Human-AI Interaction, Lecture 7

Evaluating Human-AI Systems

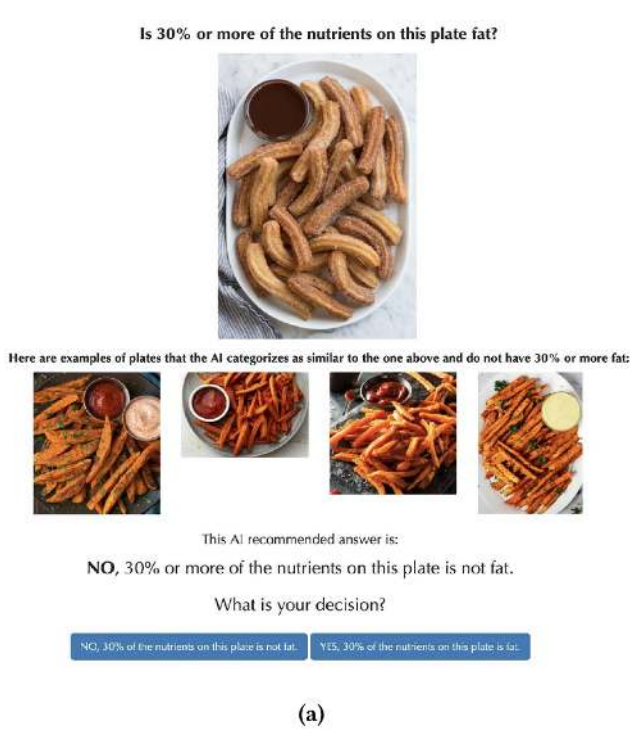
A simulated AI says whether this plate is high in fat and explains its answer

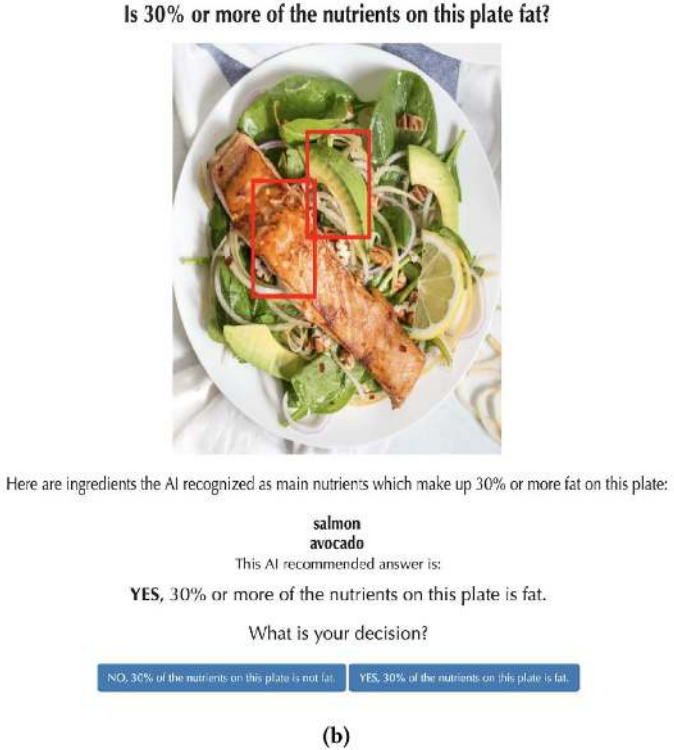
Is 30% or more of the nutrients on this plate fat?



Buçinca, Lin, Gajos & Glassman, IUI 2020, Fig. 2a (top)

Two explanation styles





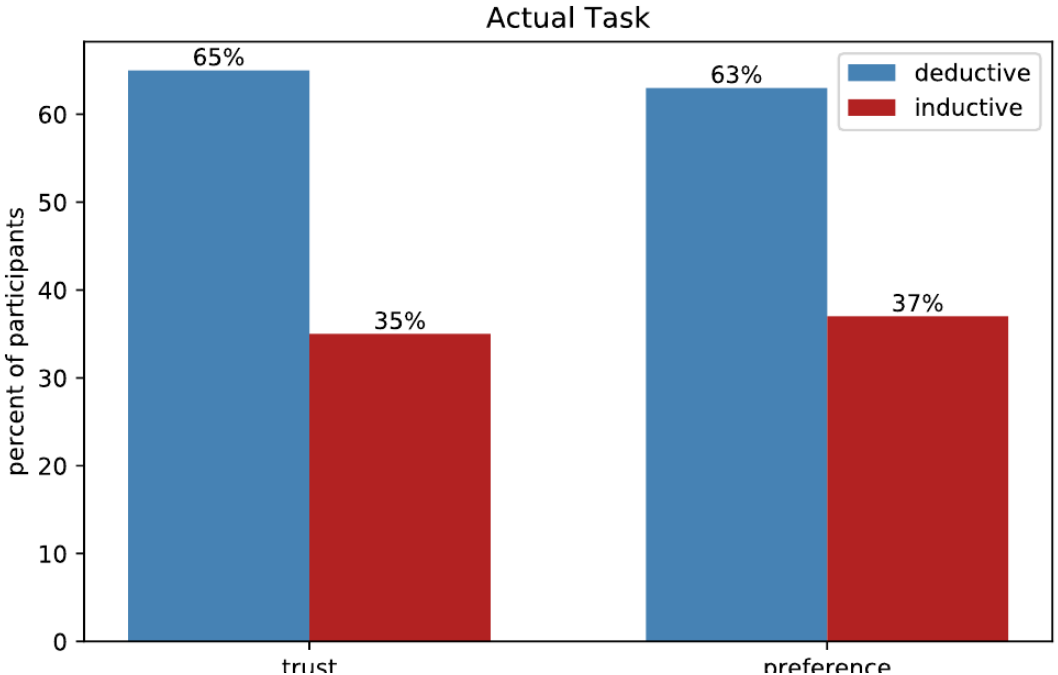
(a) plates the AI sees as similar

(b) ingredients the AI recognized

Which would you trust more?

Buçinca et al., IUI 2020, Fig. 2 (the actual task)

Trust and preference



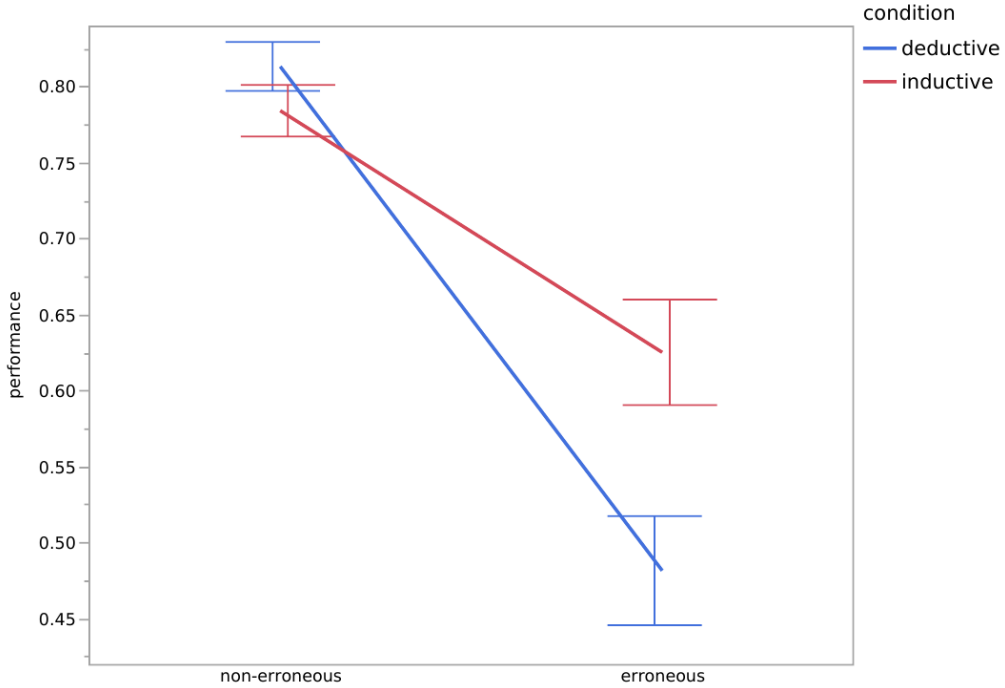
Blue: the ingredient list (“deductive”)

Red: the similar plates (“inductive”)

Most trusted and preferred the list.

Buçinca et al., IUI 2020, Fig. 5b

Accuracy when the AI was wrong



People answered correctly

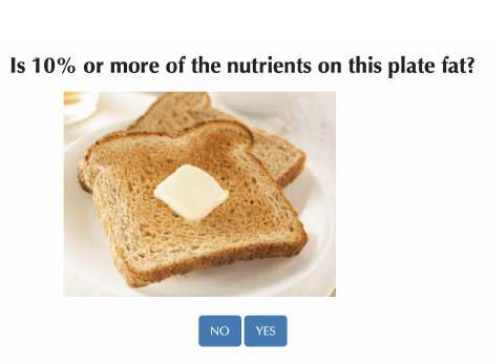
63% of the time with the plates

48% with the ingredient list

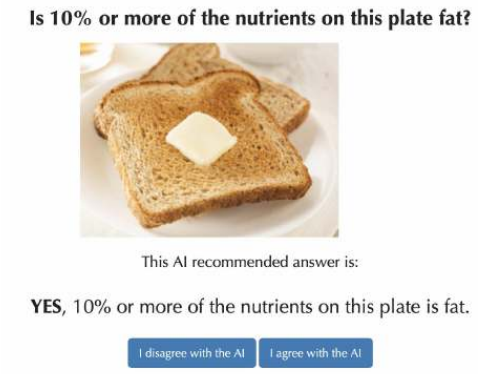
Buçinca et al., IUI 2020, Fig. 4b and Section 4.2

How can we tell whether a human-AI design works?

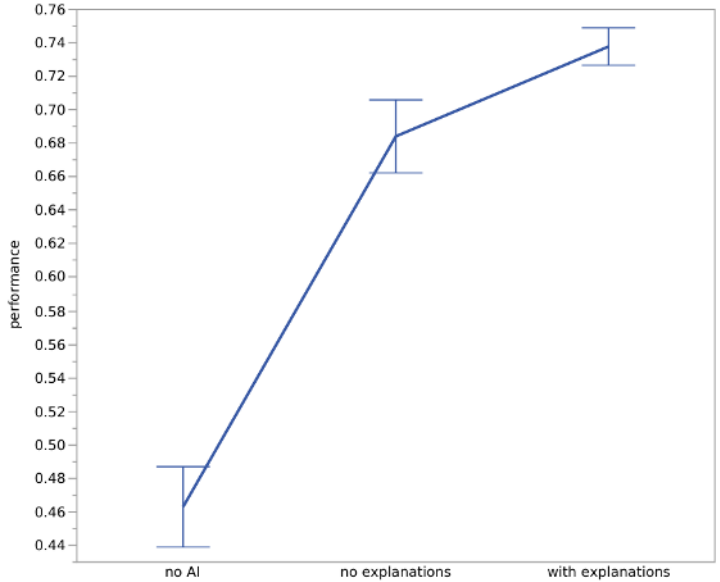
Comparison groups



(a)



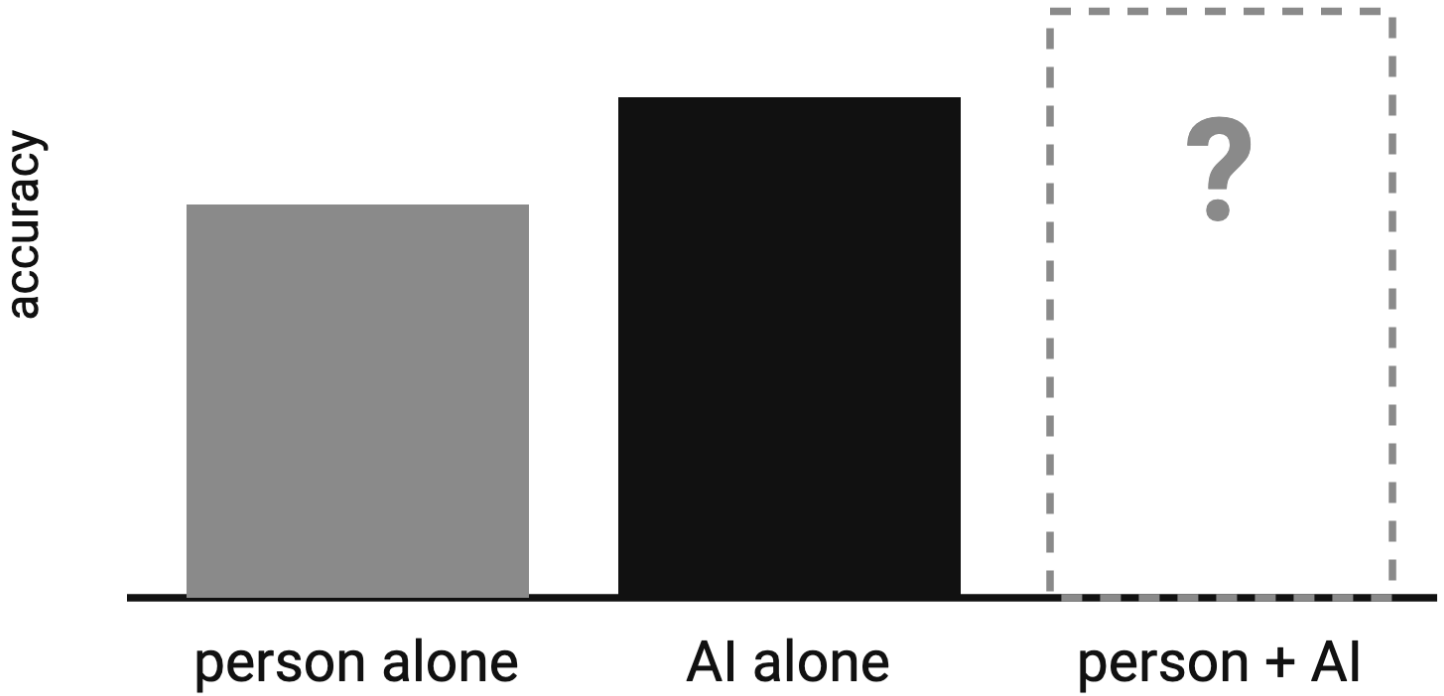
(b)



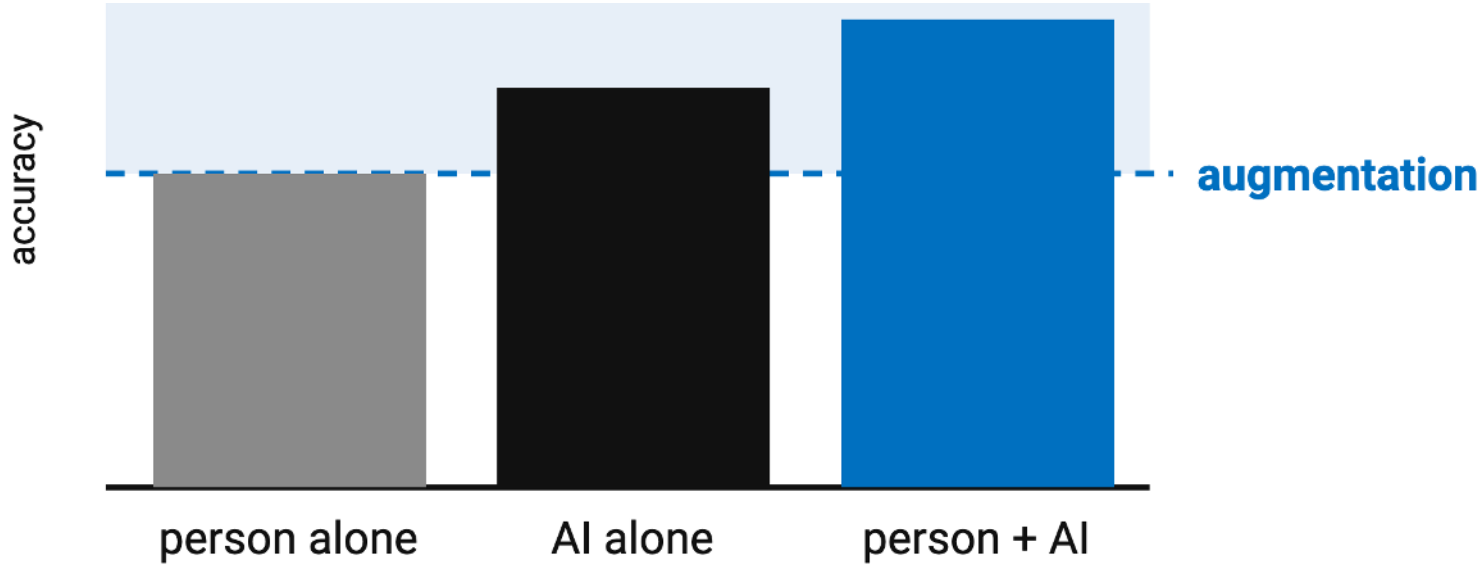
Correct answers:
no AI 46%
answer only 68%
with explanation 74%

Buçinca et al., IUI 2020, Fig. 3 (left) and Fig. 4a (right)

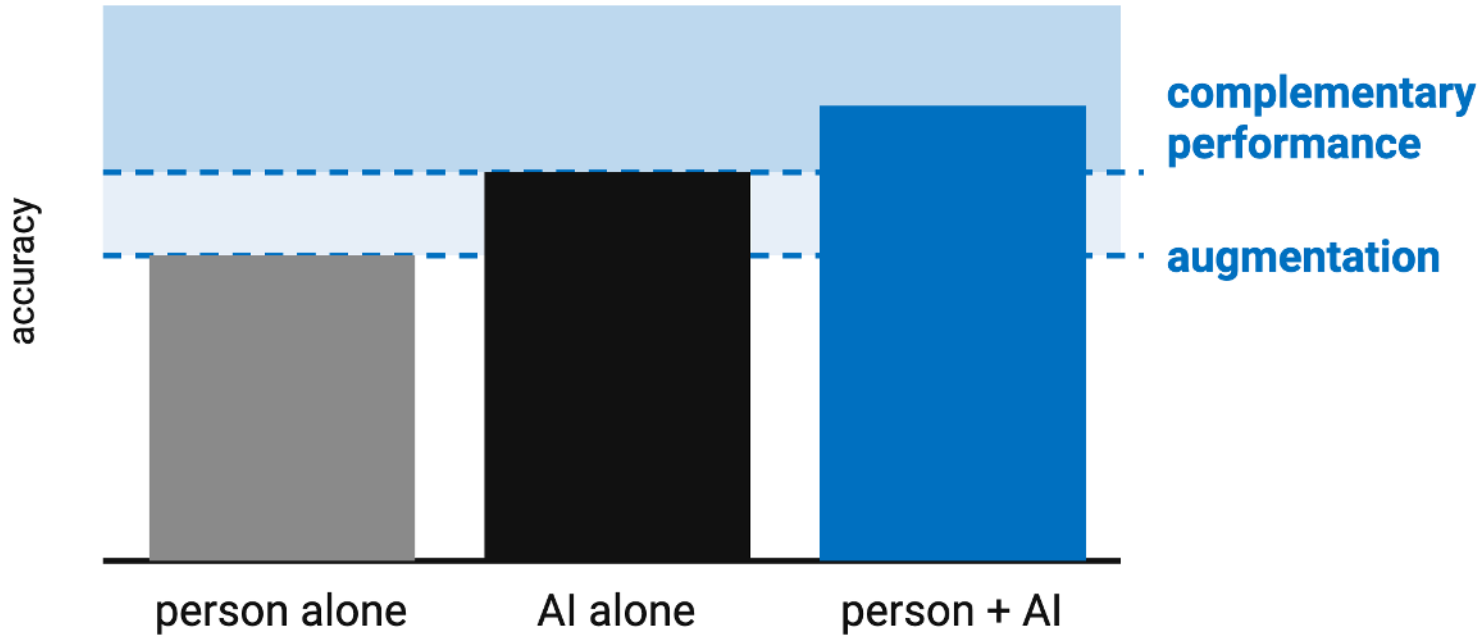
Two baselines



Augmentation



Complementary performance



Term from Bansal et al., CHI 2021

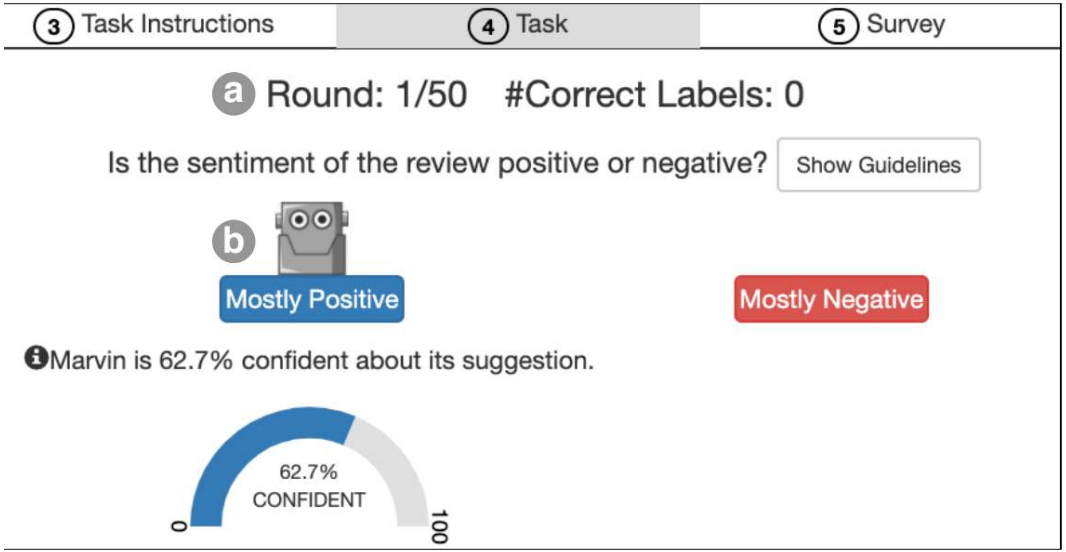
Earlier studies (Bansal et al., Table 1)

Domain	Task	Performance				
		Metric	Human alone	AI alone	Team	Complementary?
Classification	Deceptive review [43]	Accuracy ↑	51.1%	**87.0%**	74.6%	✗
	Deceptive review [42]	Accuracy ↑	54.6%	**86.3%**	74.0%	✗
	Income category [83]	Accuracy ↑	65%	**75%**	73%	✗
	Loan defaults [27]	Norm. Brier ↑	0	**1**	0.682	✗
	Hypoxemia risk [53]	AUC ↑	0.66	**0.81**	0.78	✗
	Nutrition prediction [12]	Accuracy ↑	0.46	**0.75**	0.74	✗
QA	Quiz bowl [21]	*"AI outperforms top trivia players."*				✗

The plate study:
people alone 0.46
the pair 0.74
the AI alone 0.75

Bansal et al., CHI 2021, Table 1

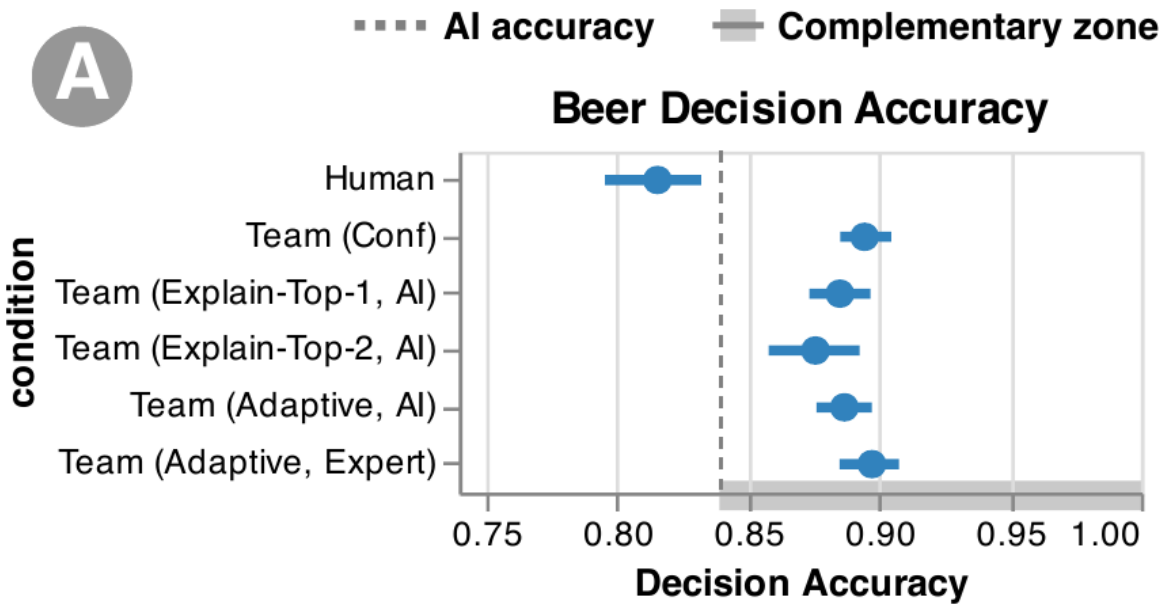
Bansal et al.: the review task



The AI was right 84% of the time, close to people on their own.

Bansal et al., CHI 2021, Fig. 2, right pane (book reviews) and Section 4.2.1

Results: beer reviews



Human: 82%

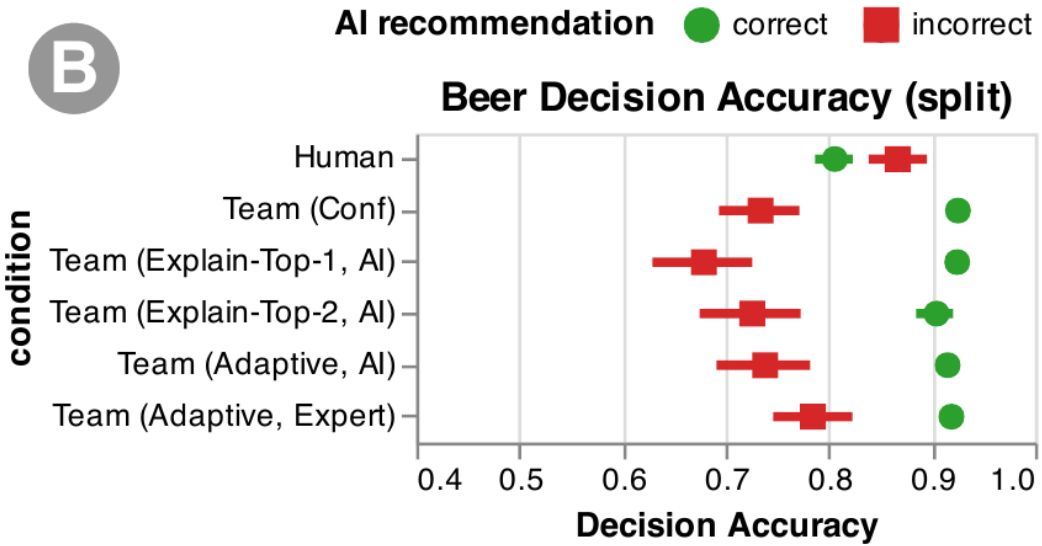
AI (dashed line): 84%

Team (Conf): 89%

The pair beat both on all three tasks.

Bansal et al., CHI 2021, Fig. 4A (beer reviews) and Section 5.1; 1,626 participants in all

Results: explanations



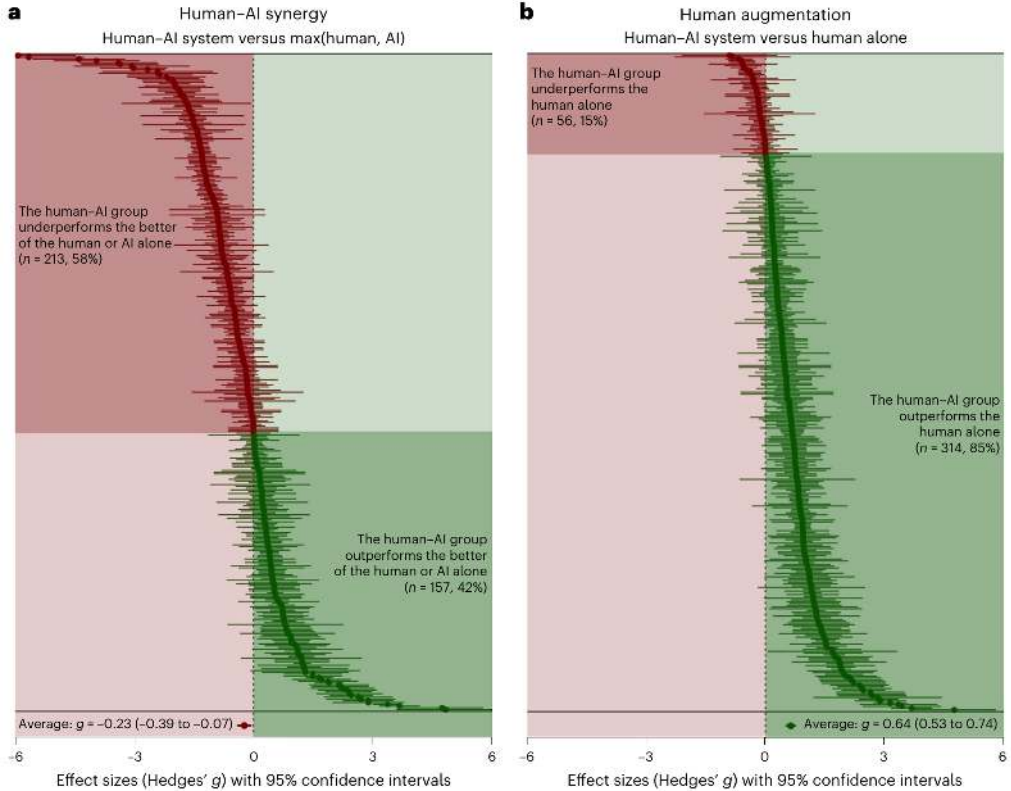
Green: the AI was right

Red: the AI was wrong

“explanations increased the chance that humans will accept the AI's recommendation, **regardless of its correctness**”

Bansal et al., CHI 2021, Fig. 4B (beer reviews) and abstract

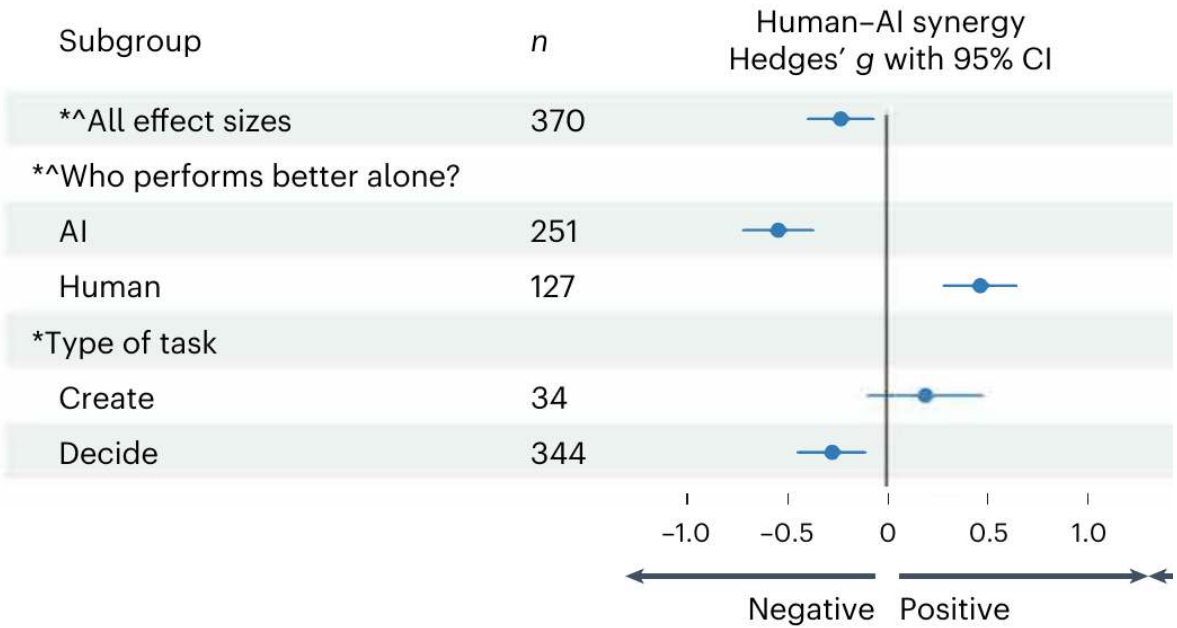
Vaccaro et al.: 106 experiments



A test should report the person alone, the AI alone, and the two together.

Vaccaro, Almaatouq & Malone, Nature Human Behaviour 2024, Fig. 1 (CC BY 4.0)

Vaccaro et al.: subgroups



Right of 0: the pair did better than the better of the two.

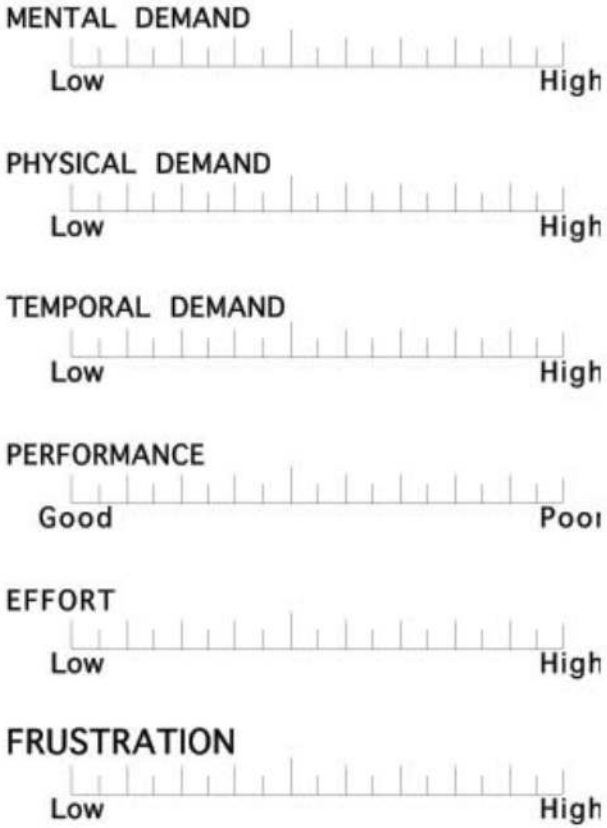
Vaccaro, Almaatouq & Malone, Nature Human Behaviour 2024, Fig. 2 (CC BY 4.0)

Measuring reliance

	kept the AI's answer	changed it
AI was right	relied on it, rightly	**ignored a right answer**
AI was wrong	**accepted a wrong answer**	**caught the error**

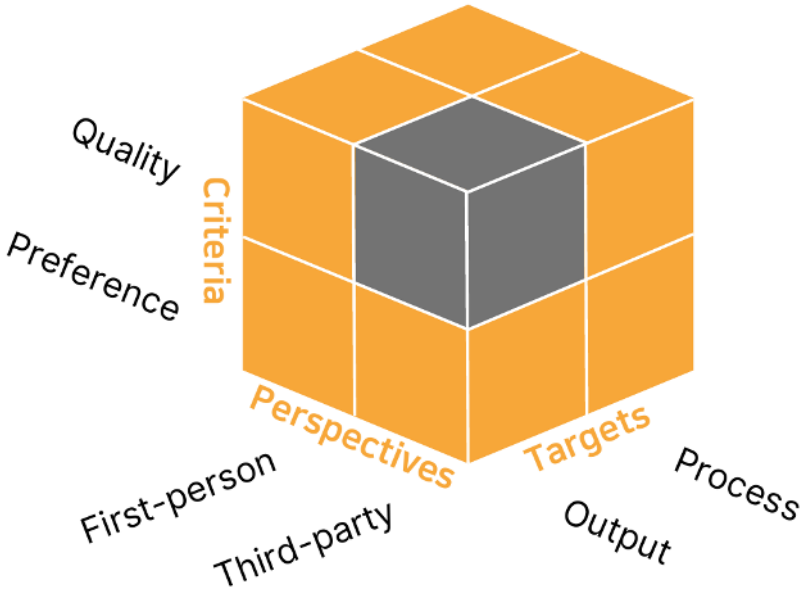
Measuring effort: NASA-TLX

People rate six scales right after the task.



NASA-TLX rating scales, from Hart, HFES 2006, Appendix

HALIE: three dimensions



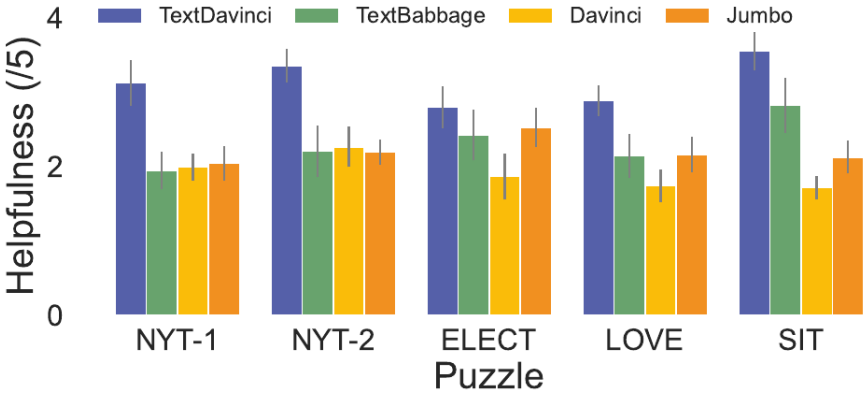
Targets: the output, or the process

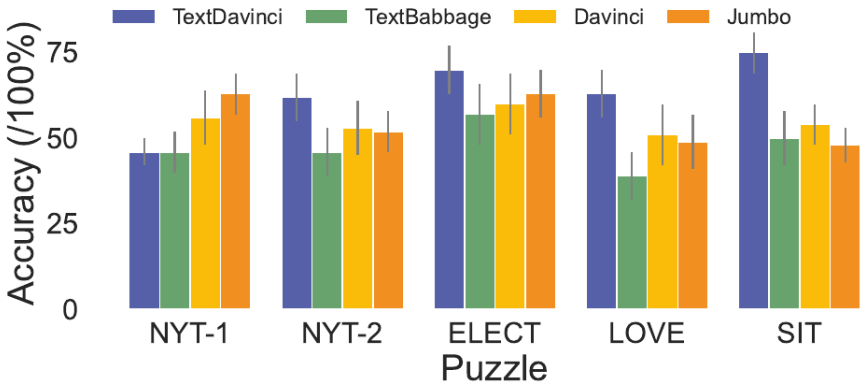
Perspectives: an outside judge, or the user

Criteria: quality, or preference

Lee et al., TMLR 2023, Fig. 1

Results: crossword puzzles

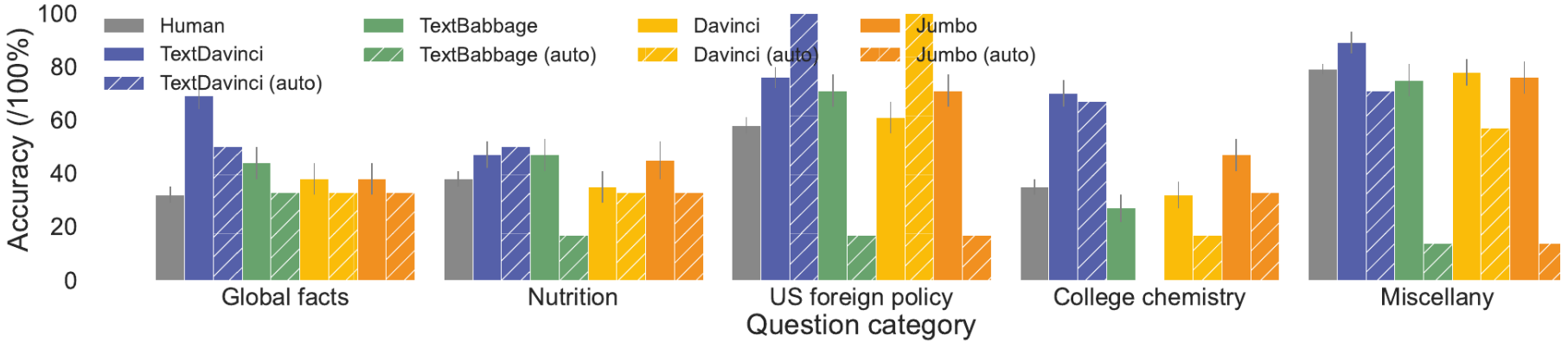




TextDavinci (blue): rated most helpful,
but **fewer correct letters on NYT-1**

Lee et al., TMLR 2023, Fig. 8 and Section 3.4 (crossword puzzles)

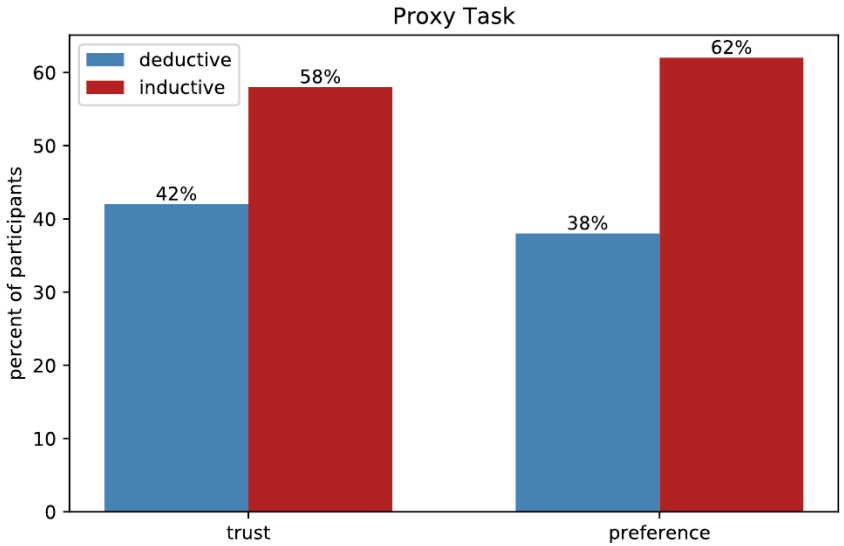
Results: nutrition questions

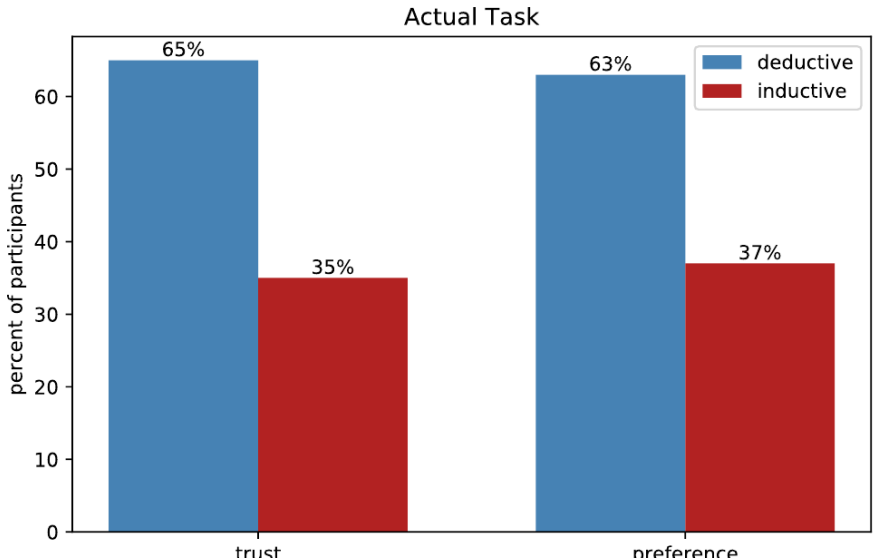


TextBabbage (green): least accurate on its own, tied for best with people

Lee et al., TMLR 2023, Fig. 5 and Section 3.3

Proxy task and actual task





Predicting the AI: most preferred the plates

Deciding themselves: most preferred the list

Buçinca et al., IUI 2020, Fig. 5 (proxy task left, actual task right)

Planted errors

We fixed the errors at questions 4, 7, 11, 16, 22 and 23, though which food the error was associated to was randomized.

You can only measure reliance if you know where the AI is wrong.

Buçinca et al., IUI 2020, Section 3.1.1

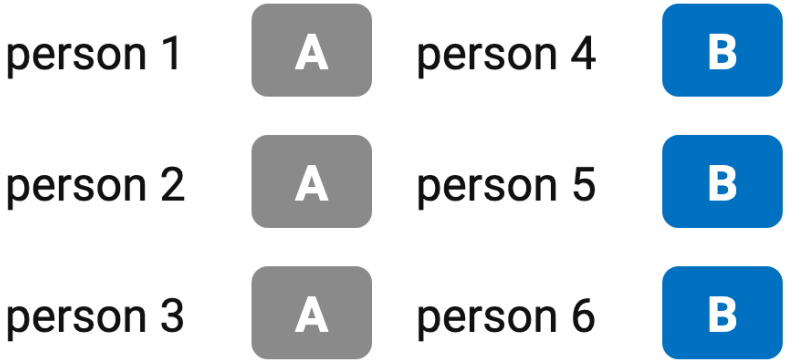
Within- and between-subjects designs

within-subjects



everyone tries both,
in a swapped order

between-subjects



each person tries one,
so you need more people

Think-aloud

Thinking aloud can change what people do.

Thinking aloud: 8 of 11 preferred the plates

Online: most preferred the list

Buçinca et al., IUI 2020, Section 5.5

Testing with two people

What it shows

where each person gets stuck

which planted errors they miss

What it cannot show

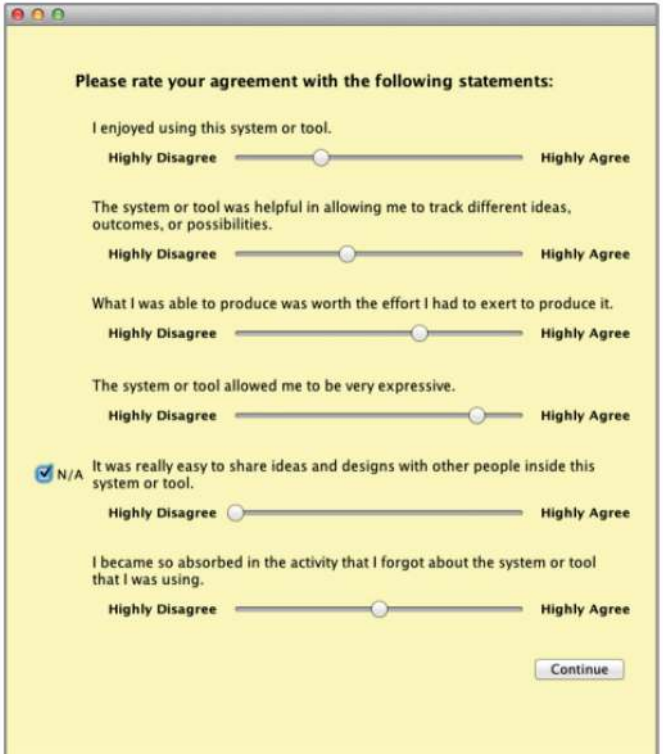
how common a problem is

that your design is better

Creativity Support Index

The CSI asks about six factors:

Exploration
Expressiveness
Immersion
Enjoyment
Results Worth Effort
Collaboration



Cherry & Latulipe, TOCHI 2014, Fig. 1

CSI part 1: agreement statements

Exploration 1. It was easy for me to explore many different ideas, options, designs, or outcomes, using this system or tool. 2. The system or tool was helpful in allowing me to track different ideas, outcomes, or possibilities.

The two Exploration statements, each rated 0 to 10

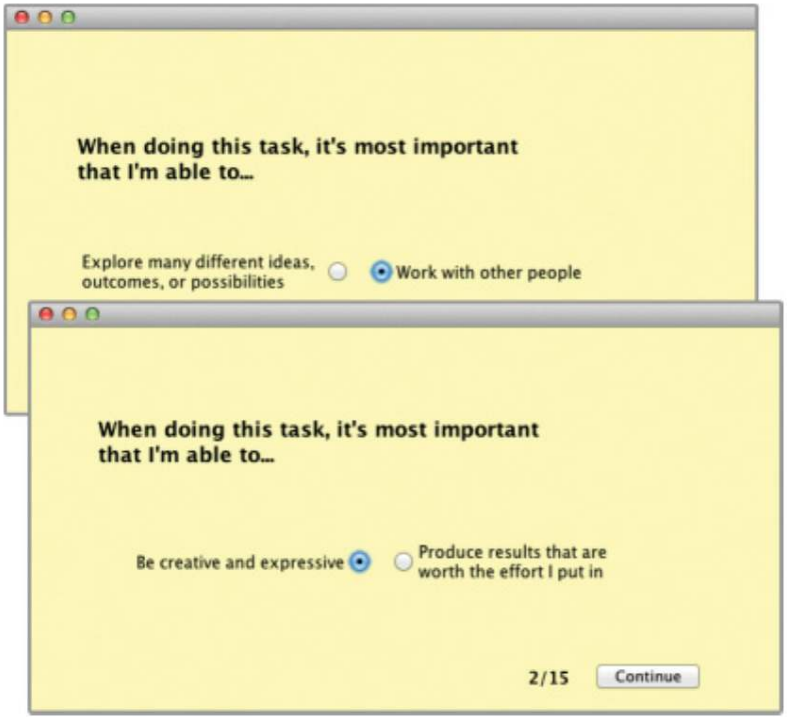
Cherry & Latulipe, TOCHI 2014, Table II

CSI part 2: paired comparisons

Table III. The Paired-Factor Comparison Test Has 15 Comparisons For each pair, a user will choose a factor description in response to the following statement: “When doing this task, it’s most important that I’m able to. . .”

1. Be creative and expressive
2. Become immersed in the activity
3. Enjoy using the system or tool
4. Explore many different ideas, outcomes, or possibilities
5. Produce results that are worth the effort I put in
6. Work with other people

15 pairs: pick the factor that matters more for this task



Cherry & Latulipe, TOCHI 2014, Table III and Fig. 2

CSI scoring

$$CSI = [(Collaboration1 + Collaboration2) * CollaborationCount +$$
$$(Enjoyment1 + Enjoyment2) * EnjoymentCount +$$
$$(Exploration1 + Exploration2) * ExplorationCount +$$
$$(Expressiveness1 + Expressiveness2) * ExpressivenessCount +$$
$$(Immersion1 + Immersion2) * ImmersionCount +$$
$$(ResultsWorthEffort1 + ResultsWorthEffort2) * ResultsWorthEffortCount] / 3.0$$

The score is out of 100.

Cherry & Latulipe, TOCHI 2014, Fig. 3 and Section 4.3

Example: writing in Google Docs

Table IV. CSI Results from a Collaborative Creative Writing Study Using Google Docs (N = 16)
The average CSI score for Google Docs in this study was 87.73 (SD = 11.30).

Scale	Avg. Factor Counts (SD)	Avg. Factor Score (SD)	Avg. Weighted Factor Score (SD)
Results Worth Effort	2.50 (1.41)	17.70 (2.30)	44.13 (27.38)
Exploration	2.56 (1.63)	16.00 (2.80)	41.00 (27.84)
Collaboration	3.63 (1.59)	18.13 (2.16)	67.63 (29.57)
Immersion	1.94 (1.34)	15.40 (4.94)	34.00 (26.00)
Expressiveness	3.19 (1.33)	17.63 (2.47)	55.44 (22.17)
Enjoyment	1.19 (0.91)	18.10 (2.19)	21.00 (16.69)

16 novice writers wrote a story together in Google Docs.

Counts: times picked, out of 5

Cherry & Latulipe, TOCHI 2014, Table IV and Section 4.5

Remy et al.: 113 papers reviewed

We found that only 24 % (27) of papers in our sample recruit domain expert participants to evaluate their CST intended for that particular area or task, as mentioned in the Results section.

To our surprise, only three of the total number of surveyed papers made use of the Creativity Support Index (CSI) [100].

41% of the evaluations lasted an hour or less.

Discussion of the readings

Bansal et al., Lee et al., Vaccaro et al.

1. When is “better than the person alone” good enough?

2. Whose view should decide what ships?

3. If everyone’s results look alike, did the tool work?

Activity: critique a test plan

1 Who	*two people from your group, and how you reach them*
2 Prototype	*Colab notebook, Wizard of Oz, or a working app*
3 Task	*one they would really do, with planted errors*
4 Comparison	*with and without your design, in which order?*
5 It works if	

Swap plans with a partner

Is there a comparison?

Are the errors planted?

Is it a real task?

What would count as working?

Name **one way** their test could fool them.

Milestone 2: Make it understandable

Due Thursday of week 5

Takeaways

1. The design people **prefer** can make errors **harder to catch**.
2. Compare the pair with the **person alone** and the **AI alone**.
3. **Plant errors** to see which wrong answers people accept.
4. Test with the **real task** that people will do.

Next: Uncertainty and Explanation

Lecture 8, on Thursday. Read before class:

- Ribeiro et al., KDD 2016
- Liao et al., CHI 2020
- Kim et al., FAccT 2024