[43] O. Scheuer, F. Loll, N. Pinkwart, and B. M. McLaren, "Computer-supported argumentation: A review of the state of the art," *International Journal of Computer-supported collaborative learning*, vol. 5, no. 1, pp. 43–102, 2010.

[44] T. Kriplean, J. Morgan, D. Freelon, A. Borning, and L. Bennett, "Supporting reflective public thought with considerit," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pp. 265–274, ACM, 2012.

[45] H. Kim, H. Kim, K. J. Jo, and J. Kim, "Starrythoughts: Facilitating diverse opinion exploration on social issues," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, Apr. 2021.

[46] Y. Jeon, J. Kim, S. Park, Y. Ko, S. Ryu, S.-W. Kim, and K. Han, "Hearhere: Mitigating echo chambers in news consumption through an ai-based web system," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–34, 2024.

[47] L. Shi, H. Liu, Y. Wong, U. Mujumdar, D. Zhang, J. Gwizdka, and M. Lease, "Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates," *arXiv preprint arXiv:2412.04629*, 2024.

[48] A. Singh, M. J. Dechant, D. Patel, E. Soubutts, G. Barbareschi, A. Ayobi, and N. Newhouse, "Exploring positionality in hci: Perspectives, trends, and challenges," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2025.

[49] L. Öhlund and M. Wiberg, "Social justice in hci: Current streams, considerations, and ways forward," *Interacting with Computers*, p. iwaf009, 2025.

[50] M. Conlen and J. Heer, "Idyll: A markup language for authoring and publishing interactive articles on the web," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 977–989, 2018.

[51] L. A. Suchman, *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[52] E. Ko, Y. Kim, and J. Kim, "Reviewaid: A scaffolded approach to supporting readers' evaluation of health news," in *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 313-320*, International Society of the Learning Sciences, 2022.

[53] N. Bakalar, "Sugary drinks linked to cancer onset," 2019, July 10.

[54] K. Doheny, "Breastfeeding may cut breast cancer risk," 2009, August 10.

[55] N. Bakalar, "Even moderate air pollution may lead to lung disease," 2019, July 10.

[56] P. R. Center, "Science news and information today," Sept. 2017.

[57] A. Dudo, "Scientists, the media, and the public communication of science," *Sociology Compass*, vol. 9, no. 9, pp. 761–775, 2015.

[58] C. E. Smith, X. Wang, R. P. Karumur, and H. Zhu, "[un]breaking news: Design opportunities for enhancing collaboration in scientific media production," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), p. 1–13, Association for Computing Machinery, 2018.

[59] S. M. Center, "Science media center: where science meets the headlines," (Accessed: June 01, 2020).

[60] A. A. for the Advancement of Science, "Sciline: Scientific expertise and context on deadline," (Accessed: June 01, 2020).

[61] HealthNewsReview.org, "Healthnewsreview.org: Improving your critical thinking about health care," (Accessed: June 01, 2020).

[62] FactCheck.org, "Factcheck.org: A project of the annenberg public policy center of the university of pennsylvania," (Accessed: June 01, 2020).

[63] S. Feedback, "Science feedback," (Accessed: June 01, 2020).

[64] T. Aitamurto, "Crowdsourcing for democracy: A new era in policy-making," *Crowdsourcing for Democracy: A New Era In Policy-Making. Publications of the Committee for the Future, Parliament of Finland*, vol. 1, 2012.

[65] T. Peixoto, "Beyond theory: E-participatory budgeting and its promises for eparticipation," *European Journal of ePractice*, vol. 7, no. 5, pp. 1–9, 2009.

[66] V. Pandey, J. Debelius, E. R. Hyde, T. Kosciolek, R. Knight, and S. Klemmer, "Docent: Transforming personal intuitions to scientific hypotheses through content learning and process training," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, L@S '18, (New York, NY, USA), Association for Computing Machinery, 2018.

[67] R. Jarman and B. McClune, *Developing Scientific Literacy: Using News Media In The Classroom: Using News Media in the Classroom*. McGraw-Hill Education (UK), 2007.

[68] B. Oliveras, C. Márquez, and N. Sanmartí, "The use of newspaper articles as a tool to develop critical thinking in science classes," *International Journal of Science Education*, vol. 35, no. 6, pp. 885–905, 2013.

[69] J. M. Brechman, C.-j. Lee, and J. N. Cappella, "Distorting genetic research about cancer: from bench science to press release to published news," *Journal of Communication*, vol. 61, no. 3, pp. 496–513, 2011.

[70] S. Dunwoody, "A question of accuracy," *IEEE Transactions on Professional Communication*, no. 4, pp. 196–199, 1982.

[71] R. Bromme, L. Scharrer, M. Stadtler, J. Hömberg, and R. Torspecken, "Is it believable when it's scientific? how scientific discourse style influences laypeople's resolution of conflicts," *Journal of Research in Science Teaching*, vol. 52, no. 1, pp. 36–57, 2015.

[72] S. Dunwoody, "Scientists, journalists, and the meaning of uncertainty," *Communicating uncertainty: Media coverage of new and controversial science*, pp. 59–79, 1999.

[73] H. P. Peters and S. Dunwoody, "Scientific uncertainty in media content: Introduction to this special issue," *Public Understanding of Science*, vol. 25, pp. 893–908, 11 2016.

[74] R. M. Lucas, R. Harris, and M. Rachael, "On the nature of evidence and 'proving'causality: smoking and lung cancer vs. sun exposure, vitamin d and multiple sclerosis," *International journal of environmental research and public health*, vol. 15, no. 8, p. 1726, 2018.

[75] P. N. Lee, B. A. Forey, and K. J. Coombs, "Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer," *BMC cancer*, vol. 12, no. 1, p. 385, 2012.

[76] S. H. Stocking, "How journalists deal with scientific uncertainty," *Communicating uncertainty: Media coverage of new and controversial science*, pp. 23–41, 1999.

[77] M.-È. Maillé, J. Saint-Charles, and M. Lucotte, "The gap between scientists and journalists: the case of mercury science in québec's press," *Public Understanding of Science*, vol. 19, no. 1, pp. 70–79, 2010.

[78] J. Concato, N. Shah, and R. I. Horwitz, "Randomized, controlled trials, observational studies, and the hierarchy of research designs," *New England journal of medicine*, vol. 342, no. 25, pp. 1887–1892, 2000.

[79] I. Boutron, R. Haneef, A. Yavchitz, G. Baron, J. Novack, I. Oransky, G. Schwitzer, and P. Ravaud, "Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit," *BMC medicine*, vol. 17, no. 1, p. 105, 2019.

[80] T. O. Notebook, "The open notebook: The story behind the best science stories," (Accessed: June 01, 2020).

[81] WFSJ, "World federation of science journalists," (Accessed: June 01, 2020).

[82] R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron, "Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news," *PloS one*, vol. 10, no. 10, p. e0140889, 2015.

[83] M. Glick and A. Carrasco-Labra, "Misinterpretations, mistakes, or just misbehaving," *The Journal of the American Dental Association*, vol. 150, no. 4, pp. 237–239, 2019.

[84] P. Smeros, C. Castillo, and K. Aberer, "Scilens: Evaluating the quality of scientific news articles using social media and scientific literature indicators," in *The World Wide Web Conference*, WWW '19, (New York, NY, USA), p. 1747–1758, Association for Computing Machinery, 2019.

[85] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, (New York, NY, USA), p. 324–332, Association for Computing Machinery, 2018.

[86] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[87] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu, "Toward computational fact-checking," *Proc. VLDB Endow.*, vol. 7, p. 589–600, Mar. 2014.

[88] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, vol. 10, no. 6, 2015.

[89] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, (New York, NY, USA), p. 675–684, Association for Computing Machinery, 2011.

[90] S. M. G. Inc., "Snopes," (Accessed: June 01, 2020).

[91] M. M. Bhuiyan, A. X. Zhang, C. M. Sehat, and T. Mitra, "Investigating "who" in the crowdsourcing of news credibility," in *Proceedings of Computation+Journalism Symposium*, C+J '20, ACM New York, NY, USA, 2020.

[92] T. Kriplean, C. Bonnar, A. Borning, B. Kinney, and B. Gill, "Integrating on-demand fact-checking with public dialogue," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, (New York, NY, USA), p. 1188–1199, Association for Computing Machinery, 2014.

[93] E. Maddalena, D. Ceolin, and S. Mizzaro, "Multidimensional news quality: A comparison of crowdsourcing and nichesourcing," in *Proceedings of 6th International Workshop on News Recommendation and Analytics*, INRA 2018, (New York, NY, USA), Association for Computing Machinery, 2018.

[94] N. Hassan, M. Yousuf, M. Mahfuzul Haque, J. A. Suarez Rivas, and M. Khadimul Islam, "Examining the roles of automation, crowds and professionals towards sustainable fact-checking," in *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, (New York, NY, USA), p. 1001–1006, Association for Computing Machinery, 2019.

[95] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, E. Bice, S. Hawke, D. Karger, and A. X. Mina, "A structured response to misinformation: Defining and annotating credibility indicators in news articles," in *Companion Proceedings of the The Web Conference 2018*, WWW '18, (Republic and Canton of Geneva, CHE), p. 603–612, International World Wide Web Conferences Steering Committee, 2018.

[96] A. T. Nguyen, A. Kharosekar, S. Krishnan, S. Krishnan, E. Tate, B. C. Wallace, and M. Lease, "Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, (New York, NY, USA), p. 189–199, Association for Computing Machinery, 2018.

[97] M. D. Greenberg, M. W. Easterday, and E. M. Gerber, "Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers," in *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, C&C '15, (New York, NY, USA), p. 235–244, Association for Computing Machinery, 2015.

[98] K. Luther, J.-L. Tolentino, W. Wu, A. Pavel, B. P. Bailey, M. Agrawala, B. Hartmann, and S. P. Dow, "Structuring, aggregating, and evaluating crowdsourced design critique," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, (New York, NY, USA), p. 473–485, Association for Computing Machinery, 2015.

[99] A. Yuan, K. Luther, M. Krause, S. I. Vennix, S. P. Dow, and B. Hartmann, "Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, (New York, NY, USA), p. 1005–1017, Association for Computing Machinery, 2016.

[100] T. J. Ngoon, C. A. Fraser, A. S. Weingarten, M. Dontcheva, and S. Klemmer, "Interactive guidance techniques for improving creative feedback," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), p. 1–11, Association for Computing Machinery, 2018.

[101] A. Xu, S.-W. Huang, and B. Bailey, "Voyant: generating structured feedback on visual designs using a crowd of non-experts," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 1433–1444, 2014.

[102] Y.-C. G. Yen, J. O. Kim, and B. P. Bailey, "Decipher: An interactive visualization tool for interpreting unstructured design feedback from multiple providers," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), p. 1–13, Association for Computing Machinery, 2020.

[103] N.-C. Wang, D. Hicks, and K. Luther, "Exploring trade-offs between learning and productivity in crowdsourced history," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, Nov. 2018.

[104] N. R. Goulden, "Relationship of analytic and holistic methods to raters' scores for speeches.," *Journal of Research & Development in Education*, 1994.

[105] A. J. Nitko, *Educational assessment of students*. ERIC, 1996.

[106] C. A. Mertler, "Designing scoring rubrics for your classroom," *Practical assessment, research, and evaluation*, vol. 7, no. 1, p. 25, 2000.

[107] S. G. Hoskins, ""but if it's in the newspaper, doesn't that mean it's true?" developing critical reading & analysis skills by evaluating newspaper science with create," *The american biology Teacher*, vol. 72, no. 7, pp. 415–420, 2010.

[108] K. Murcia, "Science in the newspaper: A strategy for developing scientific literacy," *Teaching Science*, vol. 51, no. 1, p. 40, 2005.

[109] C. A. Korpan, G. L. Bisanz, J. Bisanz, and J. M. Henderson, "Assessing literacy in science: Evaluation of scientific news briefs," *Science Education*, vol. 81, no. 5, pp. 515–532, 1997.

[110] H. Vally, "Is this study legit? 5 questions to ask when reading news stories of medical reserach," 2019.

[111] NPR, "7 questions to ask while reading health research," February 13, 2012.

[112] A. White, "How to read health news," 2014, December 23.

[113] N. C. for Complementary and I. Health, "Checklist for understanding health news stories," (Accessed: June 01, 2020).

[114] HealthNewsReview.org, "Our review criteria," (Accessed: June 01, 2020).

[115] C. Korpan, G. Bisanz, T. Dukewich, K. Robinson, J. Bisanz, M. Thibodeau, K. Hubbard, and J. Leighton, "Assessing scientific literacy: A taxonomy for classifying questions and knowledge about scientific research (tech. rep. 94-1)," *Edmonton, Alberta, Canada: University of Alberta, Centre for Research in Child Development*, 1994.

[116] U. Khan, "Tofu might harm memory in elderly," 2008, July 05.

[117] J. S. C. Leung, A. S. L. Wong, and B. H. W. Yung, "Understandings of nature of science and multiple perspective evaluation of science news by non-science majors," *Science & Education*, vol. 24, no. 7-8, pp. 887–912, 2015.

[118] P. Norris, R. Pacini, and S. Epstein, "The rational-experiential inventory, short form," *Unpublished inventory. University of Massachusetts at Amherst*, 1998.

[119] M. Wallace and A. Wray, *Critical reading and writing for postgraduates*. Sage, 2016.

[120] B. Pasian, *Designs, methods and practices for research of project management*. Gower Publishing, Ltd., 05 2015.

[121] E. Forman and C. Cazdan, "Exploring vygotskian perspectives in education," *Learning relationships in the classroom*, pp. 189–206, 1998.

[122] H. Jeong and M. T. Chi, "Knowledge convergence and collaborative learning," *Instructional Science*, vol. 35, no. 4, pp. 287–315, 2007.

[123] J. Roschelle and S. D. Teasley, "The construction of shared knowledge in collaborative problem solving," in *Computer supported collaborative learning*, pp. 69–97, Springer, 1995.

[124] J. Kim, S. Sterman, A. A. B. Cohen, and M. S. Bernstein, "Mechanical novel: Crowdsourcing complex work through reflection and revision," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 233–245, 2017.

[125] D. Retelny, M. S. Bernstein, and M. A. Valentine, "No workflow can ever be enough: How crowdsourcing workflows constrain complex work," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–23, 2017.

[126] D. Haas, J. Ansel, L. Gu, and A. Marcus, "Argonaut: macrotask crowdsourcing for complex data processing," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1642–1653, 2015.

[127] H. Schmitz and I. Lykourentzou, "Online sequencing of non-decomposable macrotasks in expert crowdsourcing," *ACM Transactions on Social Computing*, vol. 1, no. 1, pp. 1–33, 2018.

[128] E.-Y. Ko, E. Choi, J.-w. Jang, and J. Kim, "Capturing diverse and precise reactions to a comment with user-generated labels," in *Proceedings of the ACM Web Conference 2022*, pp. 1731–1740, 2022.

[129] J. B. Walther and J.-w. Jang, "Communication processes in participatory websites," *Journal of Computer-Mediated Communication*, vol. 18, pp. 2–15, Oct. 2012.

[130] J. B. Singer, "User-generated visibility: Secondary gatekeeping in a shared media space," *New Media & Society*, vol. 16, pp. 55–73, Feb. 2014.

[131] E.-J. Lee and Yoon Jae Jang, "What do others' reactions to news on internet portal sites tell us? effects of presentation format and readers' need for cognition on reality perception," *Communication Research*, vol. 37, pp. 825–846, Dec. 2010.

[132] E. Noelle-Neumann, "The spiral of silence a theory of public opinion," *Journal of Communication*, vol. 24, pp. 43–51, June 1974.

[133] G. Neubaum and N. C. Krämer, "Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media," *Media Psychology*, vol. 20, pp. 502–531, July 2017.

[134] T. Zerback, T. Koch, and B. Krämer, "Thinking of others: Effects of implicit and explicit media cues on climate of opinion perceptions," *Journalism & Mass Communication Quarterly*, vol. 92, pp. 421–443, June 2015.

[135] P. Porten-Cheé and C. Eilders, "The effects of likes on public opinion perception and personal opinion," *Communications*, vol. 45, no. 2, pp. 223–239, 2020.

[136] J. B. Walther, J.-w. Jang, and A. A. Hanna Edwards, "Evaluating health advice in a web 2.0 environment: The impact of multiple user-generated factors on hiv advice perceptions," *Health Communication*, vol. 33, pp. 57–67, Jan. 2018.

[137] S. S. Sundar and C. Nass, "Conceptualizing sources in online news," *Journal of Communication*, vol. 51, pp. 52–72, Mar. 2001.

[138] G. M. Masullo and J. Kim, "Exploring "angry" and "like" reactions on uncivil facebook comments that correct misinformation in the news," *Digital Journalism*, vol. 9, pp. 1103–1122, Sept. 2021.

[139] Y. Tian, T. Galery, G. Dulcinati, E. Molimpakis, and C. Sun, "Facebook sentiment: Reactions and emojis," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (Valencia, Spain), pp. 11–16, Association for Computational Linguistics, 2017.

[140] E. M. Sumner, L. Ruge-Jones, and D. Alcorn, "A functional approach to the Facebook Like button: An exploration of meaning, interpersonal functionality, and potential alternative response buttons," *New Media & Society*, vol. 20, pp. 1451–1469, Apr. 2018.

[141] N. J. Stroud, A. Muddiman, and J. M. Scacco, "Like, recommend, or respect? altering political behavior in news comment sections," *New Media & Society*, vol. 19, pp. 1727–1743, Nov. 2017.

[142] N. J. Stroud, E. Van Duyn, and C. Peacock, "News commenters and news comment readers," *Engaging News Project*, pp. 1–21, 2016.

[143] J. Preece, B. Nonnecke, and D. Andrews, "The top five reasons for lurking: improving community experiences for everyone," *Computers in human behavior*, vol. 20, pp. 201–223, Mar. 2004.

[144] B. Nonnecke and J. Preece, "Lurker demographics: Counting the silent," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, (New York, NY, USA), p. 73–80, ACM, 2000.

[145] C. Cheshire, "Selective Incentives and Generalized Information Exchange," *Social Psychology Quarterly*, vol. 70, pp. 82–100, Mar. 2007.

[146] H. Holländer, "A social exchange approach to voluntary cooperation," *American Economic Review*, vol. 80, pp. 1157–1167, Dec. 1990.

[147] S.-W. Huang and W.-T. Fu, "Don't hide in the crowd! increasing social transparency between peer workers improves crowdsourcing outcomes," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, (New York, NY, USA), p. 621–630, ACM, 2013.

[148] K. W. Chan and S. Y. Li, "Understanding consumer-to-consumer interactions in virtual communities: The salience of reciprocity," *Journal of Business Research*, vol. 63, pp. 1033–1040, Sept. 2010.

[149] Y. Y. Mun and Y. Hwang, "Predicting the use of web-based information systems: self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model," *International Journal of Human-Computer Studies*, vol. 59, pp. 431–449, Oct. 2003.

[150] S.-Y. Lee, S. S. Hansen, and J. K. Lee, "What makes us click "like" on facebook? examining psychological, technological, and motivational factors on virtual endorsement," *Computer Communications*, vol. 73, pp. 332–341, Jan. 2016.

[151] T. G. Sharma, J. Hamari, A. Kesharwani, and P. Tak, "Understanding continuance intention to play online games: roles of self-expressiveness, self-congruity, self-efficacy, and perceived risk," *Behaviour & Information Technology*, pp. 1–17, 2020.

[152] T. Aitamurto, "Motivation factors in crowdsourced journalism: Social impact, social change, and peer learning," *International Journal of Communication*, vol. 9, Oct. 2015.

[153] C. Dellarocas, G. Gao, and R. Narayan, "Are consumers more likely to contribute online reviews for hit or niche products?," *Journal of Management Information Systems*, vol. 27, no. 2, pp. 127–158, 2010.

[154] G. Beenen, K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. E. Kraut, "Using social psychology to motivate contributions to online communities," in *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, (New York, NY, USA), p. 212–221, ACM, Nov. 2004.

[155] P. J. Ludford, D. Cosley, D. Frankowski, and L. Terveen, "Think different: Increasing online community participation using uniqueness and group dissimilarity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, (New York, NY, USA), p. 631–638, ACM, 2004.

[156] E. Klein, *Why we're polarized*. New York, NY, USA: Avid Reader Press, 2020.

[157] D. K. Sherman, L. D. Nelson, and L. D. Ross, "Naïve Realism and Affirmative Action: Adversaries are More Similar Than They Think," *Basic and Applied Social Psychology*, vol. 25, pp. 275–289, Dec. 2003.

[158] C. A. Dorison, J. A. Minson, and T. Rogers, "Selective exposure partly relies on faulty affective forecasts," *Cognition*, vol. 188, pp. 98–107, July 2019.

[159] J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan, "Affective polarization, local contexts and public opinion in America," *Nature Human Behaviour*, vol. 5, pp. 28–38, Jan. 2021.

[160] D. J. Ahler and G. Sood, "The parties in our heads: Misperceptions about party composition and their consequences," *The Journal of Politics*, vol. 80, pp. 964–981, July 2018.

[161] T. F. Pettigrew and L. R. Tropp, "A meta-analytic test of intergroup contact theory.," *Journal of Personality and Social Psychology*, vol. 90, no. 5, pp. 751–783, 2006.

[162] Y. Kim, "Does disagreement mitigate polarization? how selective exposure and disagreement affect political polarization," *Journalism & Mass Communication Quarterly*, vol. 92, pp. 915–937, Dec. 2015.

[163] E. Pronin, T. Gilovich, and L. Ross, "Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others.," *Psychological Review*, vol. 111, no. 3, pp. 781–799, 2004.

[164] W. P. Davison, "The third-person effect in communication," *Public opinion quarterly*, vol. 47, pp. 1–15, Jan. 1983.

[165] D. M. McLeod, B. H. Detenber, and W. P. Eveland Jr, "Behind the third-person effect: Differentiating perceptual processes for self and other," *Journal of Communication*, vol. 51, pp. 678–695, Dec. 2001.

[166] E. Kim, D. Scheufele, and J. Y. Han, "Structure or predisposition? exploring the interaction effect of discussion orientation and discussion heterogeneity on political participation," *Mass Communication and Society*, vol. 14, no. 4, pp. 502–526, 2011.

[167] J. N. Druckman and M. S. Levendusky, "What do we measure when we measure affective polarization?," *Public Opinion Quarterly*, vol. 83, pp. 114–122, May 2019.

[168] S. J. Karau and K. D. Williams, "Social loafing: A meta-analytic review and theoretical integration.," *Journal of Personality and Social Psychology*, vol. 65, pp. 681–706, Oct. 1993.

[169] T. Weninger, X. A. Zhu, and J. Han, "An exploration of discussion threads in social news sites: A case study of the reddit community," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 579–583, 2013.

[170] D. Cosley, D. Frankowski, S. Kiesler, L. Terveen, and J. Riedl, "How oversight improves member-maintained communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, (New York, NY, USA), p. 11–20, ACM, 2005.

[171] J. Stray, "Designing recommender systems to depolarize," 2021.

[172] C. S. Taber and M. Lodge, "Motivated skepticism in the evaluation of political beliefs," *American Journal of Political Science*, vol. 50, pp. 755–769, July 2006.

[173] N. Marchal, ""be nice or leave me alone": An intergroup perspective on affective polarization in online political discussions," *Communication Research*, Sept. 2021.

[174] Y. Wang and N. Diakopoulos, "The role of new york times picks in comment quality and engagement," in *Proceedings of the 54th Annual Hawaii International Conference on System Sciences*, HICSS 2021, pp. 2924–2933, IEEE Computer Society, 2021.

[175] D. Yanow, "How does a policy mean?: Interpreting policy and organizational actions," *(No Title)*, 1996.

[176] D. W. Cintron, N. E. Adler, L. M. Gottlieb, E. Hagan, M. L. Tan, D. Vlahov, M. M. Glymour, and E. C. Matthay, "Heterogeneous treatment effects in social policy studies: an assessment of contemporary articles in the health and social sciences," *Annals of epidemiology*, vol. 70, pp. 79–88, 2022.

[177] C. F. Manski, "Communicating uncertainty in policy analysis," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7634–7641, 2019.

[178] S. R. Baker, N. Bloom, and S. J. Davis, "Measuring economic policy uncertainty," *The quarterly journal of economics*, vol. 131, no. 4, pp. 1593–1636, 2016.

[179] S. J. Davis, "Rising policy uncertainty," tech. rep., National Bureau of Economic Research, 2019.

[180] S. L. Popkin, *The reasoning voter: Communication and persuasion in presidential campaigns*. University of Chicago Press, 1991.

[181] R. R. Lau and D. P. Redlawsk, "Advantages and disadvantages of cognitive heuristics in political decision making," *American journal of political science*, pp. 951–971, 2001.

[182] T. Kriplean, J. Morgan, D. Freelon, A. Borning, and L. Bennett, "Supporting reflective public thought with considerit," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 265–274, 2012.

[183] C. Small, M. Bjorkegren, T. Erkkilä, L. Shaw, and C. Megill, "Polis: Scaling deliberation by mapping high dimensional opinion spaces," *Recerca: revista de pensament i anàlisi*, vol. 26, no. 2, 2021.

[184] R. Mushkani, H. Berard, and S. Koseki, "Wedesign: Generative ai-facilitated community consultations for urban public space design," *arXiv preprint arXiv:2508.19256*, 2025.

[185] M. Safaei and J. Longo, "The end of the policy analyst? testing the capability of artificial intelligence to generate plausible, persuasive, and useful policy analysis," *Digital Government: Research and Practice*, vol. 5, no. 1, pp. 1–35, 2024.

[186] L. Yun, S. Yun, and H. Xue, "Improving citizen-government interactions with generative artificial intelligence: Novel human-computer interaction strategies for policy understanding through large language models," *PLoS One*, vol. 19, no. 12, p. e0311410, 2024.

[187] M. Wang, E. Colby, J. Okwara, V. Nagaraj Rao, Y. Liu, and A. Monroy-Hernández, "Policypulse: Llm-synthesis tool for policy researchers," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2025.

[188] A. Rajiv and K. Mueller, "Legiscout: A visual tool for understanding complex legislation," *arXiv preprint arXiv:2510.01195*, 2025.

[189] A. Kang, C. Li, and S. Meng, "The impact of government budget data visualization on public financial literacy and civic engagement," *Journal of Economic Theory and Business Management*, vol. 2, no. 4, pp. 1–16, 2025.

[190] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "Autotutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Transactions on Education*, vol. 48, no. 4, pp. 612–618, 2005.

[191] C. E. Wieman, W. K. Adams, and K. K. Perkins, "Phet: Simulations that enhance learning," *Science*, vol. 322, no. 5902, pp. 682–683, 2008.

[192] R. Martinez-Maldonado, V. Echeverria, G. Fernandez Nieto, and S. Buckingham Shum, "From data to insights: A layered storytelling approach for multimodal learning analytics," in *Proceedings of the 2020 chi conference on human factors in computing systems*, pp. 1–15, 2020.

[193] S. A. Paul and M. R. Morris, "Cosense: enhancing sensemaking for collaborative web search," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1771–1780, 2009.

[194] C. Berret and T. Munzner, "Iceberg sensemaking: A process model for critical data analysis," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[195] M. Bhat and D. Long, "Designing interactive explainable ai tools for algorithmic literacy and transparency," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 939–957, ACM, 2024.

[196] C. Mancini and S. J. B. Shum, "Modelling discourse in contested domains: A semiotic and cognitive framework," *International Journal of Human-Computer Studies*, vol. 64, no. 11, pp. 1154–1171, 2006.

[197] G. E. Hein, "Constructivist learning theory," *Institute for Inquiry*, vol. 14, 1991.

[198] C. T. Fosnot, *Constructivism: Theory, perspectives, and practice*. Teachers College Press, 2013.

[199] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, *et al.*, "Scratch: programming for all," *Communications of the ACM*, vol. 52, no. 11, pp. 60–67, 2009.

[200] A. Sarkar, "Constructivist design for interactive machine learning," in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pp. 1467–1475, 2016.

[201] D. Schulz, D. Alkountar, M. Dawod, and D. Unbehaun, "Designing for engagement and immersive learning through augmented reality: A participatory design case study of virtual chemist app in an educational context," in *Companion Proceedings of the 2025 ACM International Conference on Supporting Group Work*, pp. 22–28, 2025.

[202] European Commission, "Impact assessment guidelines." European Commission Staff Working Document, 2009. SEC(2009) 92.

[203] OECD, "Introductory handbook for undertaking regulatory impact analysis (ria)," *OECD, Tech. Rep.*, 2008.

[204] P. Sabatier and D. Mazmanian, "The implementation of public policy: A framework of analysis," *Policy studies journal*, vol. 8, no. 4, pp. 538–560, 1980.

[205] M. Hill and P. Hupe, "Implementing public policy: An introduction to the study of operational governance," 2021.

[206] L. J. O'Toole Jr, "Research on policy implementation: Assessment and prospects," *Journal of public administration research and theory*, vol. 10, no. 2, pp. 263–288, 2000.

Acknowledgment

This doctoral journey was motivated by a desire to move beyond observation and to work on research that addresses real-world problems. During this process, I had the opportunity to explore questions that interested me, think deeply about problems I cared about, and try different approaches in search of answers. I am grateful for this time, which helped me grow into someone who can pursue the problems I care about.

However, this journey was far from easy. Questions about what constitutes a meaningful contribution have stayed with me throughout my doctoral studies and continue to do so even now. These questions became more intense as they intertwined with major changes in my life during the journey. During this time, I got married, had a child, and became a parent. Balancing research and life required far more energy and resolve than I had anticipated. Although these were deeply meaningful and joyful moments in my personal life, this period was a series of challenges.

This journey was made possible by the understanding and support of my advisors, colleagues, friends, and family. I am deeply grateful to all of them, and I am glad to have this opportunity to express my gratitude.

To my advisor, Juho Kim: Thank you for your guidance, support, and patience throughout this long process. Beyond research, you taught me how to approach life—how to accept challenges and grow from them.

To Professor Jeong-woo Jang: Thank you for your thoughtful guidance on my research and for the insightful and enjoyable conversations we shared. Your understanding and support as I became a mother while doing research gave me the courage I needed.

To my committee members, Professors Joseph Seering, Steven Dow, and Saiph Savage: Thank you for your time, encouragement, and thoughtful feedback on this thesis. Your comments and insights improved this work tremendously.

To Professor Jihee Kim: Thank you for giving me my first opportunity in HCI research. That experience shaped everything that followed.

To Professor Gi Woong Choi: Thank you for teaching me to view research through an educational lens and for your patient encouragement during my most challenging times.

To my collaborators on the projects in this dissertation—Yeonsu, Hyuntak, Jean, Eunseo, Jihyun, and Hyunwoo: Thank you for the countless discussions, ideas, and support that shaped this research.

To my first authors—Hyungyu, Sungchul, Hyunwoo, Jeongeon, Ju Hui, and Jihyeong: Thank you for taking on the hardest parts of the work. Thanks to you, I got to enjoy the fun parts.

To my collaborators: Thank you for the chaos, the deadlines, and the unexpected lessons along the way, and for being so kind to me through all of it.

To KIXLAB members: Thank you for making the lab such a fun place to be. I learned so much from all of you—not just about research, but about how to navigate this journey. Your friendship made everything better.

To my SIG-Social friends: Thank you for sharing research interests with me. I genuinely enjoyed our conversations exploring different work in the field. I am especially grateful to Hyunwoo and Yoonseo for patiently enduring countless iterations and questions as I worked toward completing this thesis.

To my parents-in-law and family: I am deeply grateful for the warmth, encouragement, and unwavering support you have always shown me. Being welcomed and supported as part of your family gave me great comfort and strength throughout this journey.

To my sister and brother: Whenever I am with you, I can simply be the middle one again—able to laugh, eat, and rest without any worries. I am grateful for all the time and memories we share.

To my parents: Thank you for your unconditional love and support. Knowing that you were always there for me gave me the confidence to keep going, even during the most uncertain moments. Especially during the later stages of my doctoral studies, your generous help with childcare made it possible for me to complete this journey.

To my husband Jaeyoung: Thank you for always believing in me, even when I doubted myself. I would not have been able to make it through this journey without your encouragement and support. I am so grateful to have you by my side as we move forward together.

To Jiwoo: Thank you for bringing joy into my life every day. Your bright smiles were my greatest source of strength when things got tough. I hope this work contributes, even in a small way, to making the world you will grow up in a better, more thoughtful place.