Restimate: Recovery Estimation Tool for Resilience Planning

Scott Miles[a], Megan Ly[a], Nick Terry[b], Youngjun Choe[b]
[a]Department of Human Centered Design & Engineering, University of Washington, Seattle, United States of America
[b]Department of Industrial & Systems Engineering, University of Washington, Seattle, United States of America

Abstract

The U.S. National Institute of Standards and Technology (NIST) published the Community Resilience Planning Guide in 2016. The NIST Guide advocates for a participatory process for developing a performance measurement framework for the jurisdiction's resilience against a scenario hazard. The framework centers around tables of expected and desired recovery times for selected community assets, such as electricity, water, and natural gas infrastructures. The NIST Guide does not provide a method for estimating the expected recovery times. But, building high-fidelity computer models for such estimation requires substantial resources that even larger jurisdictions cannot cost-justify. The most promising approach to recovery time estimation is to systematically use data elicited from people to tap into the wisdom of the (knowledgeable) crowd. This paper describes a novel research-through-design project to enable the computer-supported elicitation of recovery time series data. This work is the first in the literature to examine people's ability to estimate recovery curves and how design influences such estimation. Its main contribution to resilience planning is three-fold: development of a new elicitation tool called Restimate, understanding of its potential user base, and providing insights into how it can facilitate resilience planning. Restimate is the first tool to enable evidence-based expert elicitation in any community with limited resources for their resilience planning. Beyond resilience planning, those who facilitate high-stakes planning activities under large uncertainties (e.g., mission-critical system design and planning) will benefit from a similar research-through-design process.

Keywords: expert elicitation; disaster; natural hazard; infrastructure; community resilience; restoration; user-centered design; computer-supported cooperative work

Introduction

This study is the first to fill the gap in the literature to our best knowledge, where none has yet studied human ability to estimate curves of recovery from a scenario disruption (see thorough reviews of the relevant literature in [1, 2], which we find up-to-date regarding the gap as of this writing). The main research question is: *how does design influence such scenario-based estimation of curves?* This paper describes our research-through-design project to address the question through five well-established human-centered design steps that lead to an estimation

tool. We share novel insights gained along the steps to inform those who want to design a similar tool.

The academic and practical insights from this study are broadly relevant to anyone who has a similar research question as ours. The following few paragraphs describe a background of how the research question emerged, as specifically motivated by a recent trend of resilience planning initiatives.

The trend originated when SPUR (San Francisco Planning + Urban Research Association) conducted a planning process for earthquake resilience called Resilient City. It focused on the City and County of San Francisco and produced nine reports, published between 2008 and 2013 [[3], [4], [5], [6], [7], [8], [9], [10], [11]]. SPUR Resilient City inspired two state-level earthquake resilience initiatives conducted by the State of Washington Seismic Safety Committee (WASSC) and the Oregon Seismic Safety Policy Advisory Commission (OSSPAC). The Resilient Washington State (RWS) final report and the Oregon Resilience Plan (ORP) were published in 2012 and 2013, respectively [12,13]. The resilience planning approaches developed and refined by the SPUR, RWS, and ORP initiatives were synthesized and expanded on by the U.S. National Institute of Standards and Technology (NIST) Community Resilience Group. NIST published the Community Resilience Planning Guide in 2016 to support other communities to do similar resilience planning related to any type of hazard. At least four jurisdictions have used the NIST guide to conduct their own resilience planning process: San Francisco (again; different agency), Fort Collins (Colorado, USA), Boulder (Colorado, USA), and Nashua (New Hampshire, USA).

The NIST Guide is centered around collaboratively developing a performance measurement framework for assessing and monitoring the jurisdiction's resilience. The framework is a table of expected and desired recovery times for selected community assets, such as retail businesses, schools, and electricity systems, relative to different community needs or scenarios (e.g., emergency needs). The NIST Guide does not provide direction or a method for quantifying recovery times (e.g., estimating the expected recovery times or reaching a consensus on desired recovery times) to create the performance measurement framework. The approaches to recovery time quantification for each of the past conducted initiatives were widely different (from a single expert estimating times for a single system to large groups coming to consensus on times for multiple systems) based on variable quality of data and analysis. Miles compared the SPUR, RWS, and ORP initiatives' measurement frameworks [14]. Normalized estimates for recovery times differ widely across the initiatives in most instances—in some instances irreconcilably so (e.g., more than 1000 days).

Computer models for estimating expected recovery times were not used in any of the mentioned resilience planning initiatives [15]. (Computer models like Hazus-MH were used to quantify hazards and their impacts.) Computer modeling of recovery is a growing area of scholarly research [[2], [16]] but is uncommon in government and private practice. Building high-fidelity computer models requires a substantial number of resources (e.g., time, expertise, and data for validation/calibration) that even larger jurisdictions cannot afford or cost-justify. The lack of NIST guidance on how to estimate recovery times, the ad hoc approach of past initiatives, and state-of-practice of recovery computer modeling warrant development of a systematic approach to recovery time estimation that is feasible to employ.

Considering the limitations described above, the most promising approach to recovery time estimation is the development of statistical inference models based on data elicited from people (sometimes known as expert elicitation). Recently, Cao et al. [1] developed the first statistical inference model based on Gaussian process regression (GPR) for synthesizing elicited time series data into a recovery curve (i.e., constrained continuous function, not just points) with quantified uncertainty. While there has been previous work done on the topic of quantifying scenario construction and resilience assessment [[17], [18]], little work has been done on quantifying recovery trajectories for scenario disasters. The current paper describes a novel research-through-design project to perform the first-of-its-kind computer-supported elicitation of recovery time series data. The elicited data determine Cao et al.'s [1] model inference quality and, in turn, inform the high-stakes decisions for community resilience planning. The tool being researched as part of this work is called Restimate. The primary purpose of Restimate is to facilitate expert elicitation for estimating expected recovery times in the NIST Guide's framework.

The paper takes an unconventional structure in academic literature, based on the human-centered design research process taken for this project: define needs; prototype Restimate; test user experience; iterate prototype; and deploy pilot (Figure 1). The resulting main contribution of the paper is three-fold: development of a novel tool for resilience assessment (i.e., Restimate), understanding of its potential user base, and insights into its usefulness for resilience planning.. The paper concludes with a reflection on the study's contribution, recommendations for Restimate's use, discussion of limitations, and suggestions for future work.

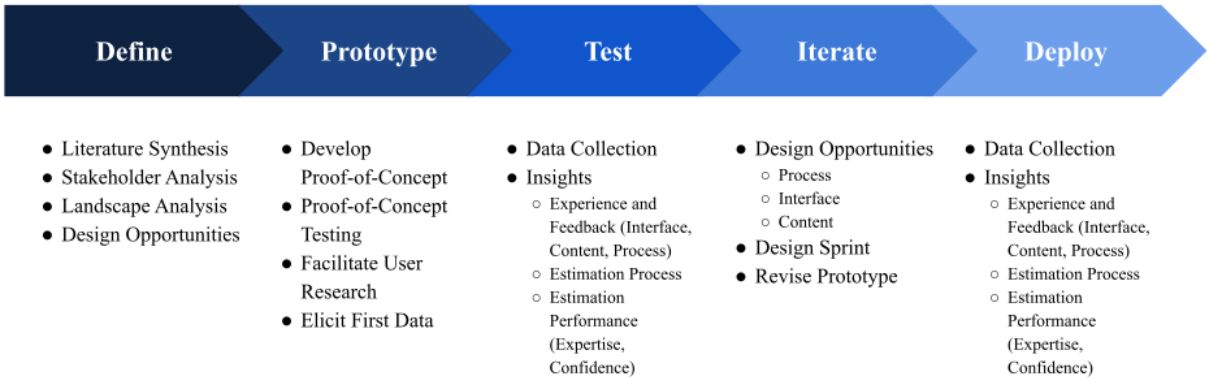

Figure 1. The overview of the human-centered design research process in this study, which is detailed in the remainder of this paper.

Define

Design process requirements for Restimate were defined through literature synthesis, stakeholder analysis, and landscape analysis.

Literature Synthesis

Here we briefly summarize our literature synthesis for design considerations. Our primary focus was on peoples' ability to estimate quantities and probabilities. The literature on human judgement and bias provided evidence-based techniques to mitigate potential biases [[19], [20], [21], [22], [23], [24], [25], [26]]. For example, anchoring bias (tendency to anchor on a starting value and insufficiently adjust the judgement afterwards [[27], [28]]) is typically countered by eliciting extreme values first before central values. Overconfidence bias (tendency to overly trust one's own judgment [29]) is often mitigated effectively by taking structured analyses such as the event tree and scenario planning [30]. We found few studies related specifically to the *use* of software to support elicitation. On this topic, Baker et al. [31] ran multiple experiments. Most relevant, they found no significant difference between computer-supported remote elicitation and in-person elicitation. Cao et al. [1] reviewed the scarce literature on expert elicitation for disaster recovery estimation to motivate the first statistical model of a recovery curve using expert-elicited data. The model's performance sensitivity analysis informs the trade-off between the increased model performance from additional data and the resulting increase in logistical burden of expert elicitation.

Stakeholder Analysis

Stakeholder analysis is essential in any human-centered design process to understand how the design will affect all involved. This study accomplished it by conducting interviews of facilitators of past resilience planning initiatives and expert elicitation processes, and by synthesizing resilience planning reports. Several relevant themes were revealed. For resilience planning and recovery estimation, expertise is difficult to define and evaluate. The topics are highly interdisciplinary and include professionals from academia, government, non-profit organizations, and for-profit companies. Further, survivors of past disasters may have more understanding of recovery than so-called experts that have never experienced, researched, or worked a large-scale disaster (common in the emergency management and engineering profession). Knowledge gaps between experts can be profound, with different theoretical, empirical, and methodological familiarity of the diversity of recovery concepts. Regardless of expertise, there can be a wide range of numeracy and familiarity or comfort with statistics and data visualization. There is no straightforward way to measure expertise about recovery. At the same time, it is unlikely that any expert could accurately recall recovery curves of past disasters. Specialized experts can question the value and validity of eliciting estimates and resist participation or not take the estimation seriously. These experts can have difficulty coming to consensus on estimates through discussion, including with a facilitator. Alternatively, such experts can also anchor to what the most senior or specialized experts on a particular topic say. The opposite is possible that an expert questions their own ability to provide an estimate because of how specialized domain knowledge and experience can be. There is not shared terminology across different disciplines and professions, potentially making content design challenging.

Landscape Analysis

We analyzed the landscape of avaialble software to decide on which one to use for the Restimate prototype development. Software exists for eliciting data to estimate unknown point values and probability distribution parameters. These software typically include data entry fields for estimators to enter quantities like minimums, maximums, and most likely values. Some software shows visualizations (e.g., of probability distributions) based on input values. Some software has

features to assist facilitators in viewing, aggregating, and analyzing estimates. Devilee and Knol [32] surveyed elicitation software available at the time. None of the surveyed software include features for eliciting time series data or functions. Further landscape analysis did not reveal software with such features made available since their survey.

We extended the landscape analysis to software developed for other data entry and visualization purposes to gauge the potential for repurposing or design inspiration. The closest class of software to elicitation needs are the variety of common spreadsheet software with plotting functions and less common plot digitization software, such as the open-source tool WebPlotDigitizer. Adjacent software includes data visualization dashboard tools, such as Plotly Dash.

We chose to develop the first Restimate prototype in Google Sheets. The choice was based on potential for rapid implementation, user familiarity, ease of sharing prototype versions for testing, and collaboration features. No other online or desktop spreadsheet software met all these criteria.

Design Opportunities

We defined the design space to explore for developing Restimate by distilling the above insights into a thematic list of guiding questions. The questions were not meant as scholarly research questions but prompts for uncovering insights and opportunities that will have the greatest impact on the ultimate design of Restimate. We clustered the design questions into three themes: 1) expertise influence, 2) content influence, and 3) process influence.

Understanding the expertise influence is key to understanding who potential users of Restimate not just will be but can be. How much of an "expert" do users need to be to make useful recovery curve estimates? (Statistically speaking, the usefulness of the estimates depends on how fast the estimates from the population of users coverge to the mean recovery curve that is representative of the best human knowledge of the unknown true recovery curve [36]. This convergence rate is faster when the population is more knowledgable about the subject. Thus, we will discuss the statistically significant difference between different populations, e.g., professionals vs. non-professionals of the subject, in later Sections to assess the importance of topical expertise. Practically speaking, the estimates are useful if they yield the mean recovery curve that is usable to inform resilience planning; i.e., the planners should be able to trust the estimated recovery curve as representative of the best possible estimation with limited resources.) What degree of expertise about what topics is needed to provide useful inputs? Does expertise require deep familiarity with geography and topical specialization? Alternatively, do general knowledge and experience allow for useful estimates across geographies and systems? Are there some topics of recovery that require more expertise and specialized than others? The less specialized and narrow expertise needs be, the larger and more diverse pool of potential estimators (i.e., users) there are. This makes it easier to solicit user participation but potentially more challenging to design for the corresponding user diversity.

Recovery and resilience intersect a wide range of complex and intersecting concepts. Quantifying and plotting these concepts require not just quantitative reasoning, but intellectual abstraction, humility, and trust. Relevant to this context, few users will possess all these traits to a high degree because even fewer users will have engaged in this or a similar estimation activity about recovery and resilience. So, what content presented in what ways can encourage and

empower users to successfully provide useful estimates? What background information about recovery is needed? How much detail should be shown about the topics, geographies, and systems that are the objects of estimation? How best can users be instructed to develop their estimates? What biases might content introduce or mitigate?

Estimation elicitation can be facilitated via many processes and modes of interaction. To what degree should the design support a process versus embody the process? What process affords the greatest time and cost efficiency that can be feasibly supported by Restimate design choices? Are there measurable differences in estimates from a group of users versus alone? Do synchronicity and human facilitation affect users' ability to make useful estimates?

Prototype

The context and problem definition described in the preceding section drove the content and design decisions for a prototype of Restimate (i.e., second step in Figure 1). The team rapidly developed the prototype in Google Sheets. The overall intent of the prototype was to conduct proof-of-concept testing, facilitate user research, and elicit the first data set of recovery curve estimates for use with the GPR model described above. The prototype consists of four pages (Figure 2). Users navigate the pages left to right via tabs at the bottom of the spreadsheet. This arrangement is meant to mimic sequential navigation of web-app pages. The first page is a sign-in page (not shown in the figure).

The second page provides background information about restoration of infrastructure after disasters. For the sake of testing, the focus of recovery was narrowed down to the restoration of electricity, drinking water, and natural gas systems, colloquially known as lifeline infrastructures. The aim of the background page is to provide users enough understanding to comprehend the subsequent pages and better inform their estimates. Other than descriptive text, two graphics are included in the page—each taken from an existing publication on the topic. The first graphic illustrates the process of restoring an electric system. The second graphic shows restoration curves for electricity, water, and natural gas after two past disasters.

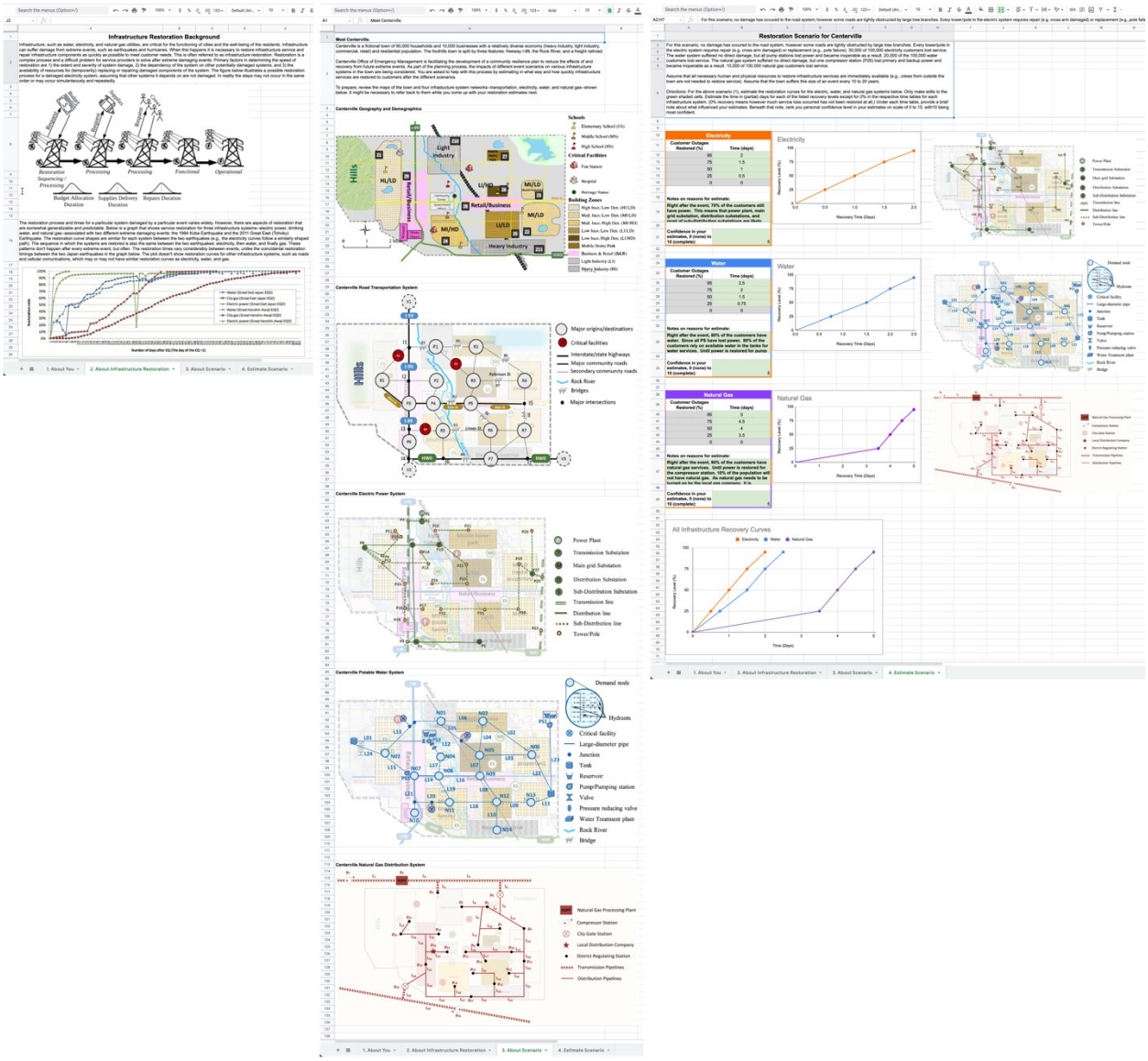

Figure 2. The pages of the Restimate prototype created on Google Sheets. From the left: background information, location information, and estimation activity page.

The third page introduces the location of the infrastructure systems that users will be making restoration estimations for. For the prototype, a fictional location was selected. The location is called Centerville and refers to a simulated city developed by researchers to benchmark computer models of disaster recovery and infrastructure restoration [[33], [34]]. A brief description is provided about Centerville and the estimation activity. Five graphics are included: maps of Centerville's land use, electric grid, water network, natural gas facilities, and roads. The maps were taken from a journal article about Centerville; no modifications were made.

The fourth page supports the estimation activity itself. The text describes a damage scenario for Centerville and its impacts on the electricity, water, and natural gas systems. The text then describes the task being asked of the user and how to interact with the remaining elements of the page. For each infrastructure system, the three primary elements are 1) a table to input estimates, a confidence rating for the estimation, and comments about the estimation; 2) a graph that

dynamically updates with entered estimate values to plot the corresponding restoration curve, and 3) the map of the infrastructure system. A dynamically updating graph is included at the bottom of the page that plots all three of the estimated restoration curves together. This is so users can conveniently compare the relative restoration sequence across the systems.

Test

Data Collection

To test the prototype, we ran fifty-four test sessions with participants directly recruited through professional and student networks of the authors. Session and participant characteristics were chosen to develop holistic insight into the design opportunities described above, particularly about the interface design, participation mode, and user type for Restimate. All sessions were conducted remotely. (In-person sessions were not possible.) We designed testing sessions such that we could statistically assess the effects of important design decisions (i.e., session mode and participant attributes) on their estimation, as shown in Table 1. Each design choice is practically relevant in terms of 1) resource/time requirements for collecting data (e.g., need for a facilitator to schedule synchronous sessions with busy professionals) and 2) topical expertise requirements (e.g., whether the participant pool should be limited to those with profesisonal engineering knowledge).

Table 1. PROTOTYPE TESTING SESSION DEMOGRAPHICS.

Session mode		Professionals		Students
		Engineers	Non-Engineers	
Async	Individual	7	7	18
Sync	Individual	5	1	13
	Group	1*	1*	2

NOTE.–* The one professional group sync session involved one engineer and one non-engineer.

32 of the sessions were conducted asynchronously with individual participants. For these sessions, each participant was emailed a link to an instance of Restimate and instructions to complete the session on their own—navigate the four pages of the prototype and complete the estimation activity. Of the 32 asynchronous sessions, 18 of them were conducted with recruited university students, most of which enrolled in the lead author's department. They represented a control group to allow for assessing the statistical significance of professionals' topical expertise. The other 14 asynchronous sessions were conducted with professionals in various disciplines and organizations who do paid work on aspects of resilience, disasters, natural hazards, or infrastructure restoration. Half (7) of the professionals in the asynchronous sessions were engineers by education. (The assumption was that engineers may be more familiar with infrastructure restoration and comfortable with quantitative tasks.)

Twenty-two sessions were conducted synchronously via video conferencing. 19 of the synchronous sessions were conducted with individual participants including 13 university students and 6 professionals (all engineers except one). The three remaining synchronous sessions were conducted with pairs of participants to approximate a group session: two of the group synchronous sessions were with university students; one was with experts (an engineer and a non-engineer).

Quantitative data includes submitted restoration curve estimates (entered as the number of days to restore at specified recovery levels) and a confidence rating of their estimates for each infrastructure system (1 to 10; 10 highest confidence). Qualitative data includes notes on observations of participants using the Restimate prototype and their solicited feedback. For asynchronous sessions, only notes from participants' emailed feedback were gathered. Additional qualitative data, which was analyzed separately, includes participants' written comments about how they derived their estimates. How these data were analyzed, and the resulting insights are described below.

Insights

Experience and Feedback

Thematic analysis of observation and feedback notes of tool usability tests resulted in nine themes that were considered for iterating Restimate (detailed in the next section). The categories are interaction, content, and process design. Themes across those categories are the following:

1. Interface: a) Layout, b) Interactivity, c) Navigation
2. Content: a) Background, b) Scenario, c) Instructions
3. Process: a) Expertise, b) Synchronicity, c) Participation

Interface

Most participants commented in some way about the density of the layout, preferring to not have so much content presented on each page. Also frequently commented on was the amount of large text blocks used throughout Restimate. Combining the text-heavy nature with the complexity of the subject matter made the experience overwhelming for some and slow for others—primarily for the background information and scenario introduction pages. Most participants who commented on the density and text-heaviness of Restimate suggested more use of hierarchy, headings, and bullet lists. Some suggested the use of icons and more graphics to break up the text. A common request among these participants was to have more organization of information by infrastructure type (i.e., electricity vs. water vs. natural gas). Multiple participants said they did not realize they should scroll down after arriving on a page. They requested that the interface and information be laid out to minimize vertical scrolling—favor more pages over longer pages. A couple participants suggested making information access progressive (e.g., collapsible text or information popups). On the final estimation page, some participants felt that it was not clear where a task began—what was passive information and what was instruction.

Most participants were able to navigate Restimate well enough to complete the estimation activity. No synchronous participant asked for assistance with navigation or commented on it while using Restimate. Some participants were unfamiliar with spreadsheet software, which required another level of familiarization. When asked for feedback, the common comment was how disruptive it is to have to tab back and forth to remind themselves about a piece of information—to lesser extent having to scroll up and down for the same reason. A suggestion was to repeat information in subsequent pages but hide repeated information behind collapsed headings or popups. A handful of comments noted that spreadsheet tabs do not provide a sense of history, progress, and where to go next other than the sequential numbering of the tabs.

The most commented upon feature was the maps of Centerville. Almost all feedback references the maps in some way. Several participants made comments about the maps while doing the test. The maps were not designed specifically for Restimate. This was directly or indirectly obvious to most participants. Few people commented that they found the maps useful for the estimation activity. Interestingly most participants made suggestions about improving the interactivity of the maps, not the content or cartography. About every interactive feature of web maps was suggested at least once—more than is worthwhile to list. Only a couple comments were made to not add interactivity and just reduce the content and simplify the design. Other comments about interactivity were largely about improving the accessibility or mobile friendliness of Restimate. The comments, however, applied to Google Sheets in general.

Content

Many participants noted that within the instructions, terminology needed to be clearer. For participants with more expertise, the comments centered around accuracy or oversimplification of definitions. (Some were deliberate design decisions; some were unintended; some are about contested terminology.) Examples include not specifying whether "water" referred to potable water or wastewater, and conflating terms like operable and functional. For participants with less expertise, the comments were more about definitions not being specific enough or about the metric of restoration.

Fewer comments were made that the tasks needed to be clearer. Some of these comments could be resolved as confusion about terminology. Other comments were about confusion on specific details like being able to specify fractions of days or whether the estimated times are what they predict will happen or what they think needs to happen. A few participants requested more instruction on what they should not do, such as not edit values in cells that were frozen or pre-filled. A handful of suggestions were made to provide instruction on how to use some of the background information (e.g., example restoration graph) or scenario content (e.g., maps) to inform their estimates. These users were not sure if they made the best use of available information and how they could do so.

Most comments and observations about the background page content were regarding the example restoration graph provided. Student participants were more likely to spend time examining the graph. Comments and suggestions came from many participants—students and professionals. A few participants felt that the graph could be better designed—specifically for informing the estimation activity—or be clearer about the purpose of the graph for the activity.

More comments focused on how generalizable the graph was, meaning could they base their estimates on the information in the graph. Several participants—mostly students—noted that they used the graph to derive their estimates. Multiple participants requested that there be graphs for more than just two past disasters, including examples from the United States and hazards other than earthquakes. It was commonly commented that the type of information conveyed by the graph was useful. However, suggestions were made to present that type of information in table format, so it was more analogous to how the estimates were entered.

Other than the example restoration curves, the other most commented upon topic was the definition and detail about the three infrastructure systems. Several participants, mostly professionals, felt that an outline should be included of the different components that make up the three different infrastructure systems. For example, an outline may describe that an electric system includes power plants, transmission towers, transmission lines, high voltage substations, low voltage substations, distribution poles, and distribution lines, with short explanations and practical implications to the activity (e.g., that a power plant takes longer to repair than a distribution pole). A couple of professionals suggested a sequential description of the "steps" necessary to restore the different systems.

A wide variety of comments and observations were made about the scenario description. The most frequent also relates to the interface layout: to simplify the scenario description - not to be conflated with shorter. The next most common comment suggests tying damage descriptions directly to infrastructure system components. For example, if it is noted that the electric system has a specific number of substations, explicitly state how many of those substations were damaged in the scenario. Unfortunately, this level of detail is difficult to obtain without deliberate (e.g., funded) efforts to develop those data—empirical and modeled. A few participants noted more exposition of assumptions and context would be useful.

Almost all participants provided feedback. In sum, the maps are not well designed, not fit-for-purpose, and relatively unused. These comments are about the map content, not the interactivity, which was discussed above. Several participants stated that the maps were confusing and contained too much superfluous detail not relevant to the scenario. A few participants suggested revisions such as designing the maps assuming users' map literacy is low. Other suggestions dealt with basic cartographic design, including better symbology, visual hierarchy, and figure-ground to communicate what is most important to consider specifically for the scenario estimation activity.

Comments specific to what to include in the maps focused most on tying map elements to the written description of the hazard areas, system components, and component damage. For component damage, the challenge of finding data—in this case modeled—is even greater than adding similar content to the written description.

Multiple participants said that maps should not be included at all. They felt that no map could be useful given the high abstraction level of the scenario, estimation, and described data. In other words, no map could provide additional information than what can be represented by textual content. Given the cost and effort to develop the data and a fit-for-purpose design, this is an appealing choice—more so considering the extensive interactivity suggestions noted above.

Process

Feedback and observations from professionals versus students were more similar than different. The most often stated or noted issue was participants' perception that Restimate is a tool for testing their knowledge and ability, rather than to elicit data for input into a statistical algorithm. While making estimates, many participants, regardless of kind, made statements they were not sure they were correct or asked if they were on the right track. Similarly, several asked what to do if they did not know the correct answer and whether it was okay to guess or that was what they had to do. While the average confidence rating across all participants was not particularly low (more below), being asked to rate their confidence often triggered their uncertainty and low level of confidence.

While comments were similar between professionals and students, the potential reason and resulting behavior were not. Of course, students had much less experience and knowledge related to the topic—intentionally so for the project's data collection strategy. Many, but not all, professionals have enough experience and knowledge that the data and state-of-the-art, including with computer simulations, currently make accurate predictions technically difficult or epistemologically impossible. Only professionals expressed resistance to even entering estimates. The only participant to refuse to enter estimates after starting the activity is a professional. Another clear difference observed between professionals and students was that students more frequently used the graphical content of Restimate.

A few synchronous users said the activity was cognitively difficult to complete because of how frequently they are disrupted in their work environment. Some synchronous users said that the content would be less overwhelming if they did not have the imposed time limits on reviewing the content—could review at their own pace. It was synchronous users who most questioned the Restimate content, the estimation activity validity, or their ability to provide correct answers. All but one asynchronous user simply made do and completed the activity without making a related comment in the email follow-up.

Given resource, time, and participant pool constraints, it was not feasible to conduct synchronous group tests representative of the group workshop format used in past resilience planning initiatives or recommended in the NIST Community Resilience Planning Guide. The group sessions for this project only involved pairs of participants. So unfortunately, the insights between group versus individual participation are not comprehensively representative. For example, no participant that expressed severe doubts about the content, activity, or abilities participated in a group session. Thus, we have no insight on whether this resistance would influence the willingness or confidence of other participants to complete the activity. We can say that the group participants enjoyed being able to share and discuss their respective estimates after completing the activity. For consistency, group participants were not allowed to share their knowledge, reasoning, or assumptions prior to making their individual estimates. This of course could be an effective process component, as done in the Delphi approach [35]. Group participants were not allowed to revise their estimates after sharing. (None volunteered that they wanted to.)

Estimation Process

Restimate prompts users to describe their reasoning and assumptions for their estimates for each of the three infrastructure systems presented (electricity, drinking water, and natural gas). All but

three participants provided comments—most three or four sentences; some over ten. We conducted a thematic analysis of these comments. The aim of this analysis is to understand participants' thought processes, particularly in relation to aspects of Restimate's interface, content, and participant attributes. This analysis also provides context for interpreting the quantitative analysis described in the next subsection.

To conduct the thematic analysis, two of the authors independently conducted open coding on the text of all comments by participants about their estimation reasoning and assumptions. The resulting open codes were shared, resolved, and explicitly defined. At that point, one author performed closed coding on the text using the established codes and definitions, followed by the next author. This back-and-forth closed coding iterated for four rounds until neither author saw need to revise the coding of the other. Sixteen codes were developed through this process. The codes and their definitions are shown in Table 2.

Table 2. THEMATIC CODES.

Code	Definition
Components	Specific physical elements of one or more infrastructure systems
Customers	Number of customers without service
Damage	Number of damaged components requiring repair or replacement
Effort	Factors that may influence how long it takes to start or conduct repairs
Examples	Past event or existing information, such as a plan, that influenced their estimate
Experience	Infrastructure provider has had similar restoration experience in the past
Generators	Backup generator used to temporarily power water or gas infrastructure components
Graph	Example graph in the background information
Interdependence	Functional relationship between infrastructure systems
Maps	Use of map or distance between components or components and people derived from map
Non-linear	Estimated curve is clearly intended not to be a straight line or approximate one
Prioritization	Organizational decision to restore an infrastructure system based on some explicit criteria
Quantification	Number and math to explain estimate reasoning
Resources	Whether equipment, supplies, money, and people are readily available
Sequencing	Statement that restoration of certain infrastructure types always finish before others
Topology	Arrangement of the network and components, such as redundancy, linearity, and connectivity

NOTE.–The thematic codes are sorted in alphabetical order.

Table 3 shows a quantitative overview of the codes as applied to the test participants, broken down for all participants, students, all professionals, non-engineer professionals, and engineer professionals. This table offers quantiative insights into the major considerations that went into their estimation reasoning and assumptions. Overall, participants cited a similar number of factors (i.e., codes) in their comments: a median of four factors for students and a median of five factors for professions (regardless of profession). As described below, there are clear differences between engineers and non-engineers. The top two most cited factors across all participants are reference to types of system components (most cited) and systems interdependence, followed with a moderate gap by damage extent and repair effort. The two least cited factors are experience and examples (tied), inching out quantitative reasoning and deterministic sequencing (i.e., natural gas comes after water which comes after electricity). The top two for students are components (most) and interdependence. The two bottom factors for students are examples and non-linear curves (least). Among all professionals, the top two factors are interdependence (most) and effort. The factors experience, sequencing, and graph are tied for the least cited across all professionals. Engineer professionals' top two factors are non-linear curves (most) and interdependence. The engineers' four factors tied for last are graph, maps, sequencing, and experience (one reference each). Non-engineer professionals' three factors tied for the top are interdependence, effort, and components. There are six factors that non-engineer professionals cited only once.

The greatest difference between students' and professionals' estimation thought process is that no students estimated a non-linear curve, which is extraordinary given that there were 99 estimated curves from students. 64% of professionals estimated non-linear curves. 36% of the students anchored their estimate to the example restoration graph. Only 9% of professionals did. 33% of students cited effort compared to 73% of all professionals. 9% of students cited resources, whereas 36% of professionals did.

Likely the most critical difference among professionals is that 77% of engineers estimate non-linear curves, compared to 44% of non-engineers. The greatest difference between engineer and non-engineer professionals was that 11% of non-engineers cited topology and 38% of engineers did. Other major differences are for the factors customers (more engineers) and resources (more engineers). Even though there were notable differences between engineer and non-engineer professionals, there was overall more difference between students and all professionals. In other words, non-engineers professionals approached estimation more like engineer professionals than students.

Table 3. CODE FREQUENCY BY TEST PARTICPANT TYPE.

Code	All (%) n = 55	Students (%) n = 33	Professionals (%) n = 22	Non-engineer Professionals (%) n = 9	Engineer Professionals (%) n = 13
Components	63.64	60.61	68.18	77.78	61.54
Interdependence	60.00	51.52	72.73	77.78	69.23
Effort	49.09	33.33	72.73	77.78	69.23
Damage	45.45	45.45	45.45	44.44	46.15

Prioritization	30.91	27.27	36.36	33.33	38.46
Graph	25.45	36.36	9.09	11.11	7.69
Non-linear	25.45	0.00	63.64	44.44	76.92
Topology	21.82	18.18	27.27	11.11	38.46
Resources	20.00	9.09	36.36	22.22	46.15
Customers	18.18	15.15	22.73	11.11	30.77
Maps	18.18	18.18	18.18	33.33	7.69
Generators	14.55	3.03	31.82	22.22	38.46
Quantification	10.91	6.06	18.18	11.11	23.08
Sequencing	10.91	12.12	9.09	11.11	7.69
Examples	9.09	3.03	18.18	22.22	15.38
Experience	9.09	9.09	9.09	11.11	7.69

NOTE.–Sorted in descending order of the ‘All’ column.

Estimation Performance

The GPR statistical model described above [1] was applied to the times series (restoration curve estimates) provided by participants to generate an overall restoration curve estimate for electricity, water, and natural gas systems. In addition to applying the GPR model to the entire data set, it was applied to the participant sample for various treatment groups. The treatment groups were defined by the following attributes: 1) student vs. professional, 2) synchronous vs. non-synchronous, 3) engineer professionals vs. non-engineer professionals, and 4) high vs. low confidence rating.

The results, visualizations, and insights are extensive (see Supplement Materials’ Round 1 Data Analysis). For brevity, only a subset of results and insights are described here. Table 4 summarizes the statistical tests comparing the treatment groups outlined above. All statistical tests were conducted at the 95% confidence level (α = .05). Since a total of 72 hypothesis tests were conducted, one could expect approximately 3.6 $(72 \times \alpha)$ Type-I errors by chance. In fact, the null hypothesis was rejected for 20 of the tests conducted, indicating that the findings are likely not attributable to statistical error.

Table 4. RESTORATION ESTIAMTES BY TEST PARTICPANT TYPE.

ELECTRICITY RESTORATION						
Participant Type	Restoration Time Mean (Days)	Restoration Time Variance	AUC Mean (Days)	AUC Variance	Confidence Mean (1 to 10)	Confidence Variance

Students	12.59	324.36	7.14	124.17	4.78	5.08
Professionals, All	13.70	145.87	9.03	73.98	5.73	3.35
Professionals, Engineers	15.73	181.35	10.82	92.72	4.99	3.69
Professionals, Non-Engineers	12.04	277.37	6.81	105.61	5.60	4.86
Synchronous	14.56	320.84	8.31	88.69	4.78	5.92
Asynchronous	11.79	199.25	7.54	118.74	5.45	3.39

DRINKING WATER RESTORATION

Participant Type	Restoration Time Mean (Days)	Restoration Time Variance	AUC Mean (Days)	AUC Variance	Confidence Mean (1 to 10)	Confidence Variance
Students	16.18	353.07	8.59	107.02	4.60	4.85
Professionals, All	8.67	63.64	5.07	30.22	6.14	2.22
Professionals, Engineers	11.23	136.96	6.97	69.81	4.90	4.14
Professionals, Non-Engineers	13.96	293.55	7.29	83.84	6.00	4.17
Synchronous	15.72	414.23	8.33	115.59	4.86	5.70
Asynchronous	11.22	117.29	6.30	50.03	5.48	3.19

NATURAL GAS RESTORATION

Participant Type	Restoration Time Mean (Days)	Restoration Time Variance	AUC Mean (Days)	AUC Variance	Confidence Mean (1 to 10)	Confidence Variance
Students	17.40	544.85	8.30	135.65	3.85	4.80
Professionals, All	6.22	42.11	3.30	22.30	5.91	2.47
Professionals, Engineers	10.20	166.31	5.22	49.02	4.41	5.10
Professionals, Non-Engineers	14.04	451.13	6.75	114.17	5.33	4.65
Synchronous	12.71	393.04	5.89	83.93	4.12	6.53
Asynchronous	13.25	369.95	6.70	108.51	5.10	3.22

NOTE.–Significant differences between a pair of participant types are highlighted in blue shade.

Expertise

There is a significant difference in mean estimated restoration time (to 95% restored; not 100% because the "full" restoration is too elusive and uncertain to be realistically estimated. See Cao et al., [1] for more discussion) between professionals and students for water and gas systems. Students estimate much longer restoration times for water and gas than professionals. The difference is not significant for electricity. The same is true for variance in estimated restoration time: significant for water and gas, but not electricity. This may be the case people are more generally familiar with electric systems and power outages, than outages of municipal drinking water networks and regional natural gas systems.

To compare restoration curve *shapes*, we use the area under the curve (AUC), which is a widely used metric of disaster resilience [36]. The unit of AUC is in time (e.g., in days) because AUC is the integration of a function that maps a post-event time (i.e., starting at time 0 at the occurrence of the event) to the recovery level (i.e., post-event system functionality), which is in the unit of percentage with respect to the pre-event system functionality. Figure 3 illustrates AUC. Suppose the estimated restoration time (to 95% restored) is 40 days. Curve A has an AUC close to 40 days because the system functionality bounces back quickly at the beginning although it takes a while to get to 95%. Curve B has an AUC of 20 days, which is 50% of the estimated restoration time because the functionality linearly restores back to 95%. Curve C has an AUC close to 0 days because the functionality stays close to 0% until its sudden jump to 95% towards the end. Note that the interpretation of AUC is *relative* to the total restoration time (i.e., 40 days in this illustration). It is straightforward that Curve A represents a more resilient recovery trajectory than Curve C because the total restoration time is identical. If it differs significantly between curves, it can lead to a significant difference between the *absolute* AUC values between the curves. The AUC value normalized by the total restoration time offers a unitless way to compare cuve shapes (e.g., more non-linear if the normalized AUC is farther from 50%; see Supplement Materials' Round 1 Data Analysis for more details).

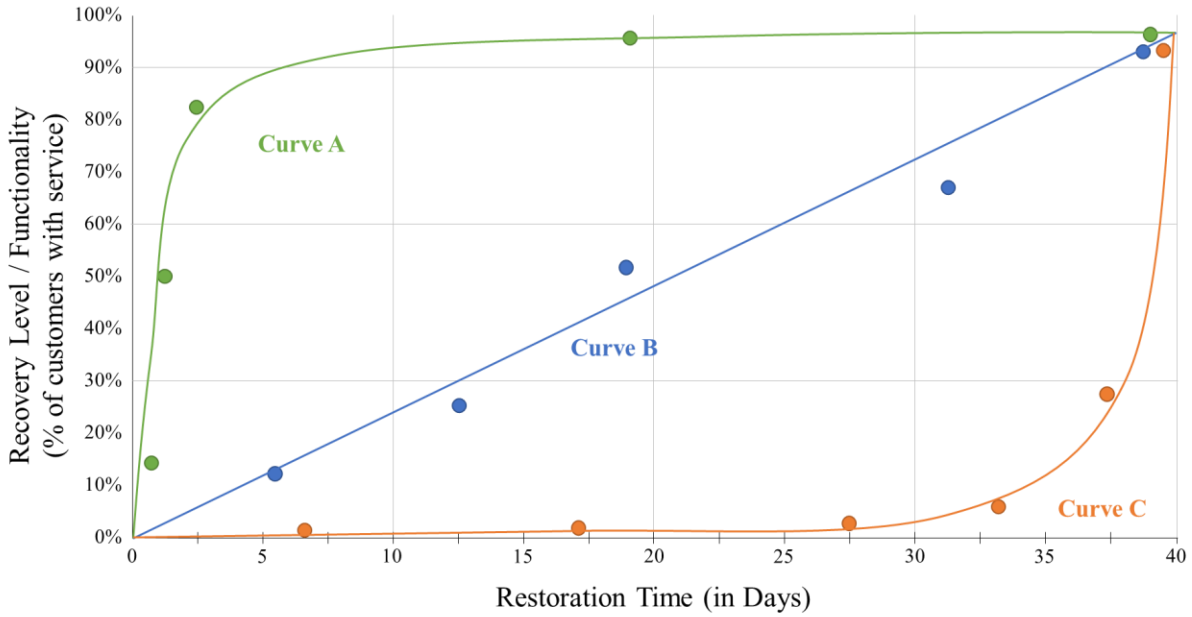

Figure 3. An illustration of the area under the curve (AUC), where Curves A through C represent different recovery trajectories in order of decreasing resilience when the total restoration time is identical across the curves. Each curve illustrates a GPR model fitted to elicited estimates (dots).

The difference in the AUC estimated by professionals versus students is statistically significant for gas (but not for electricity and water). This is likely due to the significant (or, non-significant) difference between the two groups of people regarding the estimated restoration time, as discussed in the previous paragraph. More noteworthy is that the difference in AUC estimate variance between professionals and students is significant for water and gas, but not electricity. Again, this may reflect that people in general are more familiar with power outages and electrical systems because they are more common and visually prominent. Specifically, professionals' estimates for water and gas exhibit less variance than is the case for students. (Professionals' estimate variance for electricity is quite low compared to water and gas.) Professionals are more likely to estimate non-linear restoration curves than students. As noted above, no student estimated a non-linear restoration curve for any system.

Between engineer professionals and non-engineer professionals the difference in mean estimated restoration time is not significant for electricity, water, or gas. The variance in estimated restoration time is only significantly different between engineers and non-engineers for gas. But, engineers' variance is smaller than non-engineers' for all three systems, possibly indicating a gap in relevant expertise. The difference in mean estimated AUC is not significant for electricity, water, or gas, although engineers estimate a relatively larger AUC (when normalized by the estimated restoration time) for all systems than non-engineers. The variances in AUC estimates between engineers and non-engineers are not significantly different. We find that collectively engineer professionals estimate more non-linear (s-shaped) restoration curves for all systems than non-engineer professionals.

There is no significant difference in mean estimated restoration time between synchronous and asynchronous participants. This is true for estimated AUC as well. (These results are like Baker et al. [31] who found no significant difference between online and in-person elicitation based on

multiple measures.) For water, there is a significant difference in the variance of restoration time estimates between synchronous and asynchronous, but not for electricity and gas. (Water variance is lower for asynchronous.) Variances are not significantly different for AUC estimates.

Confidence

The pattern across infrastructure systems holds for confidence rating. The difference in mean confidence rating is not statistically significant between professionals and students for electricity. The difference is significant for water and gas. For water and gas, professionals rate their confidence 1.5 to 2 points (out of 10) higher on average than students. The difference tilts higher towards professionals for electricity, as well. The difference in the variance of confidence rating made by professionals versus students is not statistically significant for electricity, water, or gas. That said, the variance is larger for students than professionals for all three systems.

Between engineer versus non-engineer professionals, there is no significant difference in confidence rating. For all systems, mean confidence is higher by 1 to 1.5 points (out of 10) for engineers. But engineers have significantly higher within-group variance in their confidence rating than non-engineers.

Between synchronous and asynchronous participants, the difference in confidence rating is not significant. This is true for variance in confidence ratings too.

We grouped participants into above median confidence rating ("high confidence") and below median ("low confidence") to potentially understand if self-rating of confidence translates into any differences in estimates. Confidence rating is meaningful. There is a significant difference between mean restoration time estimates of high confidence and low confidence participants on water and gas restoration, but not electricity. The difference in the mean AUC between high and low confidence participants for water and gas is also significant, but not for electricity. (Continuing the pattern those participants do most similarly for electricity restoration estimates.) The same is true for variance in estimates (both restoration time and AUC), with more confident participants exhibiting much less variance in estimates.

Iterate

After testing, another iteration of Restimate was developed. The first round of prototyping served to understand the feasibility of the Restimate concept, test usability, and understand how users might estimate recovery curves. The iterated prototype is intended to investigate the effectiveness of a realistic implementation, which is described in the next section.

Design Opportunities

Design opportunities for iterating Restimate were identified based on the insights from the testing described above. An overview of the more important insights and opportunities are described in the three following subsections.

Process

Estimation performance was similar between synchronous and asynchronous use. While not comprehensively tested, what insight was generated did not suggest great value in group

participation. Thus, we decided to design Restimate for asynchronous uses by individuals. This decision reduces the complexity of the design problem. A design for individual asynchronous work is likely to be equally useful for synchronous work, particularly for individual use. It also reduces the time, money, and effort to use Restimate in the future (e.g., having to schedule synchronous events, particularly with groups). A further benefit of asynchronous individual use is the greater ease in recruiting many participants—even recruiting a few professionals to participate in a synchronous group session could be difficult in many jurisdictions.

There is value in including both engineer and non-engineer professionals. The expertise of these professionals does not have to be highly specific to the infrastructure systems (e.g., electrical engineers for estimating power restoration). In other words, have all participants provide estimates for all infrastructure systems. There are some differences in how the different groups approach estimation and their respective performance. However, these are similar enough to pose a worthwhile design opportunity to promote non-engineer estimates to be more like engineers'. There are larger differences between students and professionals. We chose to continue to explore the value of having "non-expert" participants use Restimate. Improving how non-engineers estimate may also improve how non-experts estimate. There is opportunity to revise Restimate to support non-experts and non-engineers estimate like engineers, particularly by nudging them to estimate non-linear curves. This opportunity matters in practice because of the possibility of eliciting estimates from non-engineers and even non-experts when a resilience planning initiative cannot afford to engage a sufficient number of engineers.

Interface

The most important design opportunity for the interface is to facilitate a simple, more consolidated layout that is skimmable and promotes task-focused navigation. Of course, this opportunity reflects best practices for interface that were not emphasized in the rapid development of the prototype. For example, more hierarchy, structure, graphical elements, and plain language can be used. Progressive disclosure of information, particularly using interactive elements, is another opportunity for improving concision, readability, and navigability of Restimate. This is particularly important because of insights suggesting the need for additional content than already included in Restimate. For improving the usability of maps, there is also a design opportunity for increasing interactivity.

Content

It is worthwhile to explore how to reduce existing content to include additional content needed to meet other design opportunities. The scenario maps are sizable content that could be reduced or even eliminated based on how little participants referred to the maps to make their estimates, even if they did suggest many interactive features. Suggested additional content includes more instructions, more information about restoration factors and curve properties, more instructions, and more detailed system descriptions and scenarios. It was suggested to have redundancy of information across pages to alleviate the need to navigate away from the estimation input fields. Opportunities to reduce the anchoring of estimates to the example graph (for mostly students, but some non-engineers) may increase or decrease content (e.g., provide additional examples, or replace with conceptual information).

Another key opportunity for exploring is reducing user anxiety and resistance in using Restimate. What content might make it not feel like a test and be clear it is for gathering many estimates to synthesize, rather than individual estimates. The most ambitious content-related design opportunity is how to nudge users to approach their estimations more like the engineer professional participants. Thematic analysis of participants' estimation comments revealed that there are some gaps between non-engineer and engineer professionals that may be bridgeable with additional content. While the gap is larger with students, the possibility of improving their estimation performance is encouraging given how straightforward some of their problems were. For example, many students anchored their estimates on the contents of a single graphic.

Design Sprint

The described opportunities were explored through a two-week remote mixed-mode design sprint. Several iterations of low- and moderate-fidelity wireframes were developed. A sample of moderate-fidelity wireframes are shown in Figure 4.

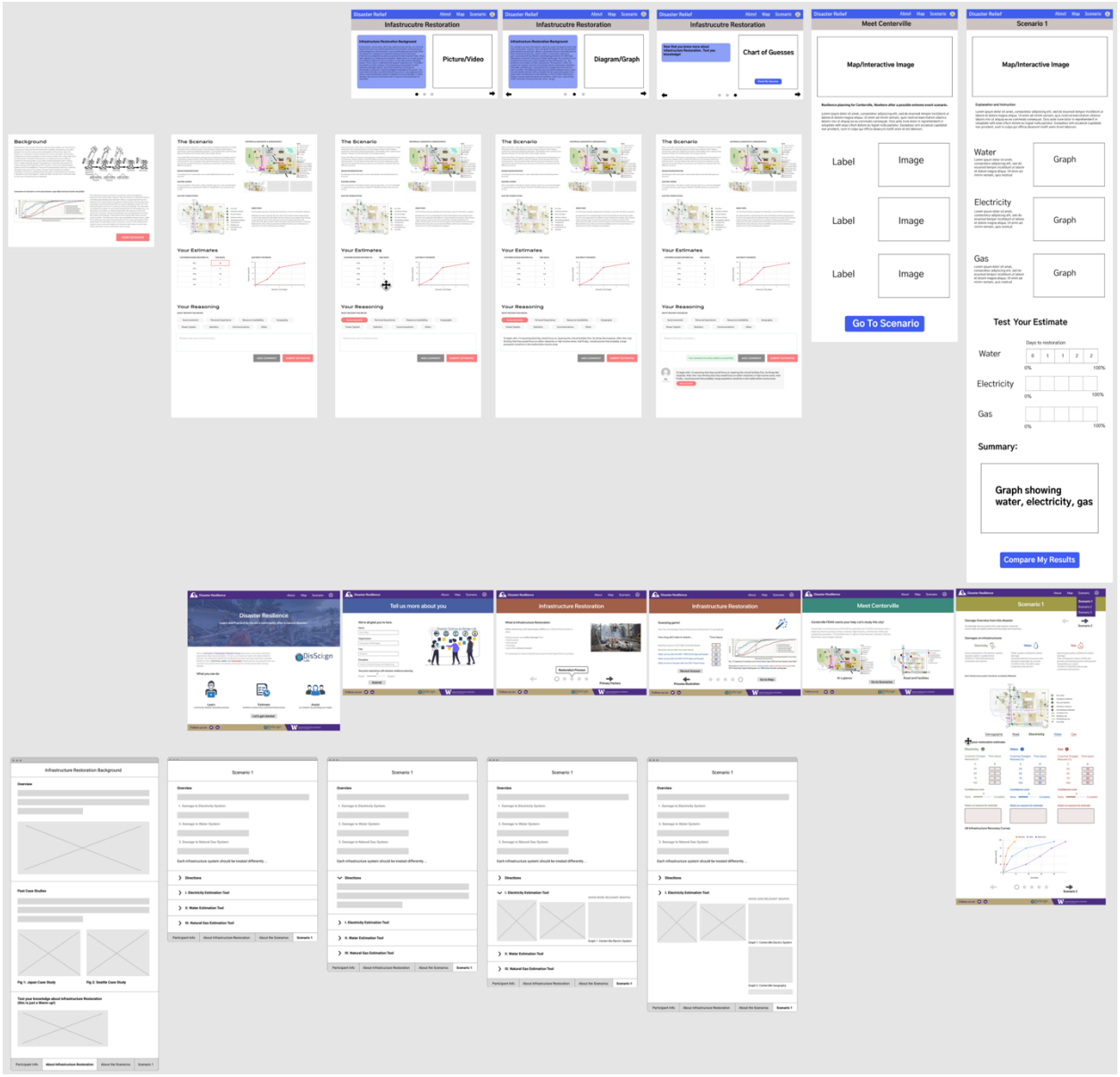

Figure 4. Moderate-fidelity wireframes for the second iteration of Restimate.

Revise Prototype

No user testing of the wireframes was done before iterating Restimate. Two authors performed a crosswalk of the moderate-fidelity wireframes against testing insights, design opportunities, and best practices. From the cross walk, a final wireframe and development environment choice were made to develop the revised prototype. We used a low-code environment called Retool. The choice balanced the selection and quality of interface elements with limited development time and resources. Given the relatively low user base and frequency of use, it may not be worthwhile to explore using a more feature-rich and usable (for end users) development environment. As a WYSIWYG tool, Retool also better facilitates localization and customization by future resilience planning facilitators, who are not likely to have extensive coding skills.

The revised prototype is shown in Figure 5. The pages in the final prototype are: 1) Sign In (not shown), 2) Consent (not shown), 3) Restoration Introduction, 4) Restoration Factors, 5) Scenario Location, 6) Scenario Estimation, and 7) Thank You (not shown). Progressive information disclosure for additional instructions, definitions, and redundant information was provided with modal popups. On-page content reduction was accomplished using plain language, editing for concision, interactive tabbed containers, and collapsible text (toggleable headings).

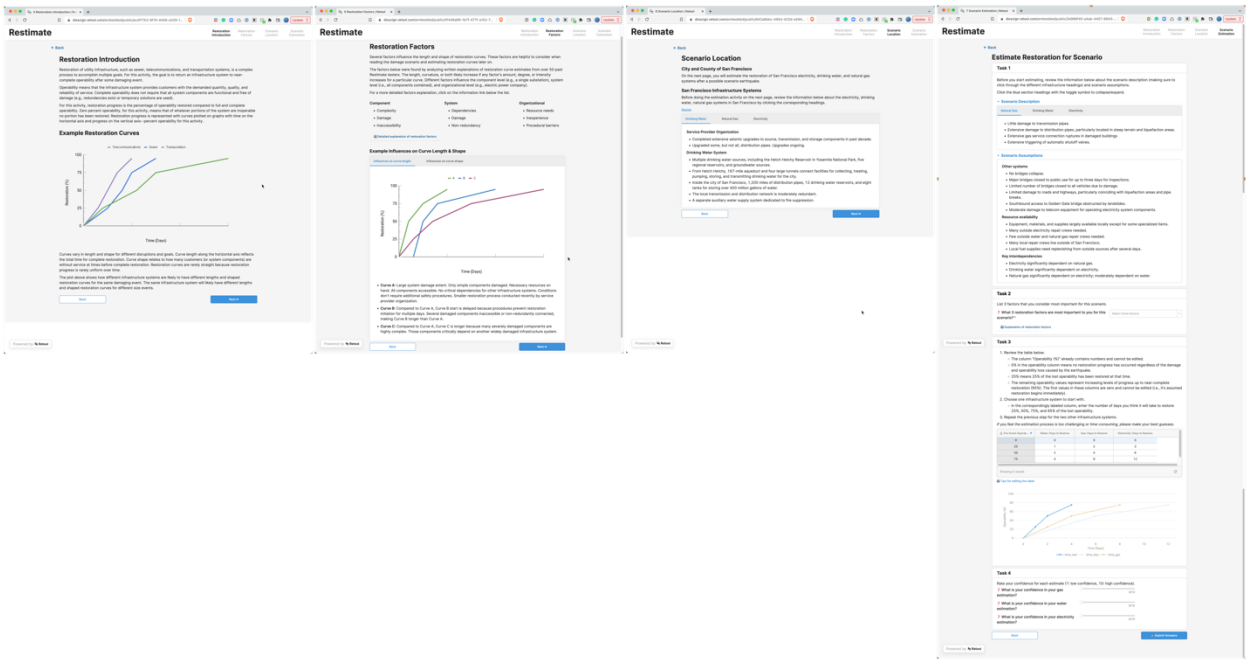

Figure 5. The pages of the revised prototype of Restimate developed on Retool. From the left: Restoration Introduction page, Restoration Factors page, Scenario Location page, Estimation Activity page.

In the previous prototype, information about infrastructure restoration was organized as "background" and "examples." For the final version, the distinction is "introduction" and "factors." The new distinction is aimed at explicitly informing users about factors to consider when estimating curves. Most notably, the example graph and associated text in the example information page were deleted. These were replaced by three graphs and text to conceptually explain restoration curves and the factors that influence them, particularly for making estimates. The example graphs were designed to prevent anchoring, for example not including units on the x-axis or referring to specific infrastructure types. The factors and their explanations were derived from the codes developed during the thematic analysis of participants' estimation comments. Not all the codes were factors that influenced infrastructure restoration (e.g. Non-linear). Some codes were split into two factors to make them more specific. For example, the code "Effort" was split into "Component Access" and "Organizational Procedures" to be clearer what might increase the necessary work effort. The modifiers "component," "system," and "organization" were added to each factor to convey what scale each factor applied to. Definitions of system components were not added because the thematic analysis suggests that participants all had a fit-for-purpose grasp of the definitions already.

The biggest change to the scenario location page is likely that the maps were eliminated. This choice was informed by the effort required to add interactivity, the effort required to redesign

maps, the scarcity of sufficient data to create detailed maps in future resilience planning initiatives, and how little participants used the maps. The fictional jurisdiction of Centerville was replaced with the City of San Francisco. The simple scenario description was replaced by a scenario developed for an actual resilience planning initiative, the San Francisco Lifelines Council's Lifelines Restoration Performance Project (LRPP) [37]. The scenario information was taken from the LRPP final report. This San Francisco scenario is multiple magnitudes more severe, widespread, and complex than the Centerville scenario. This means the corresponding estimation is more challenging.

On the scenario estimation page, input fields for estimates were combined into a spreadsheet-like table. The estimate page layout and content were redesigned to be more integrative of instructions and be task-oriented. The input fields for comments about reasoning and assumptions were eliminated. Instead, a multi-select question was added asking users to specify what factors they considered most important for making their estimates. This question appears before the estimate input fields to prompt users to explicitly consider all factors that may influence curve length and shape prior to making their estimate. Soliciting categorical data rather than free-form text also makes analysis of the data easier for future resilience planning facilitators.

Think-aloud tests of the initial version of the revised prototype were done with three professionals and two students. Multiple decisions mentioned above were made after those tests. The most significant changes from the think-aloud tests were the elimination of maps and "burying" the detailed explanations of restoration factors within a popup glossary. More aggressive copy editing for concision was also done. After the final revisions from the think-aloud tests were made, Restimate was deployed for a pilot application described below.

Deployment

We deployed the revised prototype for a final round of testing. Deployment means that Restimate is being used for a scenario developed for an actual resilience planning process (i.e., LRPP) with participants that either have participated in resilience planning or are professionals qualified to participate. In fact, restoration curves were estimated for electricity, water, and natural gas by the San Francisco Lifelines Council for the LRPP. The curves generated with Restimate and the GPR model can be compared to the LRPP curves. This section is organized the same as the Test section above. Insights about the revised prototype focus on differences and similarities compared to the insights about the initial prototype.

Data Collection

For the pilot deployment of Restimate, we recruited 102 participants. Fifty-five invitees participated in the test of the previous prototype. The other invitees were new. All newly invited participants are professionals with specialized expertise related to disasters or infrastructure restoration. Of the 102 recruits, 30 participated in the pilot. The breakdown of the participants is shown in Table 5. Unlike in the previous testing round, all sessions were done remote asynchronously with individuals only.

Pilot data included the restoration curve time series, quantitative confidence rating, and categorical values associated with the three factors participants identified as most influential to

their estimates. In addition, we did follow-up interviews with five participants, who also tested the original prototype, to gain qualitative user experience insight. Three interviewees are professionals (all engineers) and two are students. Four participants emailed unsolicited feedback. The qualitative data was transcribed and synthesized to distill final design insights and opportunities.

Insights

Experience and Feedback

All interviewees and email comments said that the experience was significantly improved, with many of their painpoints addressed. These participants felt that ease of use, navigation, readability, and language were improved. They felt the progressive disclosure of text was useful (i.e., popups, toggled text, and tabbed text). Several participants positively noted the replacement of the case study graph in the background page with the conceptual graphs and explanations. Overall, many of the design decisions were validated by the feedback.

The most common interface design request was to make information on prior pages more convenient to access on the estimation page. It was clear how to navigate back to that information, but participants felt it disrupted their estimation task to do so. So, there is further opportunity for progressive disclosure of information on the estimation page or affording direct access to prior pages. Complicating that opportunity, most feedback included desires for more detailed scenario information. Some of this information was not possible for the prototype because it was not included in the LRPP report and we were not able to find more detailed information on several issues, like location of predicted damage. In other words, inclusion of the information would depend on generating the information in any future resilience planning initiative. Other information requested was available in the report (e.g., transportation impacts). We had made a general decision to minimize this contextual information to reduce content and cognitive load.

Half of the interviewed participants mentioned that presenting the estimation fields for each infrastructure system in adjacent columns (rather than separate tables) made the task more challenging. This is because the adjacent columns cued them to consider the interdependence of the three infrastructure systems, which is desirable. All but one of the professionals described their hesitance or reluctance to provide estimates. Two invitees who agreed to participate indicated they would not make an estimate upon getting to the estimation page and pulled out of the study. We noted similar feedback from testing the original prototype. The revised prototype included multiple prompts explaining that their estimate would be combined with others and that guessing was acceptable. Regardless, there is still an opportunity to reduce users' hesitancy in making estimates. This might be accomplished with a revised content design. It also might be accomplished by deploying Restimate in synchronous sessions. A concern with group synchronous sessions is a negative spiral of hesitancy when multiple participants express their concerns. That all said, the participants' estimation process and performance were satisfactory and consistent, as described below.

Estimation Process

For the revised prototype, users are not prompted to enter freeform text to describe their reasoning in making their estimates. Instead, they are prompted to choose the three most important restoration factors based on factor explanations in the background pages. So, it was not necessary to do a thematic analysis to analyze what factors were most used by the different user groups.

The primary goal in including factor explanations (and the conceptual graphs) in the background pages was to make the estimation thought process more consistent across user groups. For example, to have students and professionals think about the task the same way. Looking at Table 5, the design choice was successful in meeting the goal. The top three factors chosen by students and professionals are the same (i.e., System Dependencies, Organizational Resources, and System Damage), with only a small quantitative difference in the order. The bottom four factors are also the same (i.e., System Redundancy, Component Complexity, Organizational Procedures, and Organizational Experience), with greater quantitative differences in the order. The differences are somewhat more pronounced between engineer and non-engineer professionals, but still similar.

Table 5. FACTORS CITED BY PARTICIPANTS.

Factor	All (%) n = 30	Students (%) n = 13	Professionals (%) n = 17	Non-engineer Professionals (%) n = 6	Engineers Professionals (%) n = 11	Bay Area Professionals (%) n = 8	Not Bay Area Professionals (%) n = 9
System Dependencies	56.67	69.23	47.06	50.00	45.45	25.00	66.67
Organizational Resources	56.67	61.54	52.94	50.00	45.45	50.00	55.56
Component Damage	46.67	38.46	29.41	66.67	45.45	75.00	33.33
System Damage	43.33	46.15	41.18	33.33	45.45	50.00	33.33
Component Access	33.33	30.77	35.29	33.33	36.36	37.50	33.33
System Redundancy	20.00	7.69	29.41	33.33	27.27	37.50	22.22
Component Complexity	16.67	23.08	11.76	16.67	9.09	0.00	22.22
Organizational Procedures	13.33	7.69	17.65	16.67	18.18	0.00	33.33
Organizational Experience	10.00	15.38	5.88	0.00	9.09	12.50	0.00

NOTE.–Sorted in descending order of the ‘All’ column.

Estimation Performance

The time series data provided by participants were used to estimate an overall restoration curve for electricity, water, and natural gas systems, respectively, using the GPR model. Like above, the GPR model was also applied to multiple treatment groups. The treatment groups were

defined by the following attributes: 1) students vs. professionals, 2) engineer professionals vs. non-engineer professionals, 3) Bay area professionals vs. non-Bay area professionals, and 4) high vs. low confidence rating. Treatment groups were compared based on the length of the estimated curves (i.e., total restoration time) and the area under the curves (i.e., curve shape). Total restoration time was defined as restoration to 95% infrastructure service level.

Description of results focuses primarily on the differences compared to analysis from the original prototype. For brevity, only an overview figure of the visualizations (Figure 6) is included and only a subset of results and insights are described here (see Supplement Materials' Round 2 Data Analysis for more). Table 6 summarizes the statistical tests comparing the treatment groups outlined above.

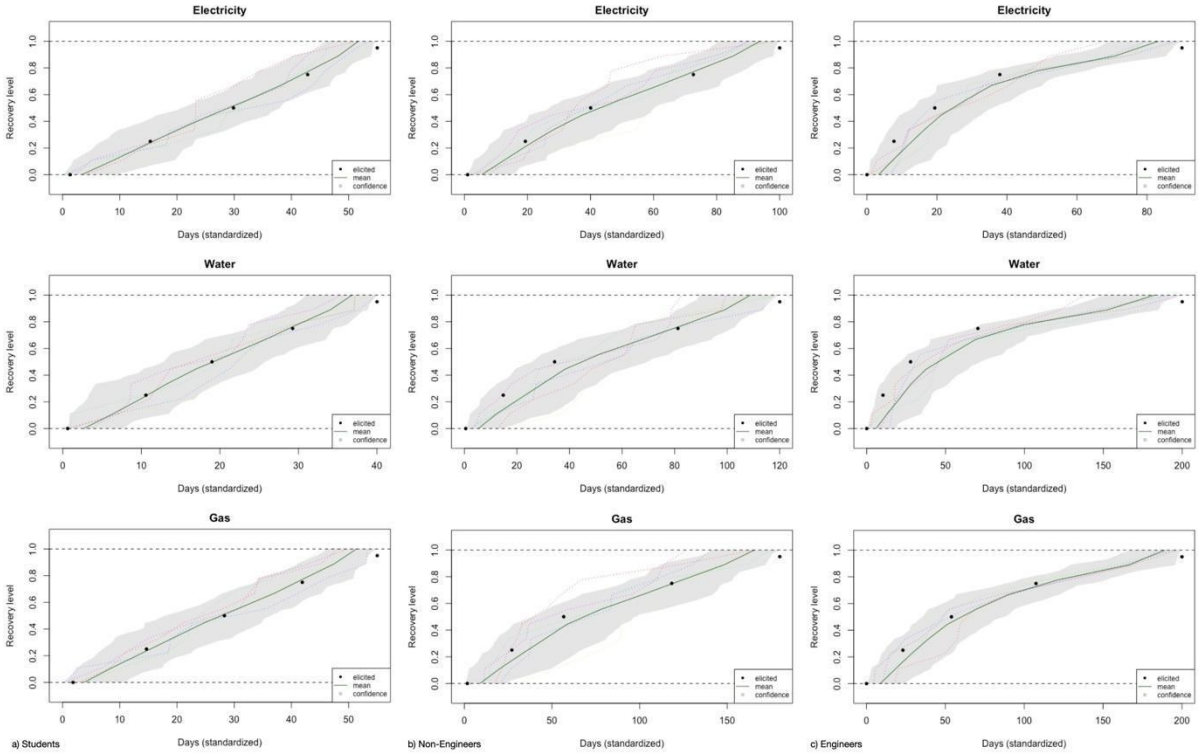

Figure 6. These are the aggregated restoration curves for each infrastructure (row: electricity, water, gas) and for each type of participant (column: students, non-engineers, and engineers).

Table 6. RESTORATION ESTIMATES BY TEST PARTICIPANT TYPE.

ELECTRICITY RESTORATION

Participant Type	Restoration Time Mean (Days)	Restoration Time Variance	AUC Mean (Days)	AUC Variance	Confidence Mean (1 to 10)	Confidence Variance
Students	17.33	231.52	8.08	54.25	5.25	4.02
Professionals, All	42.44	1189.60	27.17	488.43	4.94	4.60

Professionals, Engineers	35.45	908.67	24.10	487.43	5.36	5.05
Professionals, Non-Engineers	29.24	960.44	15.70	317.51	4.88	3.86
Professionals, Bay Area	41.88	1115.55	28.55	611.54	4.75	5.93
Professionals, Not Bay Area	27.60	828.36	15.18	269.67	5.20	3.75

DRINKING WATER RESTORATION

Participant Type	Restoration Time Mean (Days)	Restoration Time Variance	AUC Mean (Days)	AUC Variance	Confidence Mean (1 to 10)	Confidence Variance
Students	10.71	94.11	5.34	23.96	5.83	3.42
Professionals, All	65.25	3635.53	44.95	2027.37	5.06	4.60
Professionals, Engineers	60.82	4374.16	43.98	2618.49	5.27	4.82
Professionals, Non-Engineers	29.62	1607.74	17.62	662.92	5.47	3.89
Professionals, Bay Area	82.88	5723.27	59.88	3351.56	4.88	5.55
Professionals, Not Bay Area	25.47	898.78	15.21	345.75	5.60	3.62

NATURAL GAS RESTORATION

Participant Type	Restoration Time Mean (Days)	Restoration Time Variance	AUC Mean (Days)	AUC Variance	Confidence Mean (1 to 10)	Confidence Variance
Students	17.50	191.18	8.39	57.82	4.50	3.91
Professionals, All	92.88	3355.18	58.59	1293.99	4.00	3.73
Professionals, Engineers	87.82	3410.76	55.58	1187.14	3.91	4.29
Professionals, Non-Engineers	42.94	2739.43	25.10	1203.06	4.41	3.51
Professionals, Bay Area	112.50	4764.29	72.14	1752.27	3.88	4.98
Professionals, Not Bay Area	39.80	1465.12	23.04	594.07	4.35	3.40

NOTE.–Significant differences between a pair of participant types are highlighted in blue shade.

Expertise

Compared to the original prototype, the difference between students' and professionals' estimation performance is even greater. In general, students' estimates remained like their estimates for the Centerville scenario, while the professionals' estimates greatly increased, as expected for such a severe and complicated scenario. There are significant differences between

restoration time and AUC estimates for all three infrastructure systems. There is a significant difference in the variances for those estimates in every case. Interestingly, there is no significant difference between estimate confidence between students and professionals. Lastly, students again only estimated linear curves. This is the case even though an entire page of Restimate is dedicated to explaining why restoration curves are usually non-linear. So, while students and professionals may have indicated similar factors for their estimation thought process (noted in the previous section), this similarity did not translate to the actual estimates.

It appears that for water/electricity the within-group (engineer/non-engineer) mean AUCs are the same at a 95% statistical confidence. For gas there is a significant difference (p-value: 0.033), which is attributable to both 1) the almost significant difference in the mean restoration time estimate (87.82 vs. 42.94 days; p-value: 0.052) and 2) the engineers' restoration curves being more non-linear (i.e., greater AUC). There is no significant difference in the variance for all three infrastructure systems. The results are the same for total restoration time. However, the differences in the means for electricity and water are large (e.g., 61 days vs. 29 days for water) and could make a practical difference in resilience planning. There is not a significant difference in engineer versus non-engineer confidence in their estimates. Like with the students, the design revisions did not result in non-engineers estimating non-linear restoration curves. Even with the detailed explanation of how restoration is typically non-linear, only engineers provided non-linear estimates.

Confidence

Comparing the estimates provided by participants who rated their confidence higher (above median) to those who rated it lower, no significant difference was found for AUC, total restoration time, or corresponding variances. This diverges from the results from the original prototype, which indicated that confidence rating provides some utility. It is possible that the change is from the smaller sample size (30 vs. 54), increased complexity, and reality of the deployment scenario.

Conclusion

Restimate and the statistical model it is designed to elicit data for represent a major leap forward for creating recovery-based resilience measurement frameworks for future resilience planning initiatives based on NIST's Community Resilience Planning Guide. Restimate is the first published tool for eliciting and estimating time series curves of disaster recovery trends. The presented study, prototype, and statistical model were applied to estimation of infrastructure restoration curves. Restimate is technically and mathematically applicable to estimating any recovery indicator. (This does not mean that there are people capable of making useful estimates for all recovery indicators.)

The presented study took a research-through-design approach to understand Restimate and potential users of it. Through an iterative, multi-method design process the study revealed insights into development feasibility and user behavior, while producing an immediately usable web app. This is the first study in the literature that provides insight into people's ability to estimate recovery curves and how design influences those estimates.

We found evidence that estimation can be facilitated effectively with individuals remotely and asynchronously. This mode is cheaper and more convenient than group-synchronous sessions

(particularly done face-to-face). We found that expertise matters for making realistic and precise estimates. Unsurprisingly, students—"non-experts"—do not provide similar estimates to users with more topical expertise. Similarly, but to a lesser degree, professional experts who are not engineers do not provide estimates consistent with engineers who typically have more technical expertise about infrastructure systems. The inconsistency is primarily in the shape of estimated curves. This is a significant issue given the goal of Restimate (to elicit curves) and how resilience is often computed as the area under recovery curves [36]. Consistently, non-engineer professionals estimated a lower area-under-the-curve than engineers for the same estimated recovery times. Lower area is often interpreted as lower resilience.

It is important to note that most of the engineers involved in this study do not have specific expertise about electricity, water, or natural gas infrastructure systems. That we are aware of, only one has professional experience with respect to restoration of all three systems (as a researcher). Three others have professional experience with one system type. The rest (majority) are structural or soils engineers with no professional experience in infrastructure restoration. For estimating electricity, drinking water, and natural gas restoration, being an engineer with expertise in resilience matters. But having experience in infrastructure restoration matters less. Whether this extends to other recovery indicators requires further research (e.g., do estimates for economic recovery require economists with specific recovery experience?).

Restimate's strengths include that its estimates are more informative than the estimates generated during past resilience planning initiatives. Most past initiatives did not estimate recovery curves. The one initiative we are aware of that did (San Francisco Lifelines Council Lifelines Restoration Performance Project–LRPP [37]) generated linearly interpolated curves that are not comparable due to a lack of uncertainty quantification. Curves generated using Restimate are smooth curves with characterized uncertainties that can be mathematically analyzed. Restimate curves are based on significantly more expertise than the LRPP curves (each curve was generated by a different individual), which should increase estimate credibility. However, this study's evidence is limited in informing how these traits of Restimate will actually improve resilience planning as part of the NIST Guide's framework. Ascertaining this requires further research with multiple deployments of Restimate in resilience planning activities. Certainly, curves estimated by appropriate experts using Restimate will not diminish future planning outcomes because of the rigorous process embeded in the tool based on the statistical, expert elicitation, and design literature. Further, eliciting estimates with Restimate is more logistically and financially efficient than the approaches used in the other resilience planning initiatives described above.

Restimate is a better option for supporting future resilience planning unless there is a valid simulation model and suitable data developed for the scenario location (true for few, if any, locations). Our conclusion is based on our study's evidence and considering the relative ease of sufficiently revising Restimate in Retool—Restimate's development environment—using the solicited feedback described above. This is particularly true because Restimate only needs to be usable for users with significant topical expertise, meaning existing pages and content can be eliminated (e.g., background information). Our scenario description may be modified to help experts consider multi-hazard coupling effects (e.g., earthquake's ground shaking along with liquefaction, induced landslides, burst pipes, and fire) [38]. If future resilience planning organizers want to conduct a face-to-face workshop, Restimate can be deployed before or during the workshop. Beyond resilience planning, those who facilitate high-stakes planning activities

under large uncertainties (e.g., mission-critical system design and planning) may leverage the similar research-through-design process described in this paper.

Acknowledgements

The authors gratefully acknowledge the support of the U.S. National Science Foundation (NSF grants CMMI-1824681, BCS-2121616, & CMMI-2211077), NSF proposal reviewers who provided constructive feedback for the initial study design, design sprint participants in a Directed Research Group ("User Testing and Prototyping a Tool for Community Disaster Resilience") at the University of Washington in Spring 2021, Rebecca Jessup for manuscript review and helpful feedback, anonymous reviewers for their constructive comments that substantially improved this manuscript's clarity, and study participants for Restimate prototype tests who generously shared their time and expertise to help develop this tool for resilience planning.

References

[1] Cao, Q. D., Miles, S. B., & Choe, Y. (2022). Infrastructure recovery curve estimation using Gaussian process regression on expert elicited data. *Reliability Engineering & System Safety*, 217, 108054. https://doi.org/10.1016/j.ress.2021.108054

[2] Martell, M., Miles, S. B., & Choe, Y. (2021). Review of empirical quantitative data use in Lifeline Infrastructure Restoration Modeling. *Natural Hazards Review*, 22(4). https://doi.org/10.1061/(asce)nh.1527-6996.0000514

[3] Aldrich, J., Bennett, B., Carroll, S., Favetti, R., Hansen, J., & Morten, D. (2008). The Culture of Preparedness (pp. 1–45). San Francisco, CA: San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2008-06-18/culture-preparedness

[4] Barkley, C. (2009). Lifelines: Upgrading Infrastructures To Enhance San Francisco's Earthquake Resilience (pp. 1–16). San Francisco, CA: San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2009-02-01/lifelines

[5] Bonowitz, D. (2009). The Dilemma Of Existing Buildings: Private Property, Public Risk (pp. 1–27). San Francisco, CA: San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2009-02-01/dilemma-existing-buildings

[6] Hansen, J., Carroll, S., & Aldrich, J. (2008). The Hub Concept: Infrastructure For A Community Disaster Response (pp. 1–15). San Francisco Planning & Urban Research Association. http://www.spur.org/publications/spur-report/2008-10-16/hub-concept

[7] Johnson, L., Barkley, C., Boatwright, J., Brechwald, D., Chung, N.-C., Comerio, M. C., et al. (2013). On Solid Ground (pp. 1–80). San Francisco, CA: San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2013-02-06/solid-ground

[8] Maffei, J. (2009). Building It Right The First Time: Improving The Seismic Performance Of New Buildings (pp. 1–18). San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2009-02-01/building-it-right-first-time

[9] McCann, J., Avetyan, I., Barkley, C., Bruzzone, A., Fisher, L., Stadtfeld, S., and Stokle, B. (2010). After The Disaster: Rebuilding Our Transportation Infrastructure. San Francisco Planning & Urban Research Association. http://www.spur.org/publications/spur-report/2010-07-06/after-disaster

[10] Poland, C. (2009). The Resilient City: Defining What San Francisco Needs From Its Seismic Mitigation Policies (pp. 1–12). San Francisco, CA: San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2009-02-01/defining-resilience

[11] Poland, C., Barkley, C., Boatwright, J., Comerio, M. C., Dickinson, B., Dwelley-Samant, L., et al. (2012). Safe Enough to Stay (pp. 1–44). San Francisco Planning & Urban Research Association. https://www.spur.org/publications/spur-report/2012-02-01/safe-enough-stay

[12] WASSC (2012). *Resilient Washington State: A Framework for Minimizing Loss and Improving Statewide Recovery after an Earthquake*. Olympia, WA: State of Washington Emergency Management Council Seismic Safety Committee. http://mil.wa.gov/other-links/seismic-safety-committee-ssc

[13] OSSPAC (2013). *The Oregon Resilience Plan* (pp. 1–341). Salem, OR: Oregon Seismic Safety Policy Advisory Committee. https://www.oregon.gov/OMD/OEM/Pages/Resilience-Taskforce.aspx

[14] Miles, S. (2018). Comparison of Jurisdictional Seismic Resilience Planning Initiatives. *PLoS Currents*. https://doi.org/10.1371/currents.dis.42c24f29588cb4f887af021449949801

[15] Ganji, A., Alimohammadi, N., & Miles, S. (2019). Challenges in Community Resilience Planning and Opportunities with Simulation Modeling. https://doi.org/10.48550/arxiv.1904.11630

[16] Miles, S. B., Burton, H. V., & Kang, H. (2019). Community of Practice for Modeling Disaster Recovery. *Natural Hazards Review*, 20(1). https://doi.org/10.1061/(asce)nh.1527-6996.0000313

[17] Lu, X., Liao, W., Fang, D., Lin, K., Tian, Y., Zhang, C., Zheng, Z., & Zhao, P. (2020). Quantification of disaster resilience in civil engineering: A Review. *Journal of Safety Science and Resilience*, 1(1), 19–30. https://doi.org/10.1016/j.jnlssr.2020.06.008

[18] Qian, J., & Liu, Y. (2023). Quantitative scenario construction of typical disasters driven by ontology data. *Journal of Safety Science and Resilience*, 4(2), 159–166. https://doi.org/10.1016/j.jnlssr.2022.12.002

[19] Fox, C. R., & Clemen, R. T. (2005). Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior. *Management Science, 51*(9), 1417-1432. http://www.jstor.org/stable/20110430

[20] Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.

[21] Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(1), 3-19. http://www.jstor.org/stable/2988424

[22] Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*(2), 123-141. http://dx.doi.org/10.1016/0010-0277(82)90022-1

[23] Koehler, D. J., & Harvey, N. (2004). *Blackwell handbook of judgment and decision making*: John Wiley & Sons.

[24] Larrick, R. P. (2004). Debiasing. *Blackwell handbook of judgment and decision making*, 316-338.

[25] O'Hagan, A., & Oakley, J. E. (2016). SHELF: the Sheffield Elicitation Framework (version 3.0). *Sheffield, UK: School of Mathematics and Statistics, University of Sheffield.* https://shelf.sites.sheffield.ac.uk/

[26] Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

[27] Winkler, R. L. (1967a). The Assessment of Prior Distributions in Bayesian Analysis. *Journal of theAmerican Statistical Association, 62*(319), 776-800. https://doi.org/10.2307/2283671

[28] Winkler, R. L. (1967b). The Quantification of Judgment: Some Methodological Suggestions. *Journal of the American Statistical Association, 62*(320), 1105-1120. https://doi.org/10.2307/2283764

[29] Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 294--305): Cambridge University Press.

[30] Russo, J. E., & Schoemaker, P. J. (1992). Managing Overconfidence. *MIT Sloan Management Review, 33*(2), 7.

[31] Baker, E., Cruickshank, C., Jenni, K., & Davis, S. (2019). Comparing in-person and online modes of expert elicitation. *Under submission*. 620. https://scholarworks.umass.edu/mie_faculty_pubs/620

[32] Devilee, J. L. A., & Knol, A. B. (2011). Software to support expert elicitation. *RIVM letter*

report 630003001. https://www.rivm.nl/bibliotheek/rapporten/630003001.pdf

[33] Ameri, M. R., & van de Lindt, J. W. (2019). Seismic Performance and recovery modeling of Natural Gas Networks at the community level using building demand. *Journal of Performance of Constructed Facilities*, *33*(4). https://doi.org/10.1061/(asce)cf.1943-5509.0001315

[34] Ellingwood, B. R., Cutler, H., Gardoni, P., Peacock, W. G., van de Lindt, J. W., & Wang, N. (2016). The Centerville Virtual Community: A fully integrated decision model of interacting physical and Social Infrastructure Systems. *Sustainable and Resilient Infrastructure*, *1*(3-4), 95–107. https://doi.org/10.1080/23789689.2016.1255000

[35] Dalkey, N., & Helmer, O. (1963). An Experimental Application of the Delphi Method to the Use of Experts. *Management Science*, *9*(3), 458–467. http://www.jstor.org/stable/2627117

[36] Bruneau, M., Chang, S. E., Eguchi, R. T., Lee, G. C., O'Rourke, T. D., Reinhorn, A. M., Shinozuka, M., Tierney, K., Wallace, W. A., & von Winterfeldt, D. (2003). A framework to quantitatively assess and enhance the seismic resilience of Communities. *Earthquake Spectra*, *19*(4), 733–752. https://doi.org/10.1193/1.1623497

[37] San Francisco Lifelines Council. (2020). *Lifelines Restoration Performance Improvement Plan*. The Office of Resilience and Capital Planning. https://www.onesanfrancisco.org/sites/default/files/inline-files/Lifelines%20Restoration%20Performance%20Report%20Final.pdf

[38] Ba, R., Deng, Q., Liu, Y., Yang, R., & Zhang, H. (2021). Multi-hazard disaster scenario method and emergency management for urban resilience by Integrating Experiment–Simulation–Field Data. *Journal of Safety Science and Resilience*, 2(2), 77–89. https://doi.org/10.1016/j.jnlssr.2021.05.002