Carnegie Mellon University

QA: User Study

17-313 Fall 2022

Rohan Padhye, Michael Hilton, Chris Timperley, and Daye Nam

Software and Societal Systems Department

https://ninawei.medium.com/reflections-on-bad-and-good-design-in-my-daily-life-d4091067a374

https://www.ramotion.com/blog/bad-ux-design/

Why do we care about users?

Human Centered Methods

- Contextual Inquiry
- Paper prototypes
- Think-aloud protocols
- Heuristic Evaluation
- Affinity diagrams
- Personas
- …

- A/B testing
- Body storming
- Questionnaires
- Surveys
- Log analysis
- Card sorting
- …

S3D

How can we test usability

Activity

How can you test usefulness / usability of CoPilot?

GitHub Copilot

S3D

Usability Evaluation Case Study

Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models

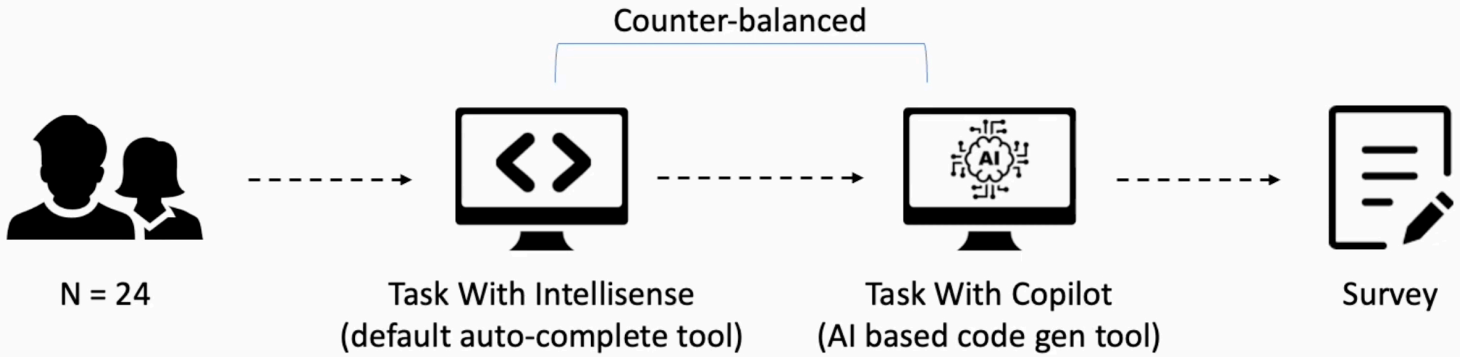
Priyan Vaithilingam
pvaithilingam@g.harvard.edu
Harvard University
USA

Tianyi Zhang
tianyi@purdue.edu
Purdue University
USA

Elena L. Glassman
glassman@seas.harvard.edu
Harvard University
USA

S3D

User Study Procedure



Participants

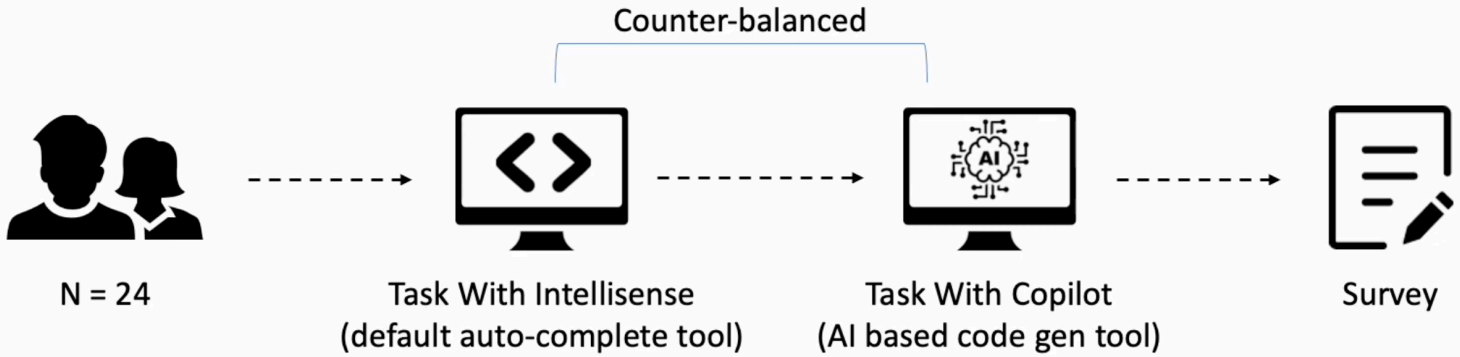
- 24 participants
  - 4 Female, 19 Male, 1 Non-binary
  - 5 master's students, 8 Ph.D. students, 1 software engineer
  - 1: <= 2 years, 14: 2-5 years, 9: > 5 years
- $20 Amazon giftcard as compensation

Task Design

- Real-world Python programming tasks with different levels of difficulty
  - T1: Edit CSV (Easy)
  - T2: Web Scrapping (Medium)
  - T3: Graph Plotting (Hard)

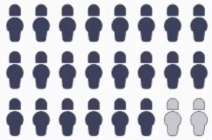
S3D

User Study Procedure



Quantitative Results

Using Intellisense



22 of 24

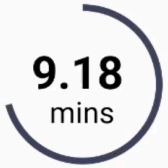
participants finished the task successfully.

Average Time

Using Copilot

19 of 24

participants finished the task successfully.



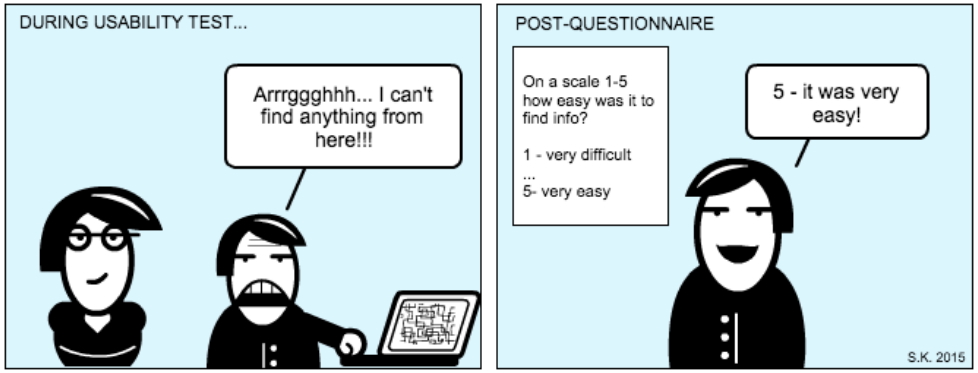
Average Time

19 of 24

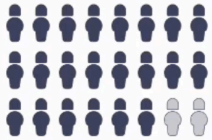
participants **prefer using Copilot**

Not statistically significant

Quantitative Results



Using Intellisense



22 of 24

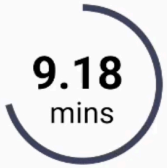
participants finished the task successfully.

Average Time

Using Copilot

19 of 24

participants finished the task successfully.



Average Time

19 of 24

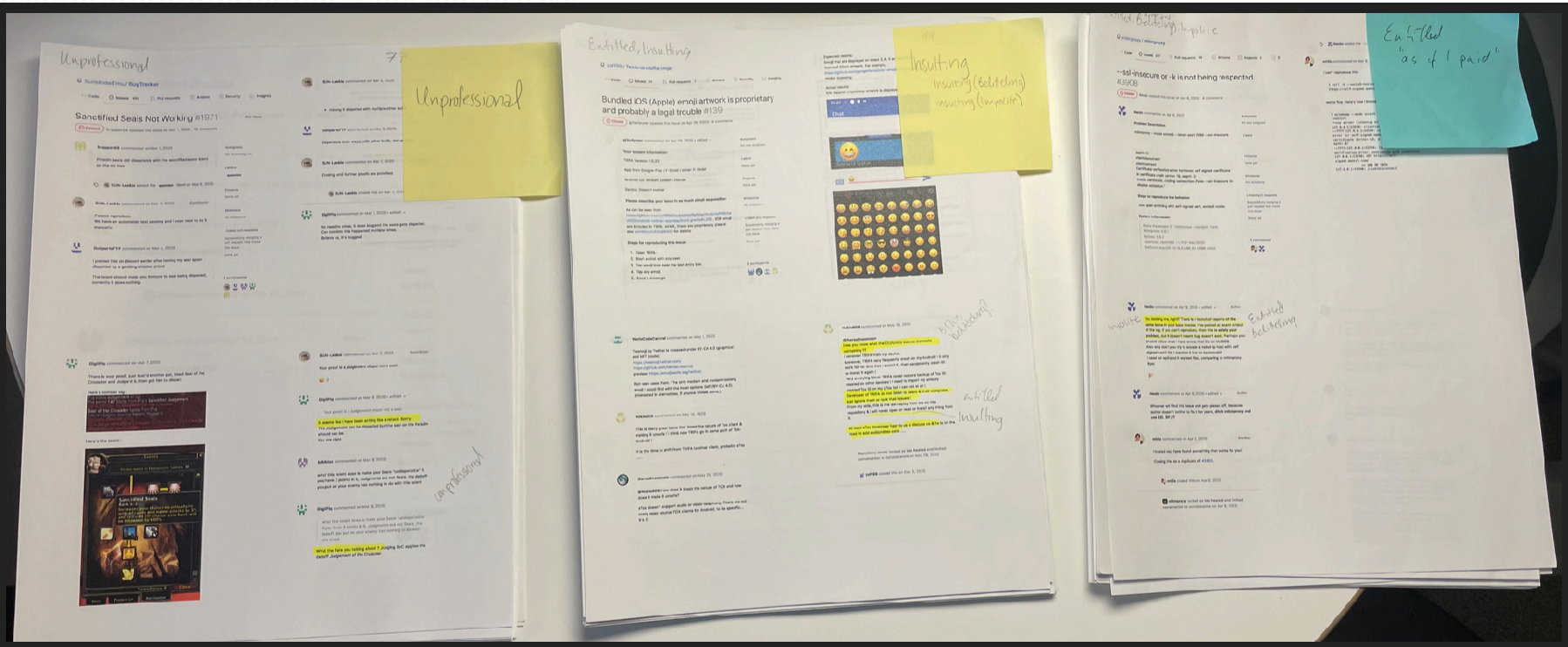
participants **prefer using Copilot**

Not statistically significant

+ Qualitative analysis

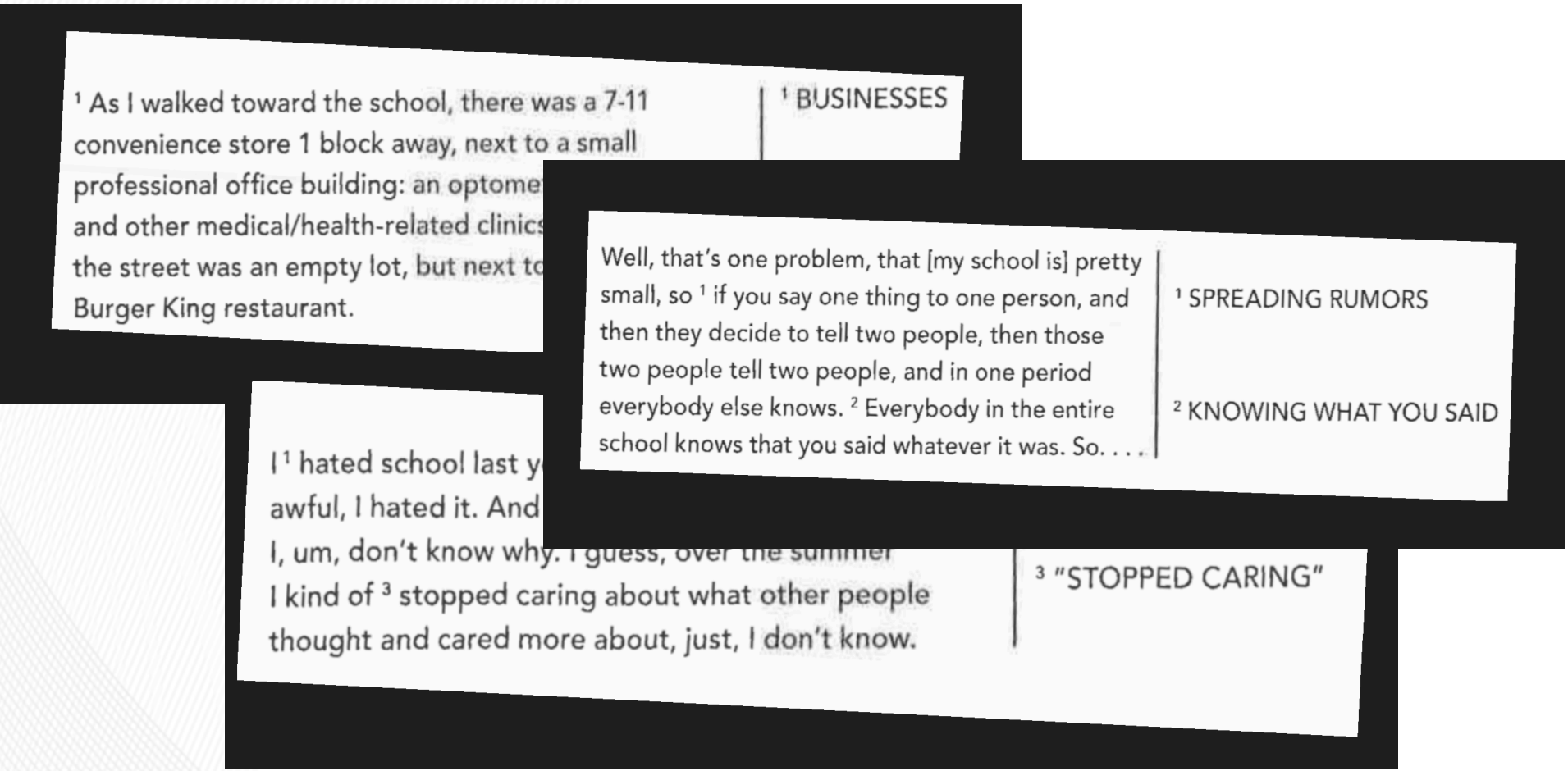
- Piles of qualitative data, mostly text
  - What to do with it?
  - From journalism to science – how?
- Step 1: Abstraction
  - Attach "codes" (labels) to chunks of data
  - Characterize / summarize the data
- Step 2: Finding patterns
  - Use these abstractions to find meta-patterns, craft a theory ("grounded theory"), …
  - Interpret the data
- This is difficult, but very doable with practice

+ Qualitative analysis



Miller, C., Cohen, S., Klug, D., Vasilescu, B., & Kästner, C. (2022). "Did You Miss My Comment or What?" Understanding Toxicity in Open Source Discussions. In In 44th International Conference on Software Engineering (ICSE'22).

+ Qualitative analysis



Qualitative Insights

Copilot provides a useful starting point for users unlike Intellisense

12 users found code generated by copilot **hard to understand and change**

5 users found it **hard to trust code** generated by copilot

Users of Copilot **constantly switch modes** between reading and writing code

Many users used copilot as a **subsititution for internet search**

8 users **over-relied on copilot**, i.e., **ignored validating** the code generated by copilot

Design Implications

Several participants requested to see **multiple code suggestions** for a given prompt

Explanations with inline comments for code generated by copilot can help in understanding

Users relying on internet search compare multiple sources unlike users using copilot. It is worthwile **integrating online search with code generation**

We can also help users debug the code by **automatically generating test cases** and test data for users to validate

We observed copilot is more accurate for simpler prompts. It's worth exploring **task decomposition** for better accuracy

Activity

What are the other aspects to consider?

GitHub Copilot

S3D