Concept Testing Result

Hi Design,

The usability testing result is here!

Please read the high-level summary first. If you are curious of how we get here. You can read a little bit more on the evidences.There are two parts of our results: quantitative and qualitative.

High-level Summary of Usability Testing

The research involved five (5) moderated one-on-one usability tests with members of the Public Health Communications Collaborative (PHCC). Data was collected through Maze (for quantitative data and heatmaps) and Google Meet (for qualitative conversations, recordings, and transcripts).

Two additional sessions that were conducted for usability testing converted to an interview format because:
- participants arrived with their colleagues, and
- had technical difficulties running Maze on their computers.

To gather their insights moderators shared their own screen and asked the participants to walk them through the process of compelling the tasks. TO protect the integrity of the research, the quantitative metrics from these sessions were not used for data analysis. However, the qualitative insights were retained for analysis.

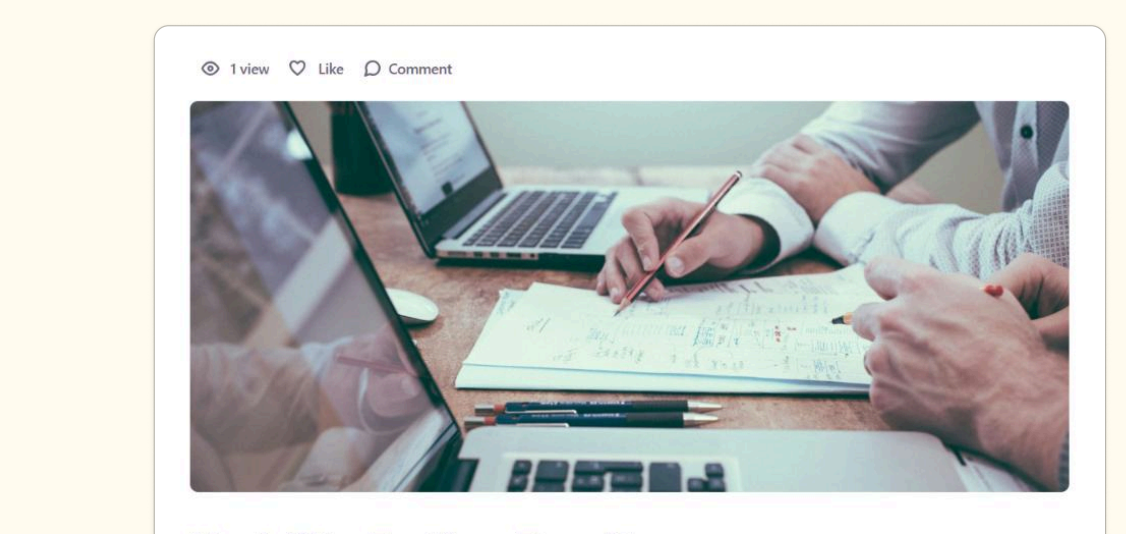
All sessions were transcribed and analyzed in Dovetail for qualitative insights. The qualitative feedback indicated mixed user sentiment. Users generally appreciated:
- the prototype's user-friendly interface,
- clean design, and
- credible sources, highlighting its versatility and efficiency.

However, challenges included
- complexity,
- unclear task expectations,
- navigation difficulties, and
- readability issues on smaller screens.

Recommendations for improvement focus on customization, trending topics, content generation features, language options, cloud storage integration, tutorials, and deeper location-specific data. Users also stressed the importance of human oversight and credible sourcing for AI-generated content.

The quantitative data was captured in Maze and was analyzed in that environment (e.g., heat maps) and Excel. The quantitative results, derived from heatmaps, post-task surveys, the System Usability Scale (SUS), and effectiveness/efficiency metrics, revealed specific areas for improvement.
- Heatmaps highlighted issues with the "use style" button's visibility and user preference for text entry over feature buttons.
- Post-task surveys showed varying perceived difficulty across tasks, with some tasks consistently rated as more challenging.
- SUS scores indicated mixed perceptions of usability, with some users encountering accessibility issues.
- Effectiveness metrics showed a generally low success rate across tasks, particularly for certain tasks, indicating significant usability issues.
- Efficiency data, measured by task duration, pointed to confusion in tasks that took longer to complete, revealing discrepancies between objective efficiency and subjective ease.

Qualitative



Usability Testing Result

We analyzed qualitative data from usability testing in Dovetail. The process was similar to analyzing the interview results. Key differences:
- We developed new tags for analyzing usability tests,
- We kept the udsability test data separate from the rest of the interviews. So, the insights shared here include ONLY the feedback received during usability testing.

Below is the summary of qualitative results. If you would like to read the full Insight in Dovetail with quotes, you can find it here:

Dovetail insight page link here

Qualitative Feedback Summary

The prototype received mixed feedback from users, with both positive aspects and areas for improvement identified.

Users appreciated the prototype's user-friendly interface, clean design, and credible sources. They found it versatile for different roles and valued its transparency regarding risks and regulations. The time-saving aspects and efficiency in tasks like creating press releases were highlighted as benefits.

However, users also encountered usability issues. Some found the tool complex, with unclear task expectations and difficulty navigating the interface. The small font size and readability on smaller screens were problematic for some users.

Recommendations for improvement include:

1. Customization options for better integration with existing workflows

2. Providing trending topics based on user location

3. Features to generate press releases and infographics

4. Options to adjust language and readability levels

5. Integration with cloud storage platforms

6. Offering a demo or tutorial for less tech-savvy users

7. Allowing users to dig deeper into location-specific data

Users emphasized the importance of maintaining human oversight when using AI tools and verifying the accuracy of AI-generated content. They also stressed the need for credible, trustworthy sources and proper citation of information.

Overall, while the prototype shows promise, addressing the identified usability issues and implementing the suggested improvements could significantly enhance its effectiveness and user satisfaction.

Quantitative

Three things to look at:
- Heat map
- Surveys: post-task survey, post-test survey
- Effectiveness & efficiency

1. Heat map

The provided heat maps are based on clicks. We do not have data on cursor movement.
Other than task 1, all heat maps are from participants who did not successfully complete the task.
This is an opportunity for us to investigate their click behavior and identify our usability issues.

Please go through each task and read our sticky notes. But for a quick overview, here's our most significant findings in heat map:
- "use style" button is "invisible" for most participants, or they didn't understand what this button means so it doesn't make sense to click on it.
- It's most intuitive for participants to write prompt in the text entry box to interact with this system (some of them mentioned our tool looked similar to ChatGPT). They would first click on the text entry box, rather than trying to find the feature buttons like "translation", "share".
- The horizontal feature buttons in content studio are not catching people's attention. We saw very few clicks on any of those buttons.

One participant was very enthusiastic and eager to explore. He suggested trending topics should be based on area/location. #reccomendation

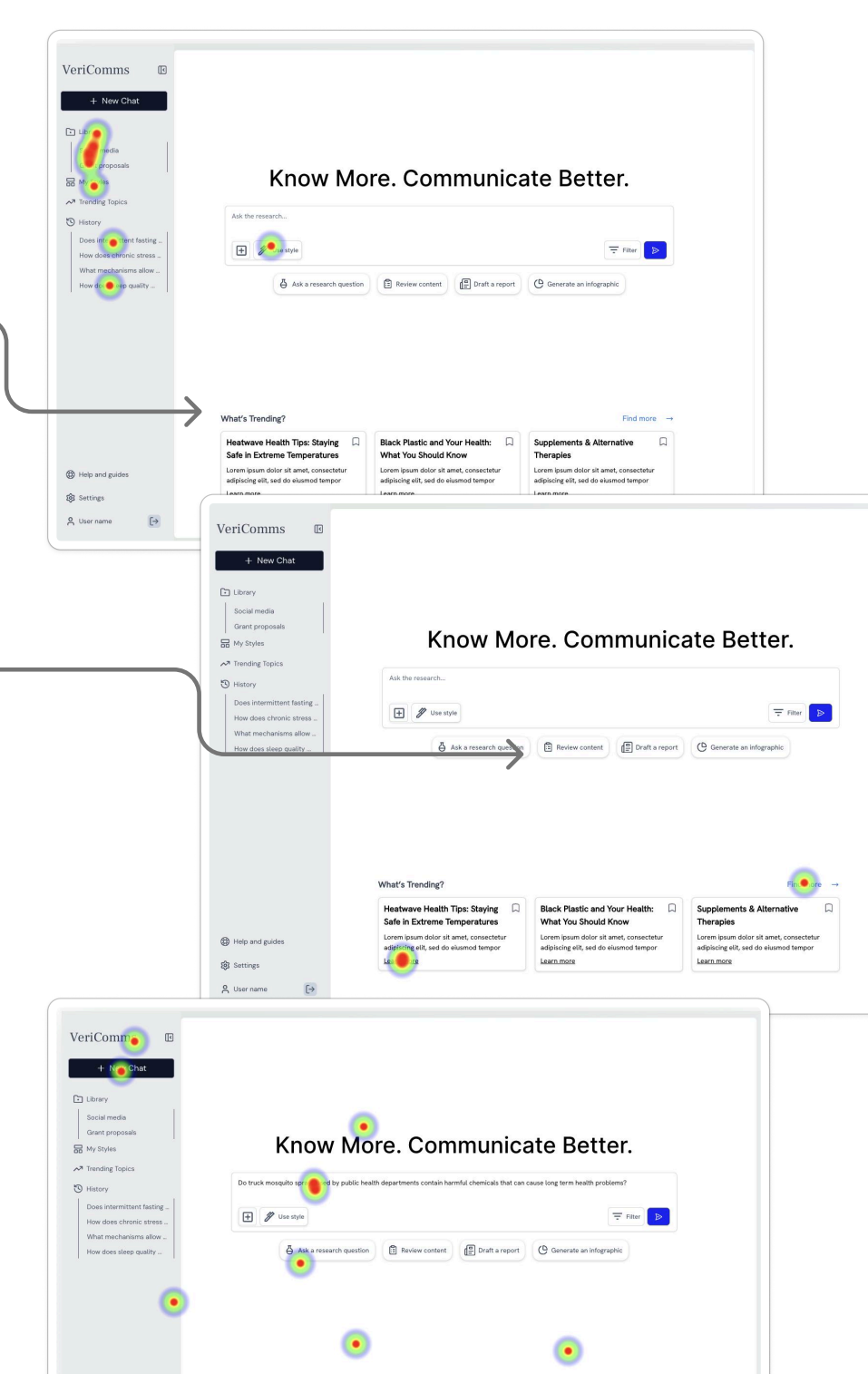
khadijat ibrahim

Enthusiastic participant assumed the way "review content" button worked would require him to first upload a doc or deliverable. #mentalmodel

khadijat ibrahim

Task 1. Dashboard

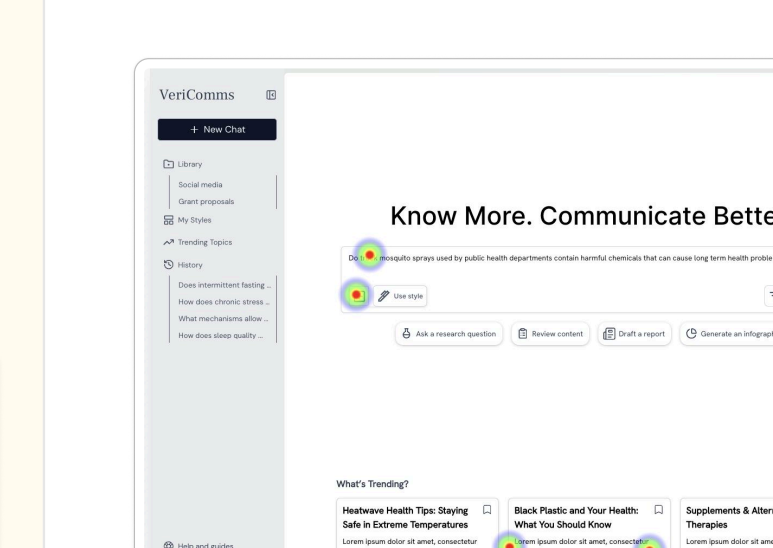
Click heat map



Easiness completing this task (1-5): 4.29±1.11

Task 2. AI output

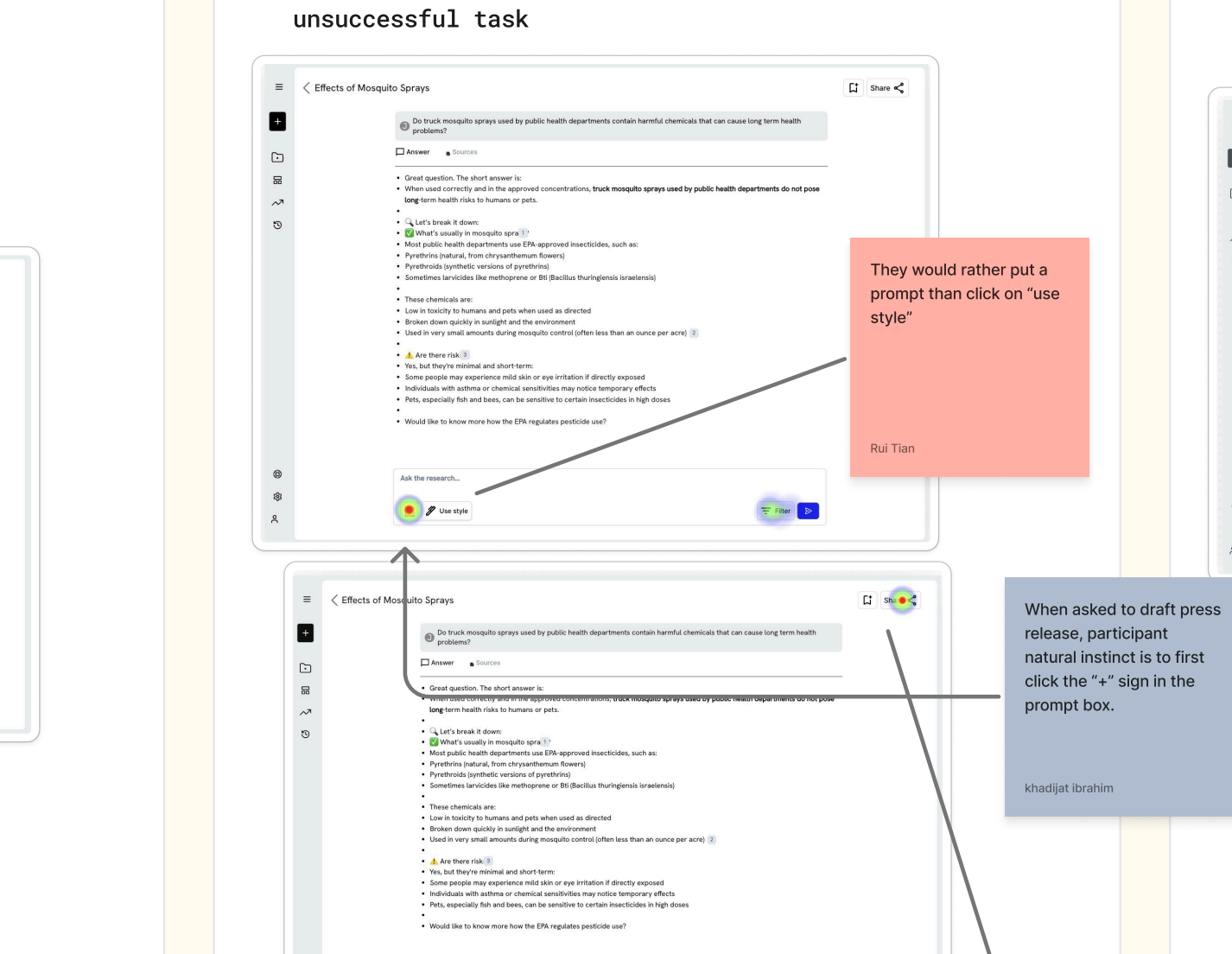
unsuccessful task



Easiness completing this task (1-5): 4.57±1.13

Task 3. Press release draft

unsuccessful task



They would rather put a prompt than click on "use style"

Rui Tian

When asked to draft press release, participant natural instinct is to first click the "+" sign in the prompt box.

khadijat ibrahim

Participant tried to click every possible button other than "use style".

Rui Tian

see the AI output→ change style

not natural to participant!!!

Rui Tian

Easiness completing this task (1-5): 3.43±1.27

One participant went back to the dashboard because they saw a "Draft a report" button there, but it was not available after the output was produced.

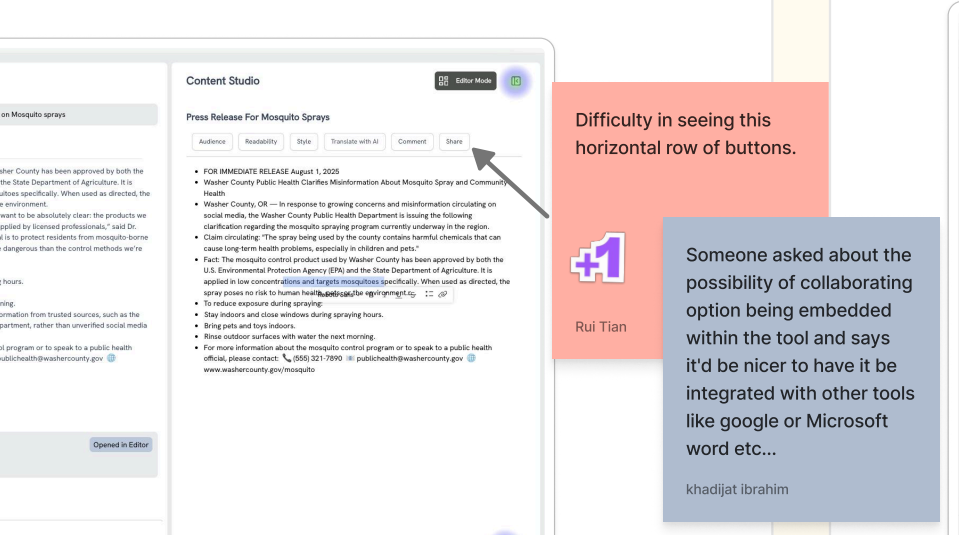
Katerina

Task 3 is the most difficult task, with spread out answers.

Rui Tian

Task 4.1. Sharing

unsuccessful task



Difficulty in seeing this horizontal row of buttons.

+1

Rui Tian

Someone asked about the possibility of collaborating option being embedded within the tool and says it'd be nicer to have it integrated with other tools like google or Microsoft word etc...

khadijat ibrahim

Participant are excited about this feature, they were thinking if they should copy+paste to share. this is their natural reaction to solve this problem (sharing). traditional,

Rui Tian

couldn't find button

Or, They did not see the "share" button.

Rui Tian

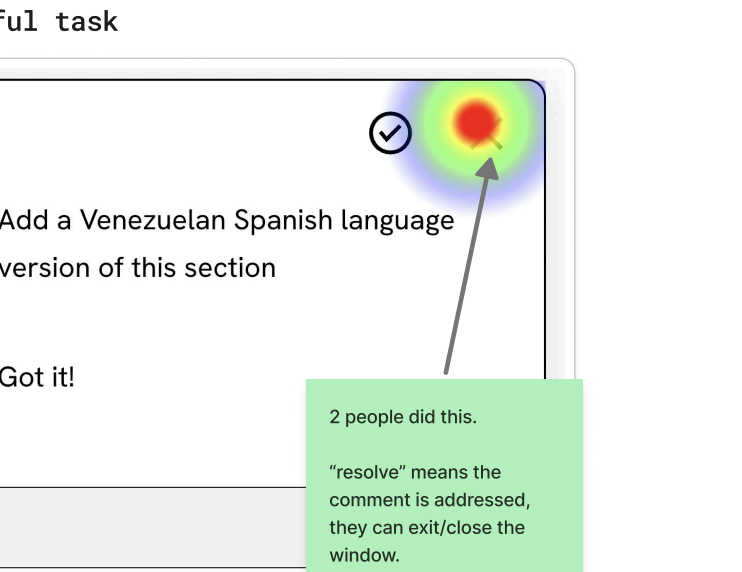
Another participant was thinking of downloading the output, e.g., as a word or PDF. This again highlights a more traditional approach to sharing documents.

Katerina

Easiness completing this task (1-5): 4 ±1.29

Task 4.2. Collaborating

unsuccessful task



K Add a Venezuelan Spanish language version of this section

J Got it!

2 people did this.

"resolve" means the comment is addressed, they can exit/close the window.

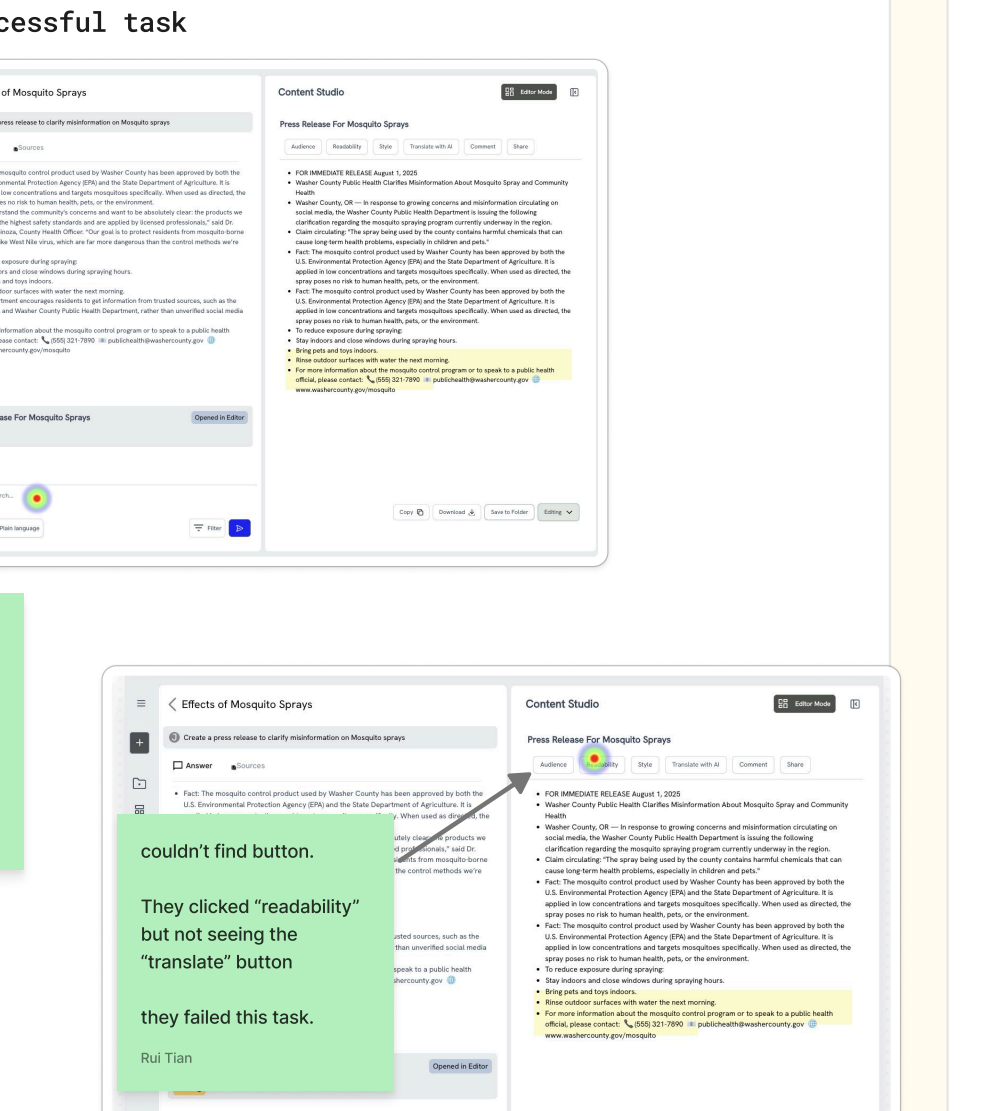
people did not click the "check" button.

Rui Tian

Easiness completing this task (1-5): 4.29 ±1.11

Task 5. Translation

unsuccessful task



writing a prompt seemed more intuitive to them.

Rui Tian

couldn't find button.

They clicked "readability" but not seeing the "translate" button

they failed this task.

Rui Tian

Easiness completing this task (1-5): 4.71 ± 0.78

Task 5 & 6 are the most easy task, closely clustered.

Rui Tian

Task 6. Review

All succeeded!!

Task 5 & 6 are the most easy task, closely clustered.

Rui Tian

Easiness completing this task (1-5): 4.86 ± 0.38

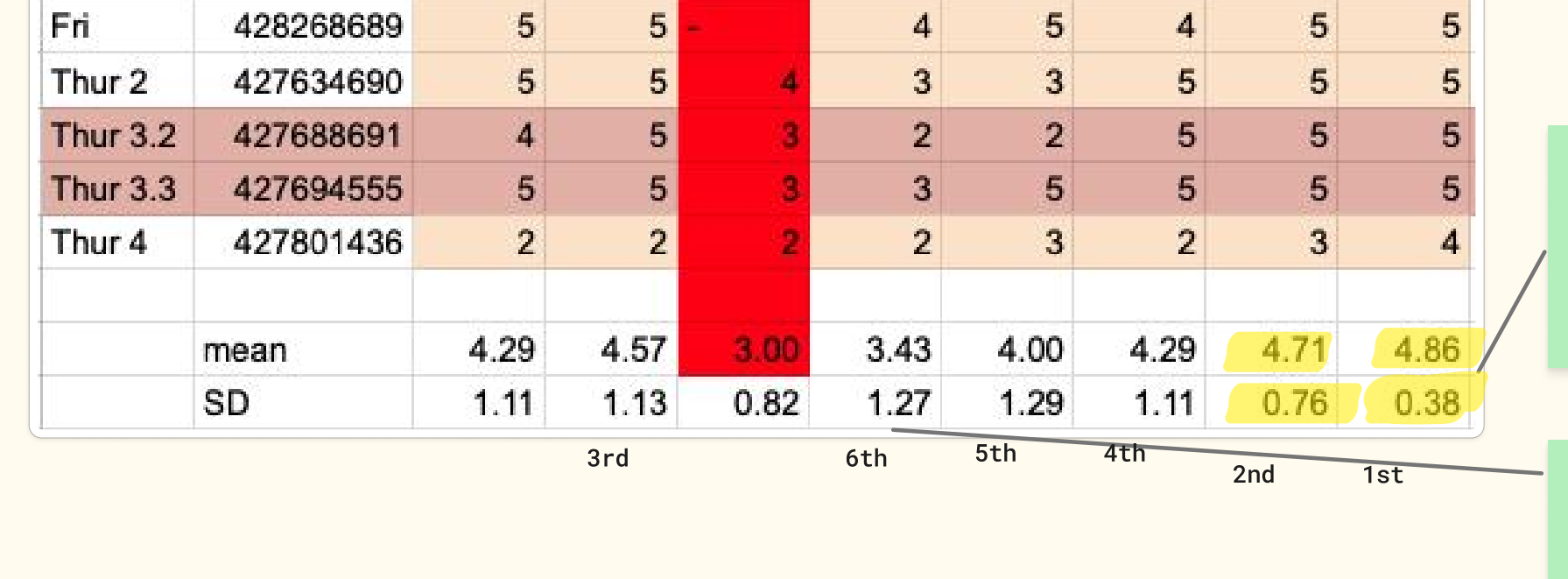
2. Surveys

1. Post-task Survey: "How would you rate your experience completing this task?"
Participants rate how difficult or easy they felt about completing the task on a scale from 1 (difficult) to 5 (easy). So higher score means a positive attitude. The two highlighted red rows are not ideal results, but we decided to keep them here.
Summary of finding:
- They feel task 3 is the most difficult one, but because the standard deviation is high, not all of them agreed.
- They feel task 5 and 6 are the easiest ones, and standard deviation is low, they tend to agree.

2. System Usability Scale (SUS)
This is a standardized post-usability test questionnaire that consists of 10 Likert-scale questions. Participants rate how strongly they agree or disagree with statements on 1-5 scale (1 - strongly disagree to 5 - strongly agree). SUS helps measure the perceived ease of use for any system.
Summary of finding:
- They do not all agree. Some of them think we had excellent usability, some of them think it's "meh".
- One participant gave some low scores. They struggle to read texts on screen and the moderator had to read it out loud to them. This one suggest that we should pay attention to accessibility of our prototype, and consider how our product would look like to a not so tech-savvy person.

3. Emotional experience
In addition to the SUS scale, we asked all participants to express their agreement with the following statement: I feel good when using this tool.
This question also used the scale from 1 to 5 (1 - frowning face to 5 - wide smiling face).
Overall, we see that users emotional experience (whether they feel good or not so good about using the product) aligns with their SUS answers:
2 on the feel good scale ⟶ 52.5 on SUS (1 person)
3 on the feel good scale ⟶ 82.5 on SUS (2 people)
4 on the feel good scale ⟶ 85.0 on SUS (1 person)
5 on the feel good scale ⟶ 92.5 on SUS (1 person)

Post Task: How easy



Scale		Task 1	Task 2	reliability	Task 3	Task 4.1	Task 4.2	Task 5	Task 6
Fri	428106261	5	5	-	5	5	4	5	5
Fri	428237553	4	5	-	5	5	5	5	5
Fri	428268689	5	5	-	4	5	4	5	5
Thur 2	427634690	5	5	4	3	3	5	5	5
Thur 3.2	427688691	4	5	3	2	2	5	5	5
Thur 3.3	427694555	5	5	3	3	5	5	5	5
Thur 4	427801436	2	2	2	2	3	2	3	4
	mean	4.29	4.57	3.00	3.43	4.00	4.29	4.71	4.86
	SD	1.11	1.13	0.82	1.27	1.29	1.11	0.76	0.38
			3rd		6th	5th	4th	2nd	1st

Task 5 & 6 are the most easy task, answers are closely clustered (consistent).

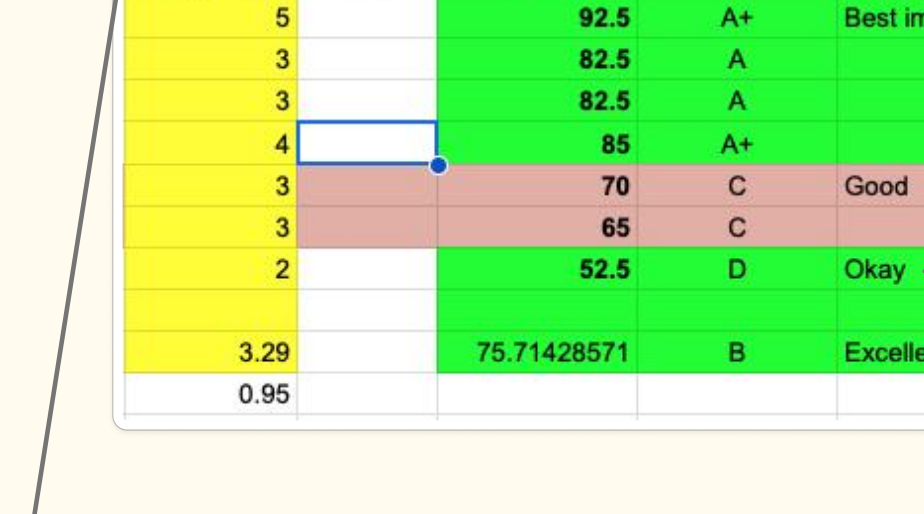
Rui Tian

Task 3 is the most difficult task, with spread out answers.

Rui Tian

SUS

emotional experience



U	V	W	X	Y
I feel good when I use this		SUS Sum		
5		92.5	A+	Best imaginable
3		82.5	A	
3		82.5	A	
4		85	A+	
3		70	C	Good
3		65	C	
2		52.5	D	Okay
3.29		75.71428571	B	Excellent
0.95				

Big variance in SUS

Rui Tian

This participant cannot see clearly, the moderator read the screen for them.

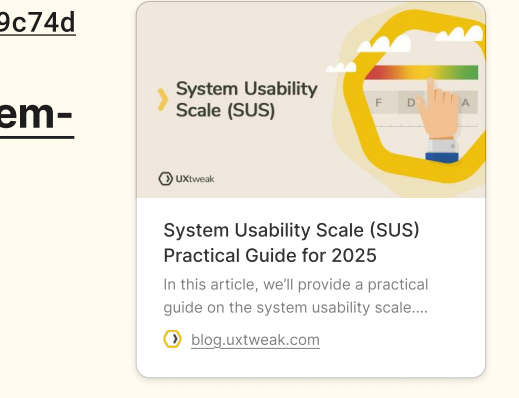
Accessibility issue
Tech literacy

Rui Tian

https://uxdesign.cc/measuring-user-trust-a30b56a9c74d

https://uiuxtrend.com/measuring-system-usability-scale-sus/

https://blog.uxtweak.com/system-usability-scale/
Interpreting the SUS scores



System Usability Scale (SUS) Practical Guide for 2025

In this article, we'll provide a practical guide on the system usability scale...

blog.uxtweak.com

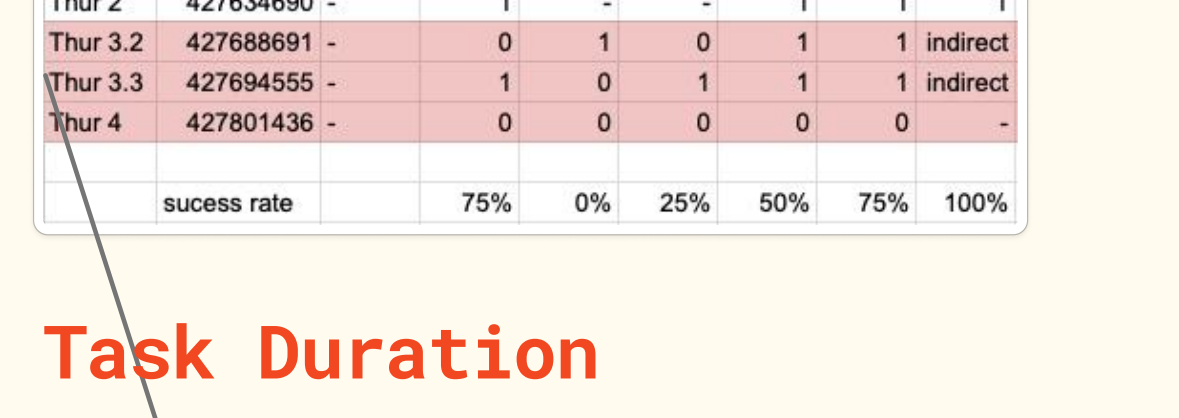
3. Effectiveness & Efficiency

Notes:
- We did not count the highlighted red rows in calculation.
- 427634690 had issue with the Maze system which was an error on our side. They technically succeed in the two missing data point. But for accuracy, we considered them as missing data.

1. Effectiveness: Success Rate (4 participants)
In general, 78% is an average success rate. We have 5 out of 6 tasks' success rate is lower than this. This suggest that there's still room for improvement and a lot to work on.
- Task 3 has a 0% success rate. Task 4.1 has a 25% success rate. Most likely there are some major usability issues.

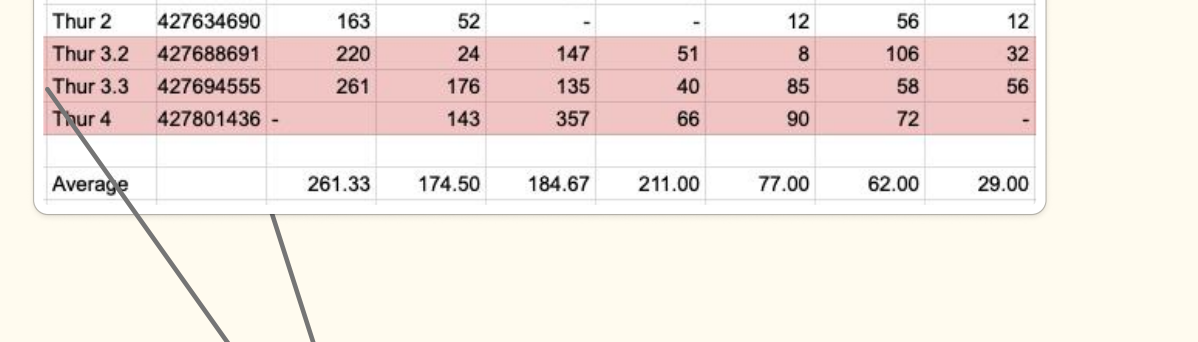
2. Efficiency: time (4 participants)
- **Task Duration (in seconds):** Ideally, we want them to do it as quickly and simply as possible. When you see that a certain task takes longer to complete (task 4.1), this might be a sign of them experiencing confusion, probably due to the usability issues.
- **Efficiency**: This is how quickly and easily users can complete tasks. A slightly complicated math formula was calculated. Check out the highlighted yellow row: the higher the value, the more efficient participants were in a task.
- We found a discrepancy in efficiency ranking vs. post-task survey (section 2).
  - Efficiency (from high to low): task 6 → 4.2 → 5 → 2 → 4.1 → 3
  - Post- task (from high to low): 6 → 5 → 2 → 4.2 → 4.1 → 3
  - Efficiency is "objective time", and post task survey is "subjective ease". This may suggest that cognitive load and confidence influence perception. here's our suggestions:
- Task 4.2: make instructions to make users feel as confident as their speed suggests.
- Task 2: explore why people perceive it as harder despite a decent speed. maybe it's more mentally loaded?

Success Rate



Success =1, Fail =0		Task 1	Task 2	Task 3	Task 4.1	Task 4.2	Task 5	Task 6
Fri	428106261	-	1	0	0	0	0	1
Fri	428237553	-	0	0	0	0	1	1
Fri	428268689	-	1	0	1	1	1	1
Thur 2	427634690	-	1	-	-	1	1	1
Thur 3.2	427688691	-	0	1	0	1	1	indirect
Thur 3.3	427694555	-	1	0	1	1	1	indirect
Thur 4	427801436	-	0	0	0	0	0	-
	sucess rate		75%	0%	25%	50%	75%	100%

Task Duration



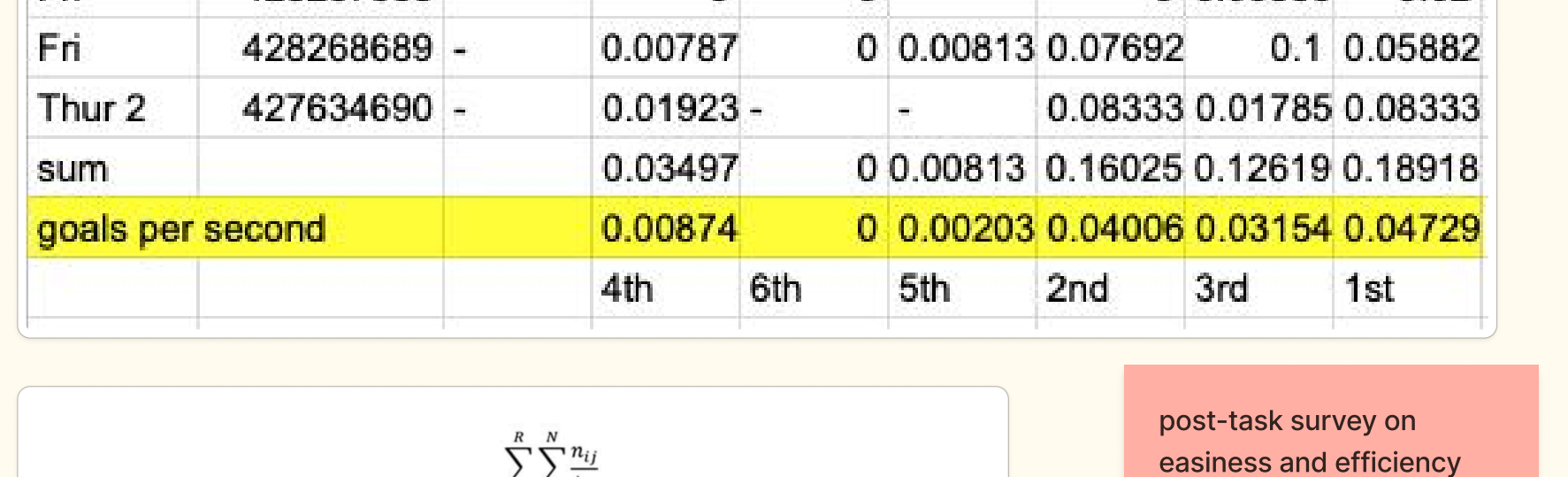
duration		Task 1	Task 2	Task 3	Task 4.1	Task 4.2	Task 5	Task 6
Fri	428106261	557	127	118	299	214	62	37
Fri	428237553	-	392	6	-	69	120	50
Fri	428268689	64	127	430	123	13	10	17
Thur 2	427634690	163	52	-	-	12	56	12
Thur 3.2	427688691	220	24	147	51	8	106	32
Thur 3.3	427694555	261	176	135	40	85	58	56
Thur 4	427801436	-	143	357	66	90	72	-
Average		261.33	174.50	184.67	211.00	77.00	62.00	29.00

Discard the red highlighted rows

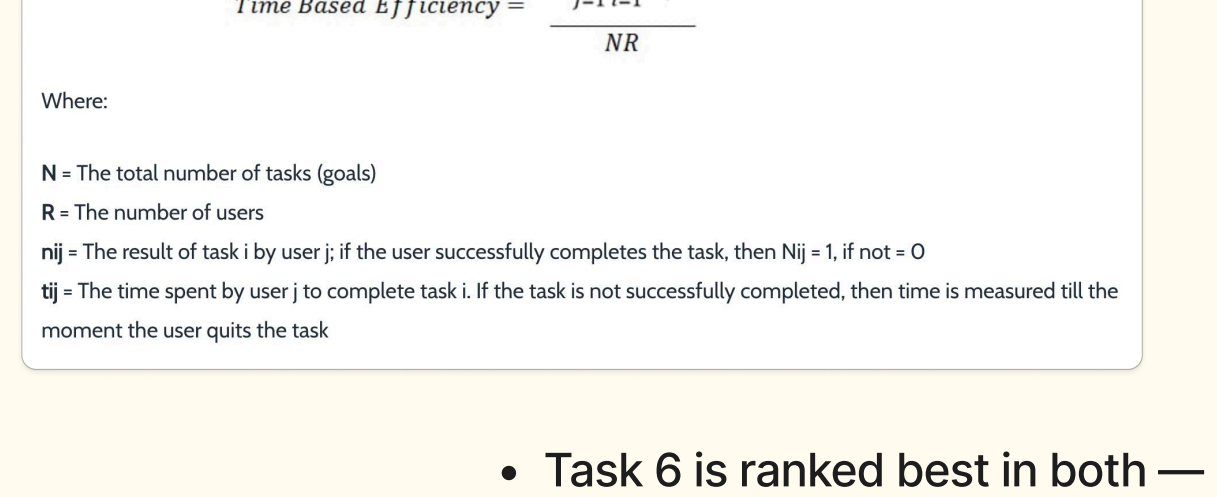
Rui Tian

This is the average time a user takes to successfully complete each task.

Time-based Efficiency (goals per second)



		Task 1	Task 2	Task 3	Task 4.1	Task 4.2	Task 5	Task 6
Fri	428106261	-	0.00787	0	0	0	0	0.02702
Fri	428237553	-	0	0	-	0	0.00833	0.02
Fri	428268689	-	0.00787	0	0.00813	0.07692	0.1	0.05882
Thur 2	427634690	-	0.01923	-	-	0.08333	0.01785	0.08333
sum			0.03497	0	0.00813	0.16025	0.12619	0.18918
goals per second			0.00874	0	0.00203	0.04006	0.03154	0.04729
			4th	6th	5th	2nd	3rd	1st



$$\text{Time Based Efficiency} = \frac{\sum_{j=1}^{R}\sum_{i=1}^{N}\frac{n_{ij}}{t_{ij}}}{NR}$$

Where:

N = The total number of tasks (goals)
R = The number of users
nij = The result of task i by user j; if the user successfully completes the task, then Nij = 1, if not = 0
tij = The time spent by user j to complete task i. If the task is not successfully completed, then time is measured till the moment the user quits the task

post-task survey on easiness and efficiency data doesn't match perfectly:

people didn't know what is expected of them of our system.

Rui Tian

- Task 6 is ranked best in both — users did it fastest and also felt it was easiest.
- Task 4.2 was objectively the second fastest but was perceived as harder than tasks 5 and 2.
- Task 5 was objectively slower than 4.2 but felt easier.
- Task 2 was faster than Task 5 objectively, but users felt it was harder.
- Task 4.1's relative position is consistently near the lower middle.
- Task 3 is ranked lowest in both — slowest and felt hardest.