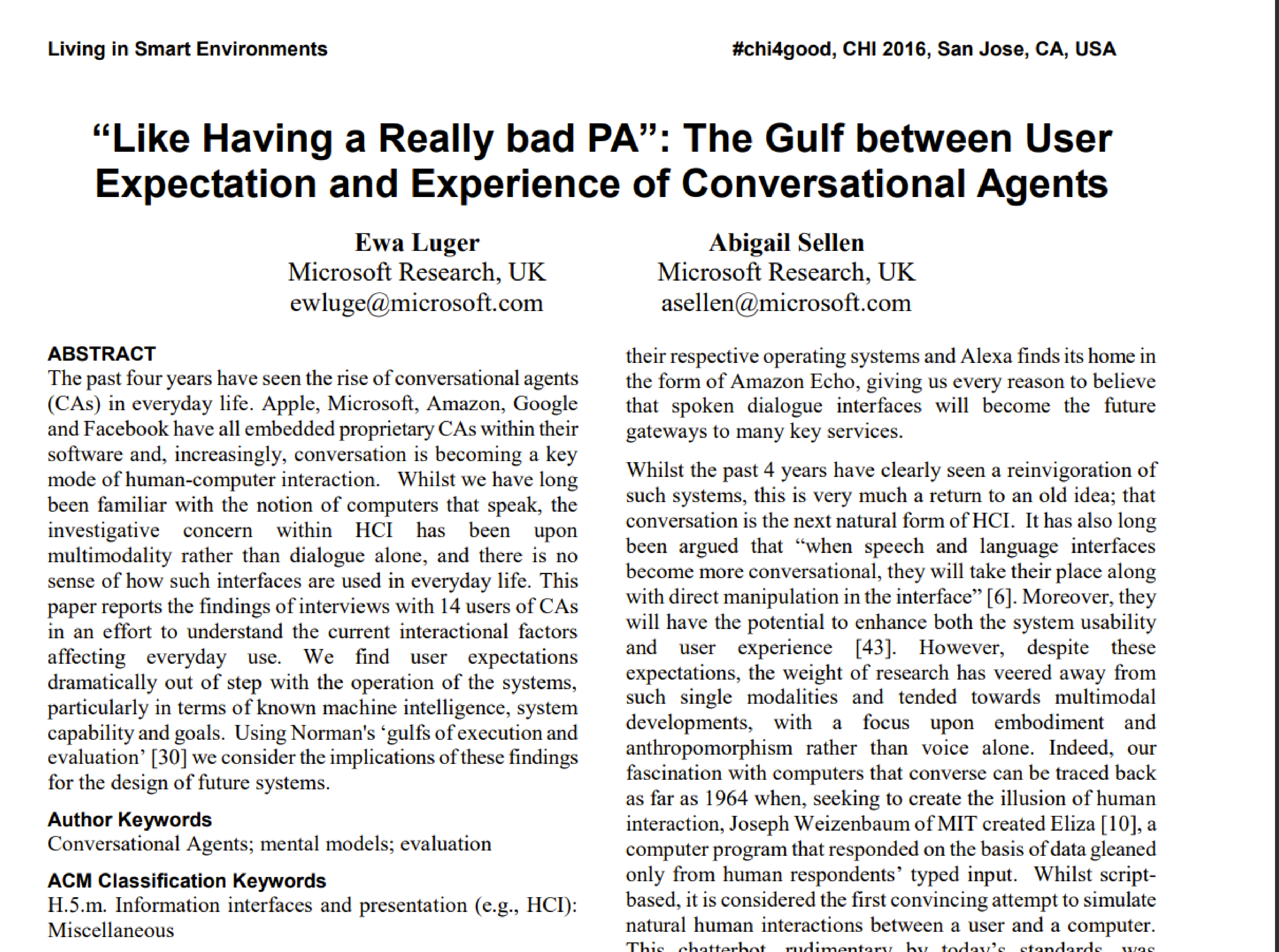
Proactive Speech Agents
(and how not to design them)

Justin Edwards – University of Oulu

slido

Join at slido.com
#1667378

ⓘ Start presenting to display the joining instructions on this slide.



"Like Having a Really bad PA": The Gulf between User Expectation and Experience of Conversational Agents

Ewa Luger
Microsoft Research, UK
ewluge@microsoft.com

Abigail Sellen
Microsoft Research, UK
asellen@microsoft.com

ABSTRACT

The past four years have seen the rise of conversational agents (CAs) in everyday life. Apple, Microsoft, Amazon, Google and Facebook have all embedded proprietary CAs within their software and, increasingly, conversation is becoming a key mode of human-computer interaction. Whilst we have long been familiar with the notion of computers that speak, the investigative concern within HCI has been upon multimodality rather than dialogue alone, and there is no sense of how such interfaces are used in everyday life. This paper reports the findings of interviews with 14 users of CAs in an effort to understand the current interactional factors affecting everyday use. We find user expectations dramatically out of step with the operation of the systems, particularly in terms of known machine intelligence, system capability and goals. Using Norman's 'gulfs of execution and evaluation' [30] we consider the implications of these findings for the design of future systems.

Author Keywords
Conversational Agents; mental models; evaluation

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

their respective operating systems and Alexa finds its home in the form of Amazon Echo, giving us every reason to believe that spoken dialogue interfaces will become the future gateways to many key services.

Whilst the past 4 years have clearly seen a reinvigoration of such systems, this is very much a return to an old idea; that conversation is the next natural form of HCI. It has also long been argued that "when speech and language interfaces become more conversational, they will take their place along with direct manipulation in the interface" [6]. Moreover, they will have the potential to enhance both the system usability and user experience [43]. However, despite these expectations, the weight of research has veered away from such single modalities and tended towards multimodal developments, with a focus upon embodiment and anthropomorphism rather than voice alone. Indeed, our fascination with computers that converse can be traced back as far as 1964 when, seeking to create the illusion of human interaction, Joseph Weizenbaum of MIT created Eliza [10], a computer program that responded on the basis of data gleaned only from human respondents' typed input. Whilst script-based, it is considered the first convincing attempt to simulate natural human interactions between a user and a computer. This chatterbot, rudimentary by today's standards, was

PhD Study 1 – Elicitation Paradigm

Design:

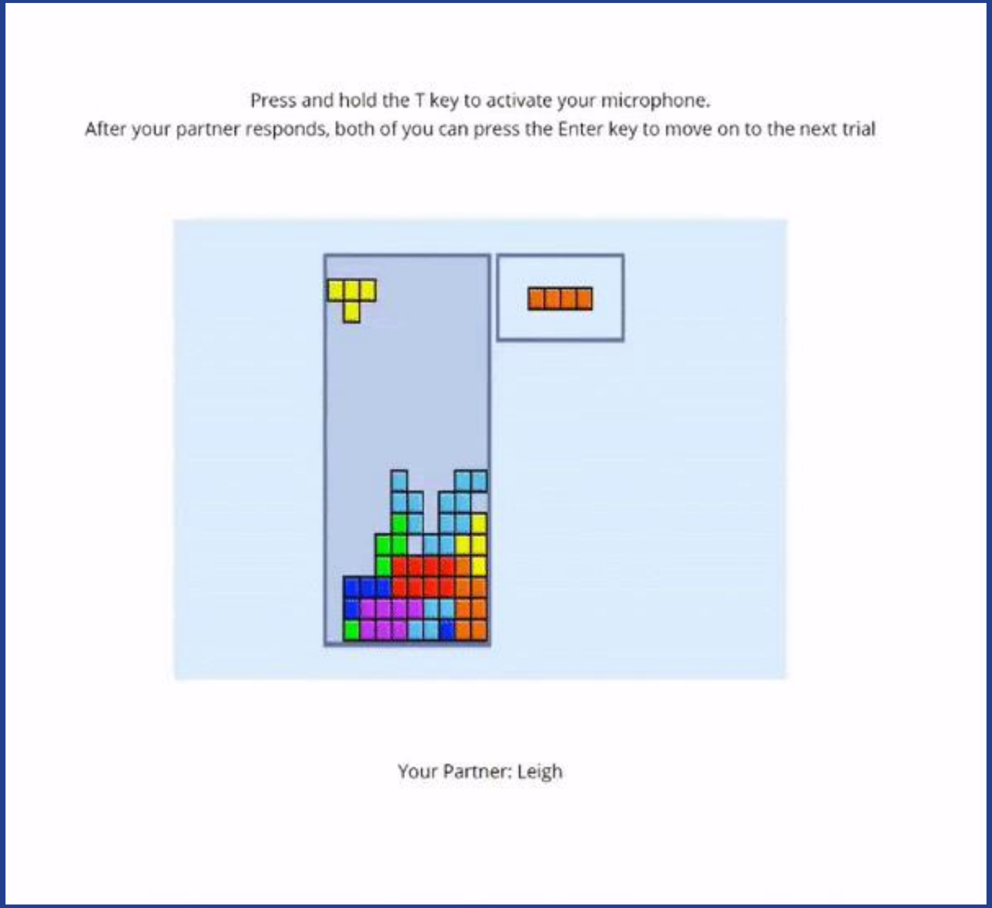
Participants interrupt a recorded "partner"

Tetris - complex continuous task

Spoken interruptions - questions based on Kubose et al., (2006)

Spoken replies from Irish (recorded) voices

Gamified scoring system, based on Landesberger et al., (2020)



Press and hold the T key to activate your microphone.
After your partner responds, both of you can press the Enter key to move on to the next trial

Your Partner: Leigh

Participants:

52 participants (26 W, 24 M, 2 prefer not to specify;

M_{age} = 29.4, SD = 7.9

AMT crowdworkers ($10 credit, ~20 minutes)

British and Irish, native English speakers

PhD Study 1 - Quantitative Hypotheses

Urgent interruptions will have …

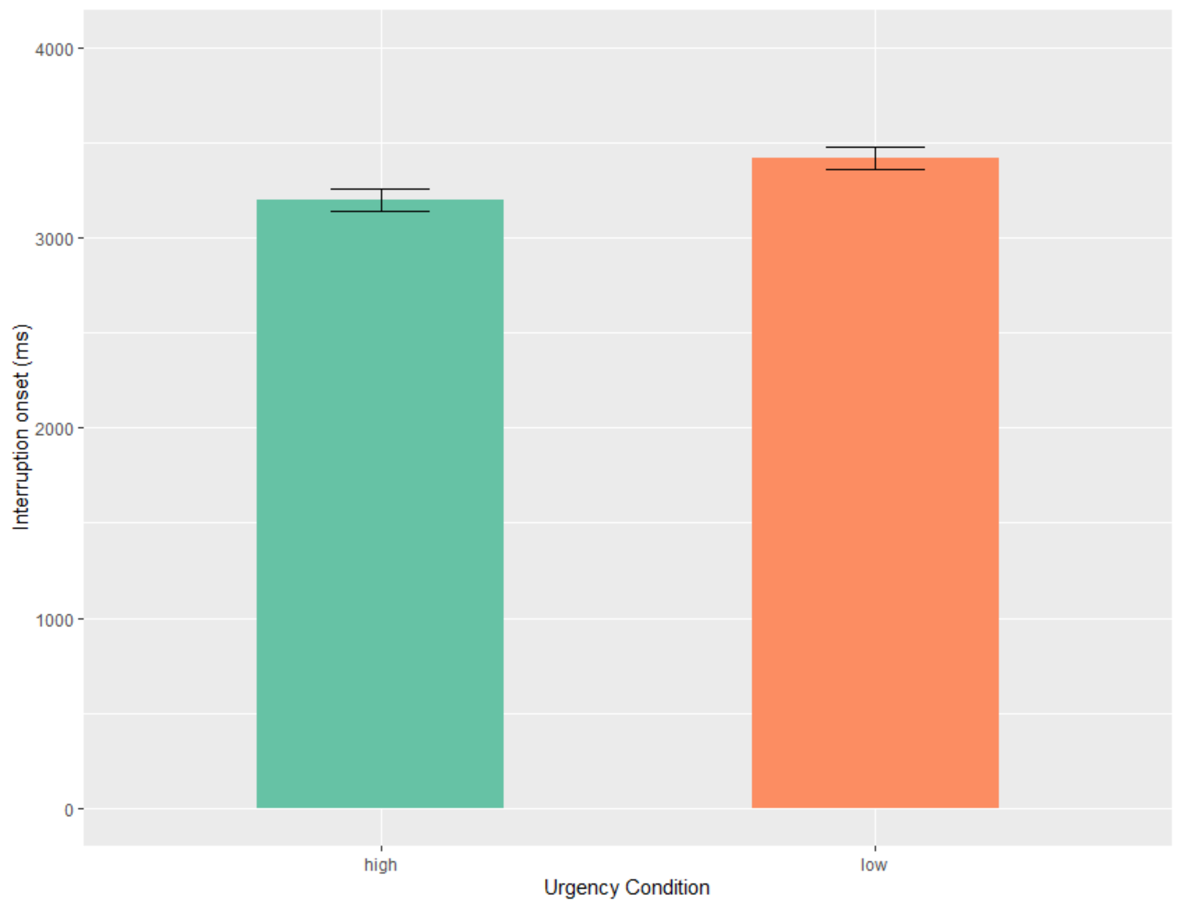
Delay ⬇️ (H1)

Duration ⬇️ (H2)

Access Rituals ⬇️ (H3)

… compared to non-urgent interruptions

PhD Study 1 – Quantitative Results



Urgent Interruption Delay ⬇ (H1):
(Unstandardized β=232.83, p=.04)

~~Urgent Interruption Duration ⬇ (H2):~~

~~Urgent Access Rituals ⬇ (H3):~~

PhD Study 1 – Timing Strategies

Prioritising Speed

“[I interrupted] as soon as possible, the timing of Tetris didn't occur to me” (P09)

Prioritising Accuracy

“[I] Took my time deciding on how to word and when to deliver the question” (P28)

“[I] just decided to say it casually. not make him feel like he needs to answer too quickly. for the low urgency trials.” (P44)

Tetris task characteristics

“As soon as a block was placed and a new one was at the top of the screen” (P12)

“[I interrupted] when I felt she had selected a spot for the falling piece.” (P29)

Message content

“I tried to wait until a piece had been played if it was a longer question, if it was a simple and short question I asked it straight away” (P51)

No strategy

PhD Study 1 – speech strategies

Phrasing

“I used as few words as possible, so she didn't have to think about it” (P15)

“[I asked] the questions I would normally ask an acquaintance.” (P23)

Delivery

“[I used] a calm voice to not startle my partner” (P24)

“[I spoke] clearly so she can understand.” (P47).

Message content

“[My priorities varied] based on the type of question.” (P13)

No strategy

I tried to make my questions as clear as possible, but in hindsight I think I probably should've made an effort to make my questions shorter as though I started when I thought it was a good time to talk, actually by the time I'd finished asking and it was time for her response it was in the middle of what I'd consider a high risk moment in the game!
-- (P16)

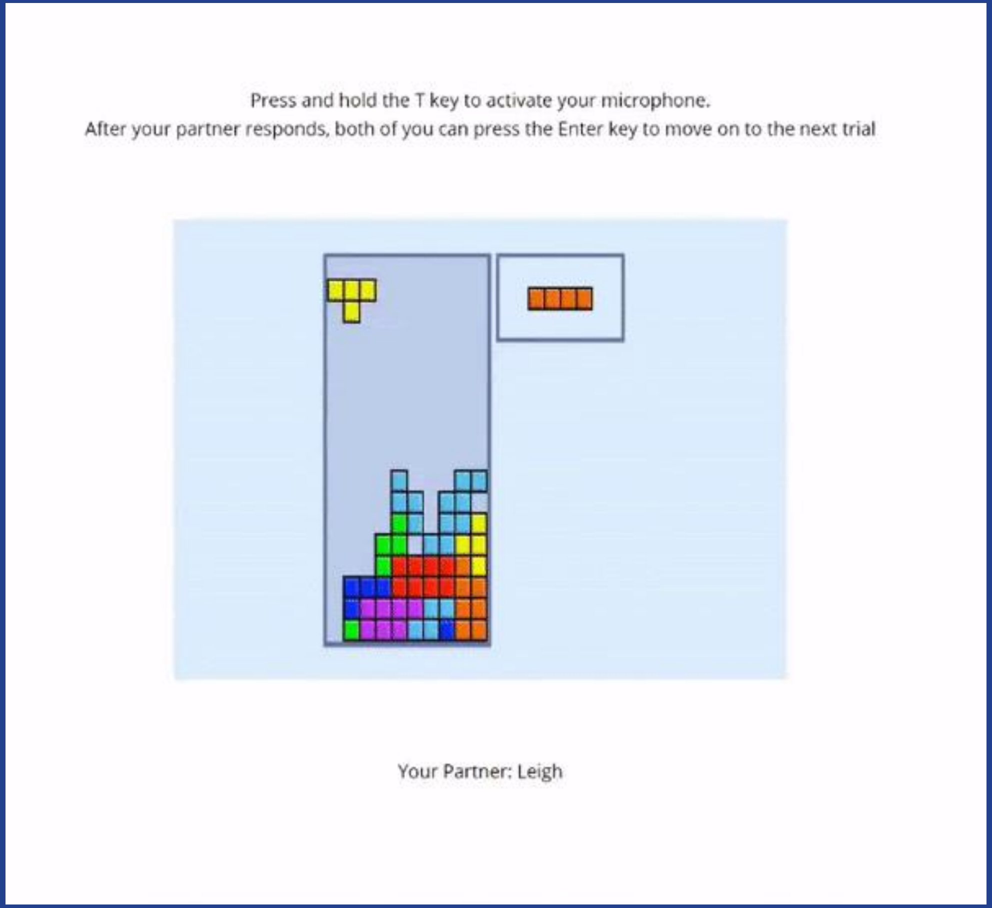
PhD Study 2 – Elicitation Paradigm

Design:

Same paradigm as study 1

2X2 design:
Explicit IV - Urgency (urgent vs not urgent)

Implicit IV – Game Difficulty (hard vs easy games)



Press and hold the T key to activate your microphone.
After your partner responds, both of you can press the Enter key to move on to the next trial

Your Partner: Leigh

Participants:

93 participants (45 W, 48 M)

M_{age} = 33.9, SD = 10.2

Crowdworkers ($10 credit, ~20 minutes)

British and Irish, native English speakers

PhD Study 2 - Quantitative Hypotheses

Easy games will have

Delay ⬇️ (H1) Duration ⬇️ (H2)

… compared to interruptions of difficult Tetris games.

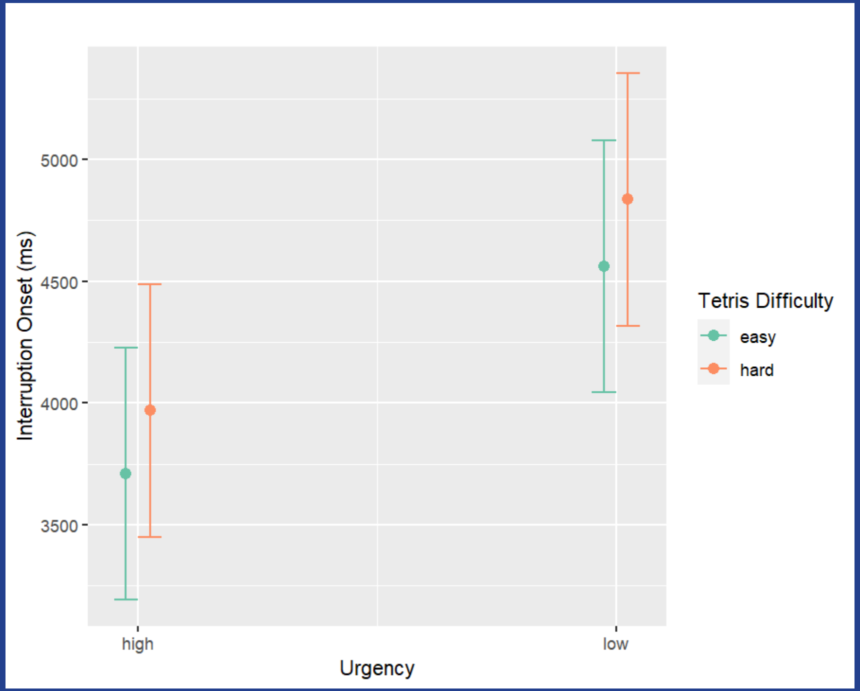
Urgent interruptions will have …

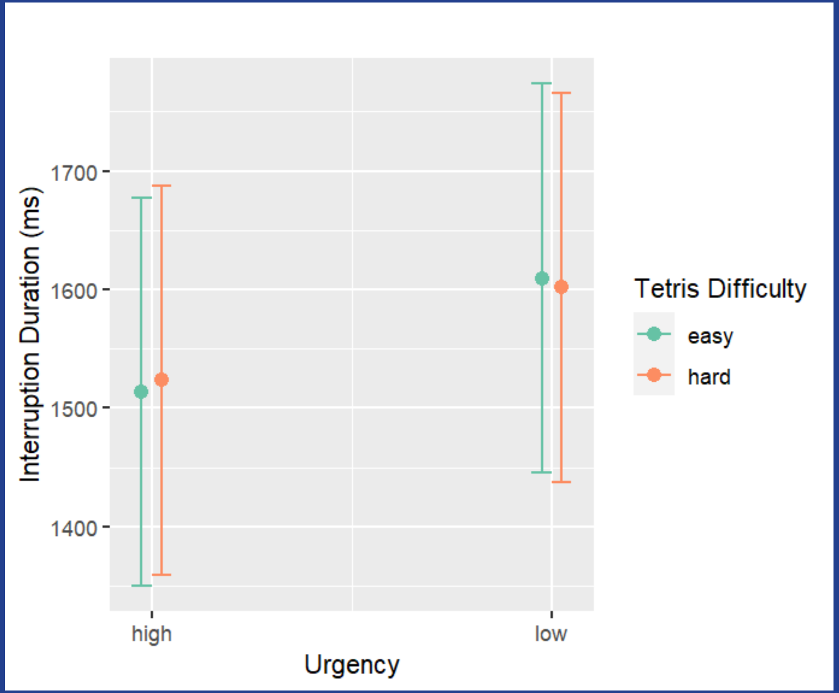
Delay ⬇️ (H3) Duration ⬇️ (H4)

… compared to non-urgent interruptions.

Easy games will have Access rituals ⬆️ compared to interruptions of hard Tetris games (H5)

Study 2 – Quantitative Results





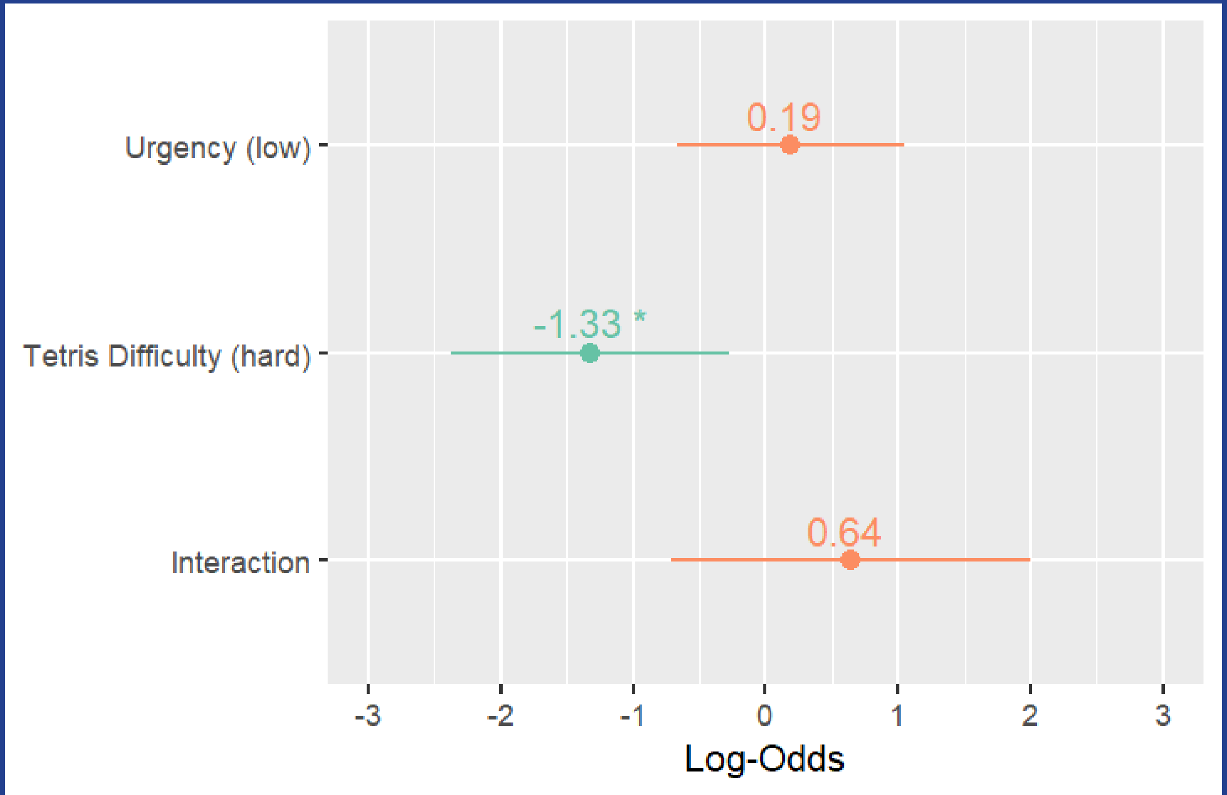
~~H1 – Easy games delay~~ ⬇️

~~H2 – Easy games duration~~ ⬇️

H3 – Urgent interruption delay ⬇️
(Unstandardized β=851.80, p<.001)

H4 – Urgent interruption duration ⬇️
(Unstandardized β=96.03,, p<.01)

PhD Study 2 – Quantitative Results



H5 – Easy Tetris games access rituals ⬆️
(Log-odds = 1.33, p=.01)

PhD Study 2 – Qualitative themes

Interruption timing strategies:

Interrupting as soon as possible

“There’s never going to be a good time, so [I] just go for it.” (P02)

Interrupting at an appropriate moment

“I picked the point where the other person just laid their last brick down, so they had the most time before the next needed placing.” (P20)

No strategy

Interruption structure strategies:

Communicating urgency

“I tried to word the question so to almost direct them to a quick answer not to give them too much scope for having to think through a variety of possible answers.” (P84)

Communicating calmness

“I decided to be laidback. The message was non-urgent and so there was no reason for my tone to be urgent either.” (P47)

No strategy

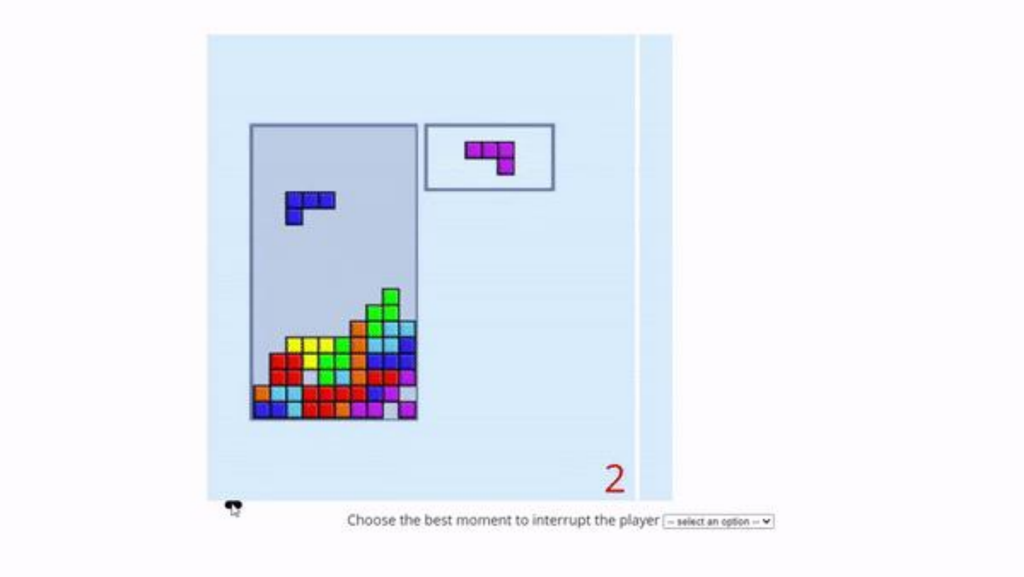
PhD Study 3 – Categorizing interruptible moments

Design:

Participants select a moment to begin spoken interruption

Participants first watch full-speed clips of Tetris game, then select best still frame for starting an interruption

Participants have unlimited time to select specific game moment interrupt



Choose the best moment to interrupt the player -- select an option --

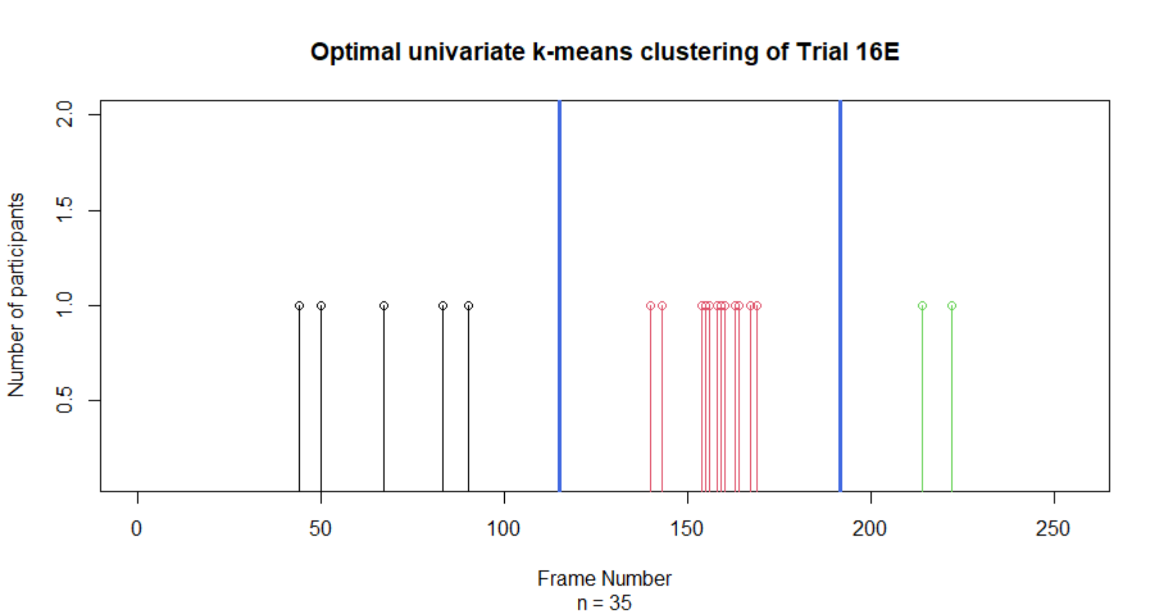
Experimental aims:

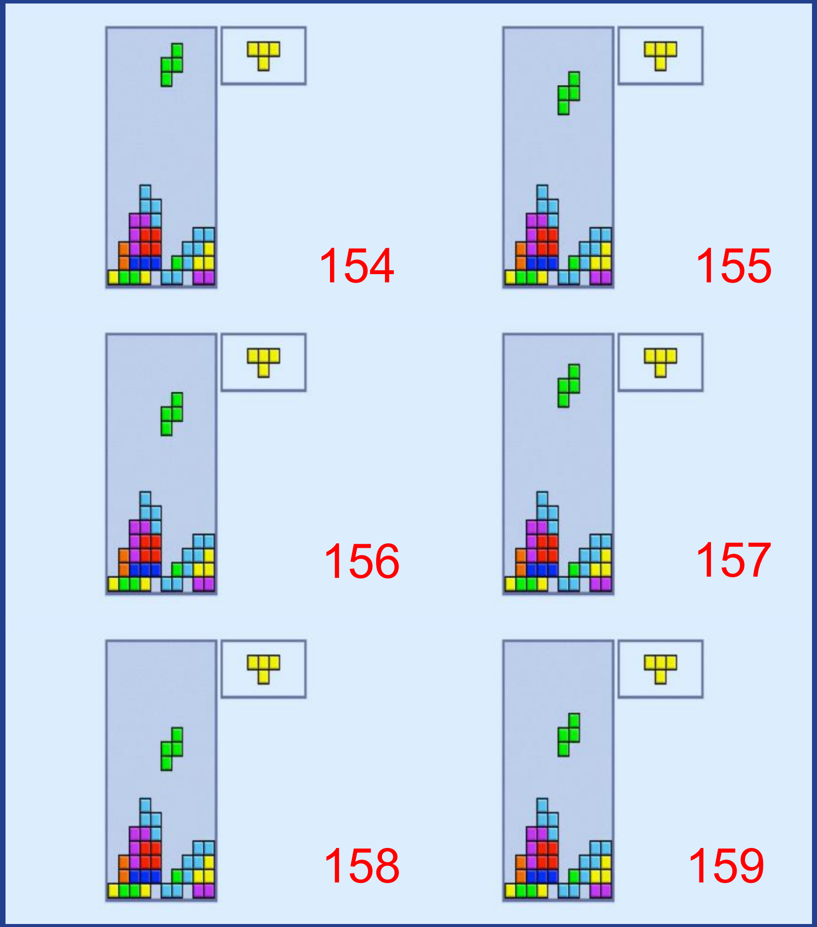
Clusters of adjacent selections illustrate interruptible windows

Analysis of common themes between these interruptible windows yields general descriptions of interruptible moments

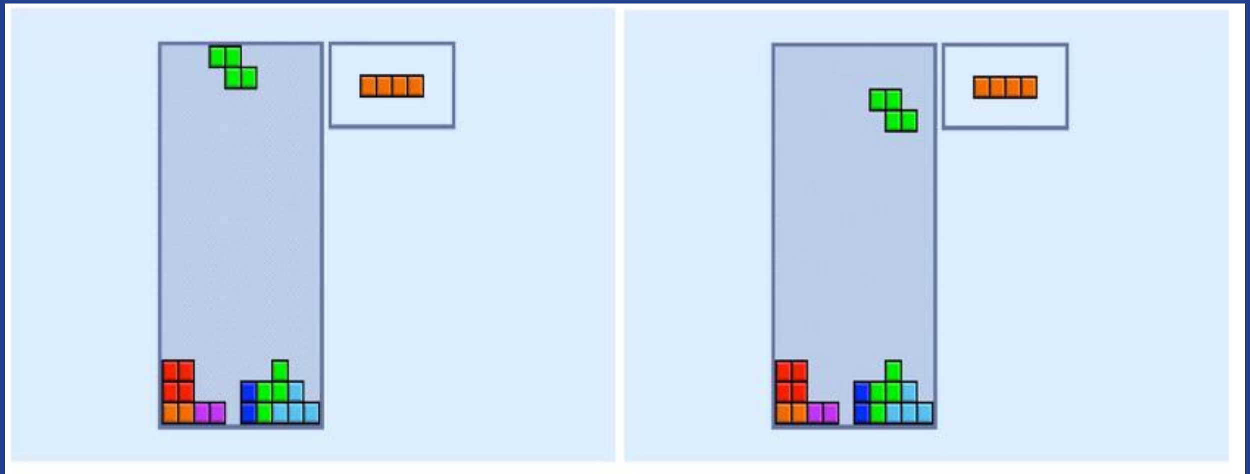
Categorization of interruptible windows allows for analysis of prior interruption data

PhD Study 3 – Quantitative Results

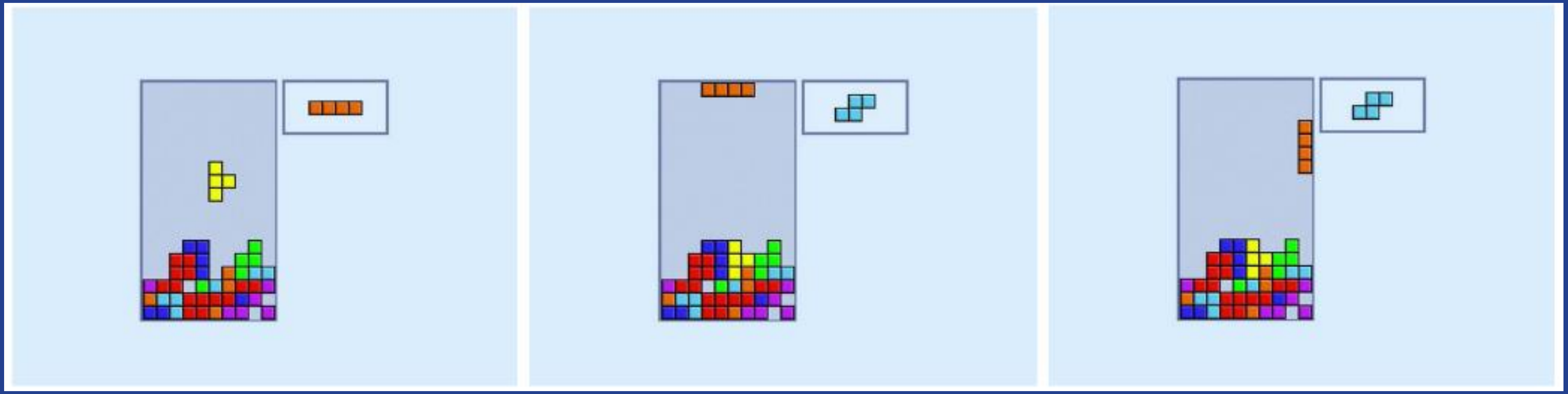




PhD Study 3 – Quantitative Results

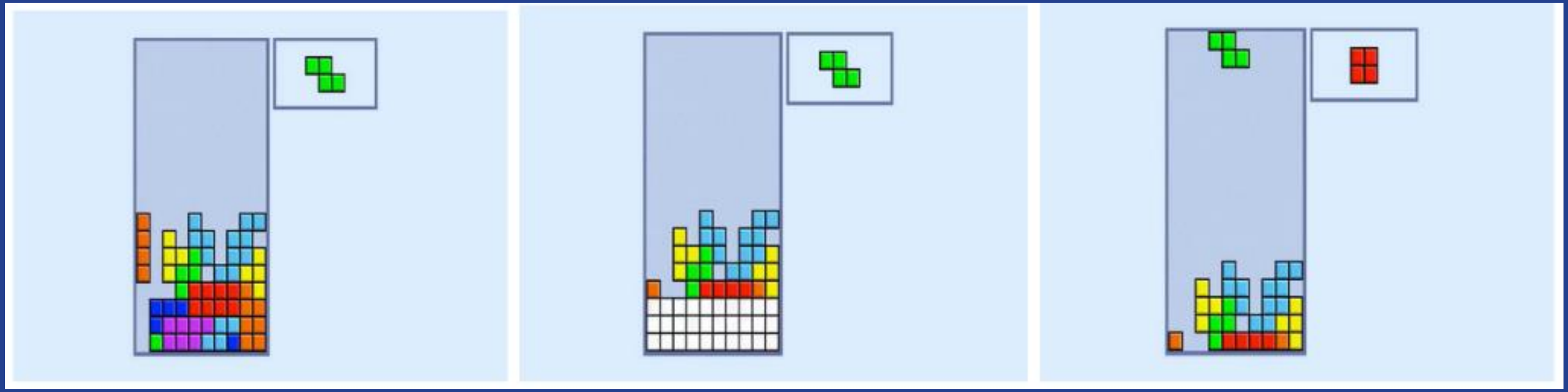


No spin

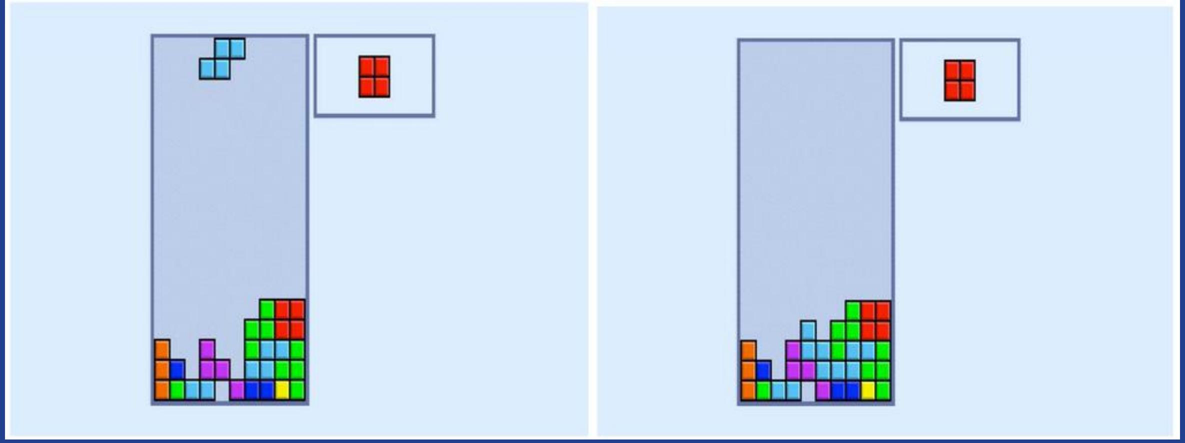


One spin

Study 3 – Quantitative Results



Line Clear



Calm episode

Study 3 – Quantitative Results

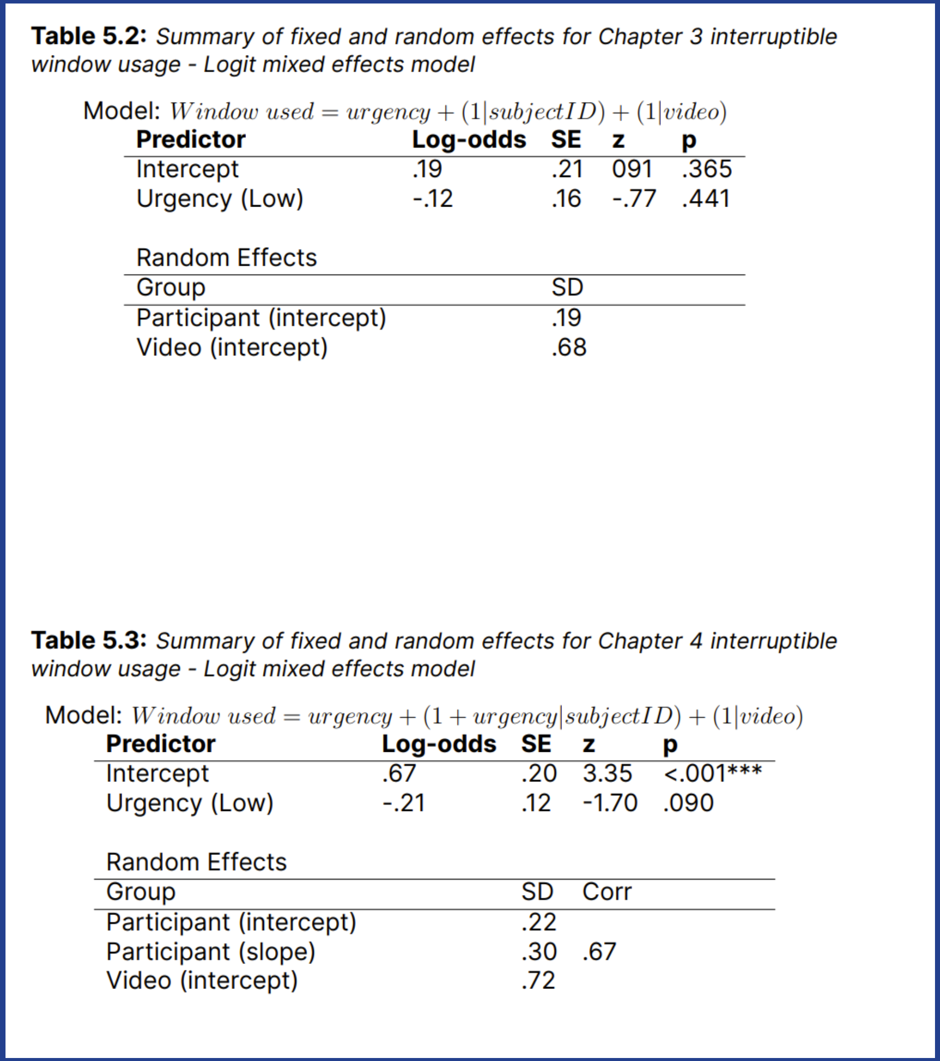


Table 5.2: *Summary of fixed and random effects for Chapter 3 interruptible window usage - Logit mixed effects model*

Model: $Window\ used = urgency + (1|subjectID) + (1|video)$

Predictor	**Log-odds**	**SE**	**z**	**p**
Intercept	.19	.21	091	.365
Urgency (Low)	-.12	.16	-.77	.441

Random Effects

Group	SD
Participant (intercept)	.19
Video (intercept)	.68

Table 5.3: *Summary of fixed and random effects for Chapter 4 interruptible window usage - Logit mixed effects model*

Model: $Window\ used = urgency + (1 + urgency|subjectID) + (1|video)$

Predictor	**Log-odds**	**SE**	**z**	**p**
Intercept	.67	.20	3.35	<.001***
Urgency (Low)	-.21	.12	-1.70	.090

Random Effects

Group	SD	Corr
Participant (intercept)	.22	
Participant (slope)	.30	.67
Video (intercept)	.72	

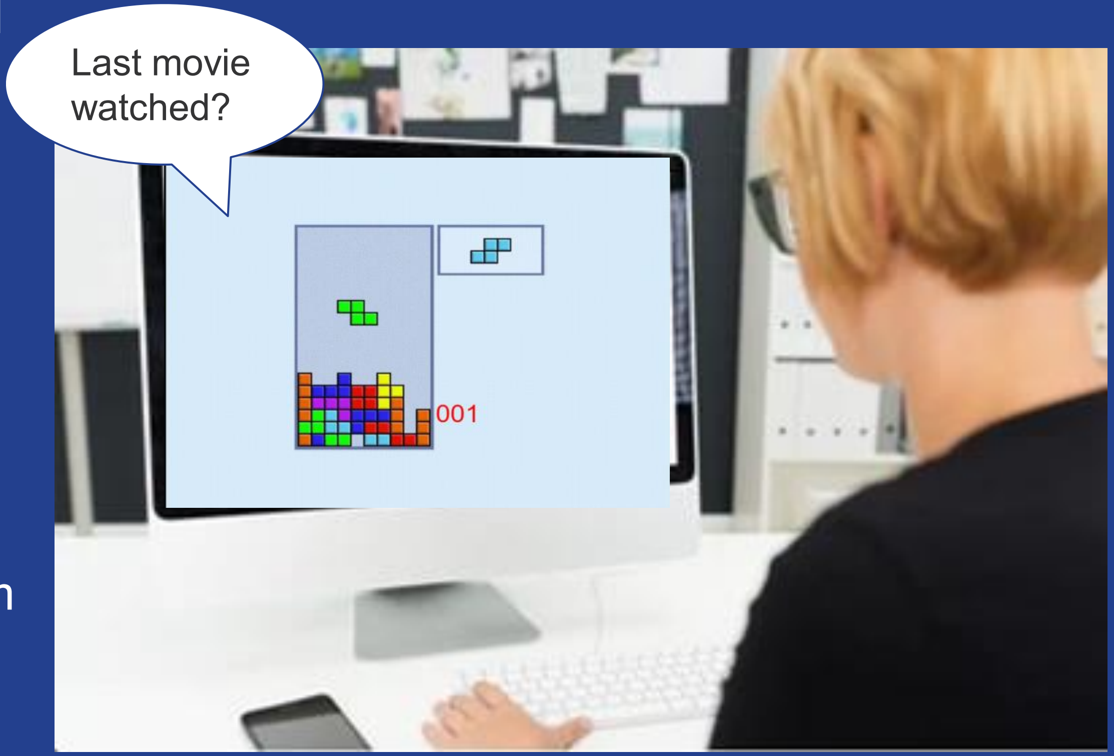
PhD Study 4 – Perceptions of humanlike interruptions

Design:

Participants watch a recorded Tetris game

A speech agent interrupts the player with questions like in studies 1 and 2

Two agent conditions: agent either considers the cues from prior experiments and adapts like human interrupters or ignores them



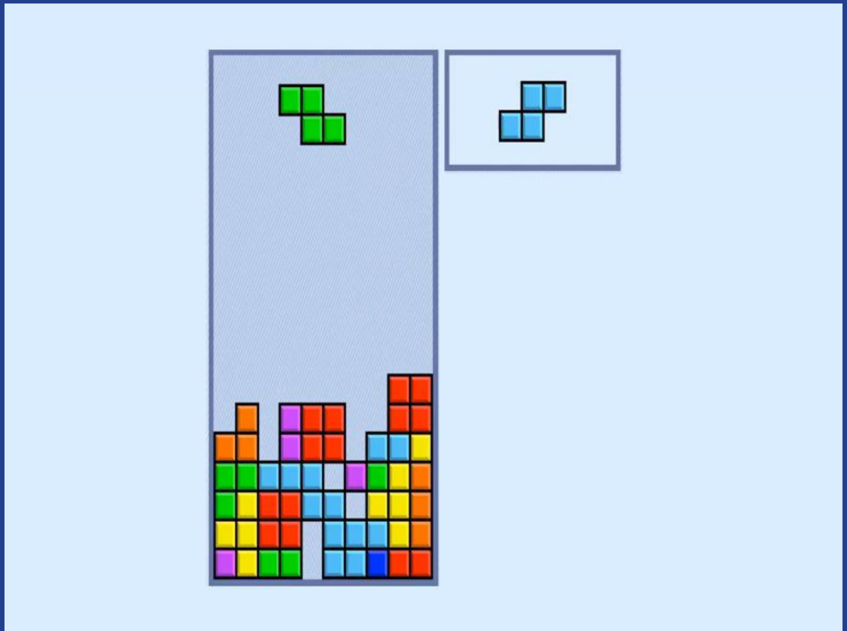
Participants:

Quantitative comparison of partner models for adaptive and static agents

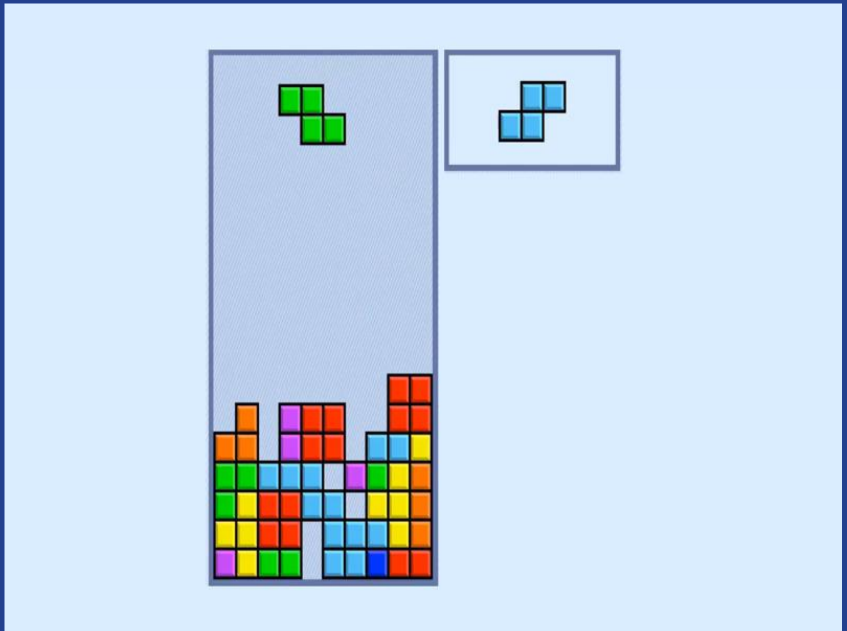
Quantitative comparison of perception of agents' interruption timing and style

Qualitative comparison of perceptions of each agent

PhD Study 3 – Agent examples



Static



Adaptive

PhD Study 4 - Quantitative Hypotheses

Adaptive agent interruptions will be rated as …

Timing (H1) ⬆️

Appropriateness ⬆️ (H2)

… compared to interruptions from a static agent.

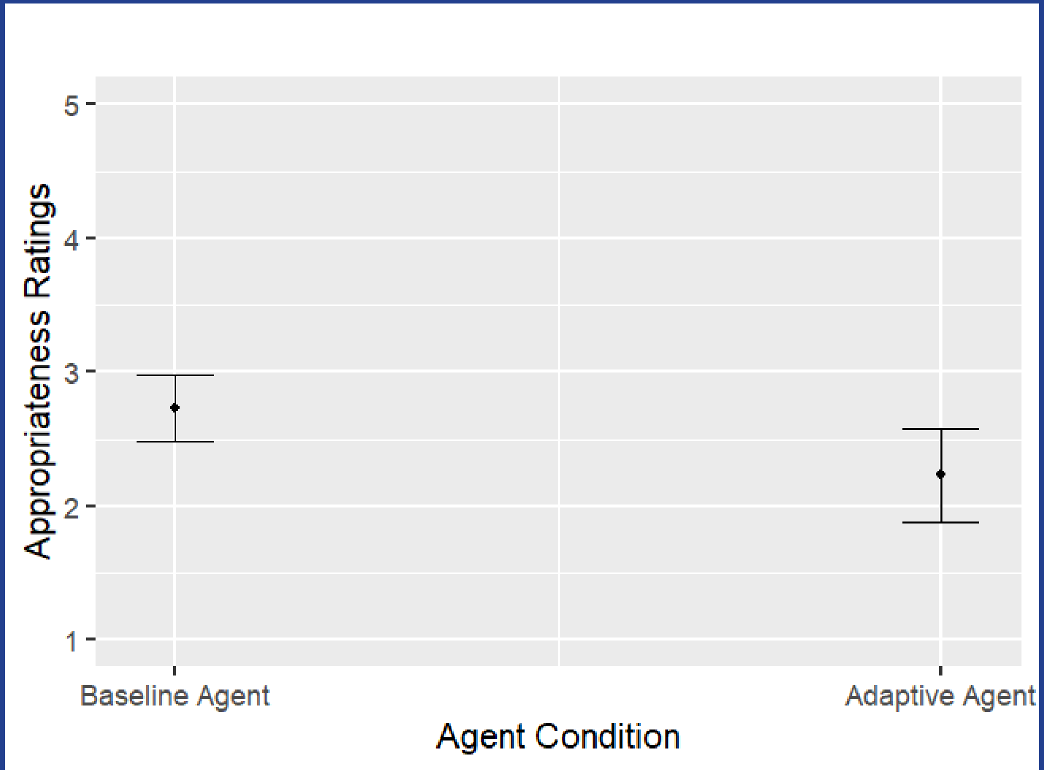
Adaptive agents will be seen as having …

capability as a dialogue partners (H3) ⬆️

… compared to interruptions static agents.

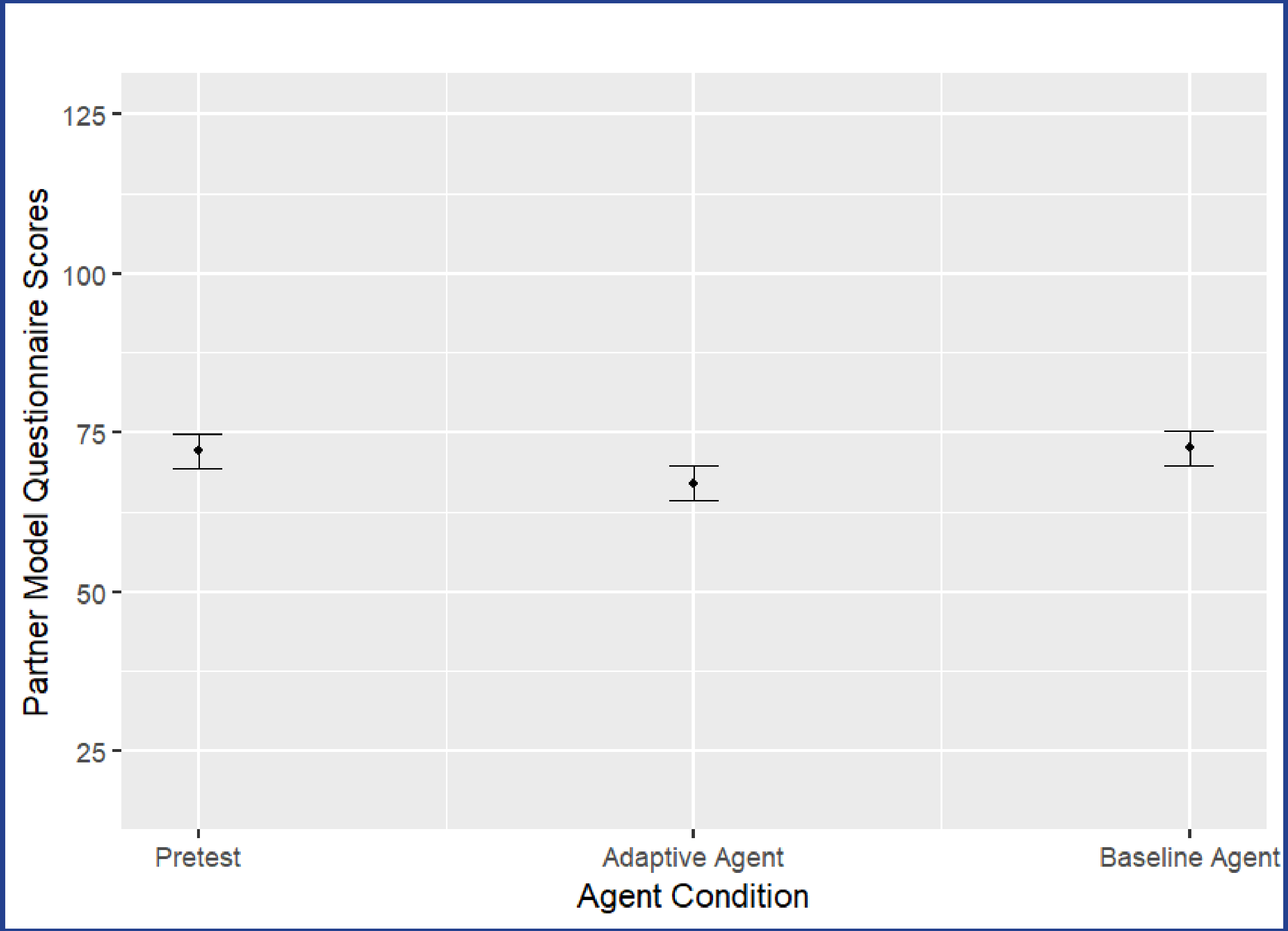
Study 4 – Quantitative Results

PhD Study 4 – Quantitative Results



~~Timing (H1) ⬆️~~

~~Appropriateness ⬆️ (H2)~~



~~Capability as a dialogue partners (H3) ⬆️~~

PhD Study 4 – Qualititative Results

Clarity, directness, politeness

“The [adaptive] was much more pleasant and polite, the [static] was a lot more cold and blunt” (P08)

Rushed speech

“the [adaptive] one was more informal and rushed. it also spoke in less cohesive sentences.” (P23)

Appropriateness

“I found the [static] more appropriate with tone, and asked the questions in a slower more human way.” (P02)

Human mimicry

“The [adaptive] speech assistant seemed to be making an attempt at sounding more human. I think it failed in this and the attempts became a tad annoying.” (P10)

Inconsistency

“The [adaptive] one was inconsistent, some questions were read very quickly at times” (P73)

“The [adaptive] was unclear and inconsistent. The [static] was far better” (P21)

Describing without judging

“[The adaptive] version varied speed of question depending on urgency and also inserted extra comments like "excuse me".” (P80)

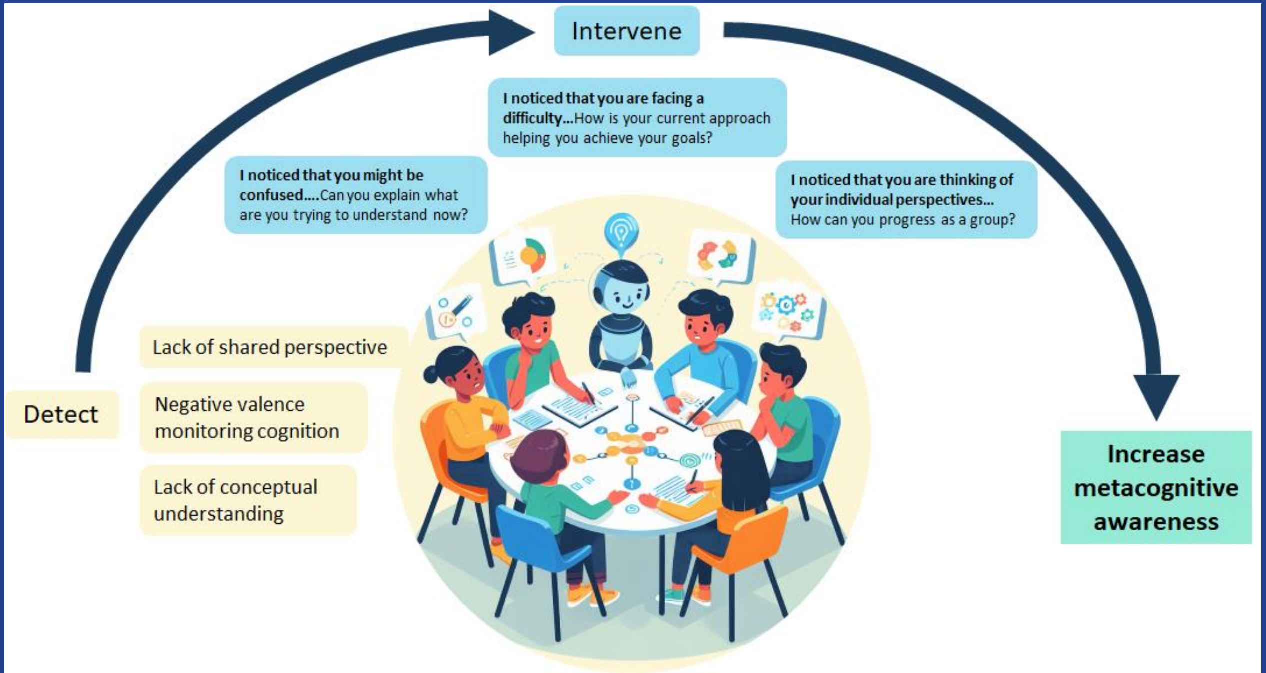
No differences

“I actually thought they were the same. I think i was concentrating more on the task.” (P35).

Learning Regulation with AI – Promoting Adaptive K-12 Learners (LEAD)

Sanna Järvelä, Andy Nguyen, Joni Lämsä, Marta Sobocinski, Justin Edwards, Ridwan Whitehead, Belle Dang, Anni-Sofia Roberts

Designing MAI – a Metacognitive AI agent


Intervene
I noticed that you are facing a difficulty...How is your current approach helping you achieve your goals?
I noticed that you might be confused....Can you explain what are you trying to understand now?
I noticed that you are thinking of your individual perspectives... How can you progress as a group?
Detect
Lack of shared perspective
Negative valence monitoring cognition
Lack of conceptual understanding
Increase metacognitive awareness

Trigger events: challenging cognitive or emotional moments that require a strategic response (and adaptation) from the group

MAI is a **metacognitive AI agent**

Prompts learners when it detects:

- Lack of shared perspective
- Difficulties progressing towards task goals (negative valence in monitoring)
- Lack of conceptual understanding

MAI Study 1 research design

52 pre-service teachers

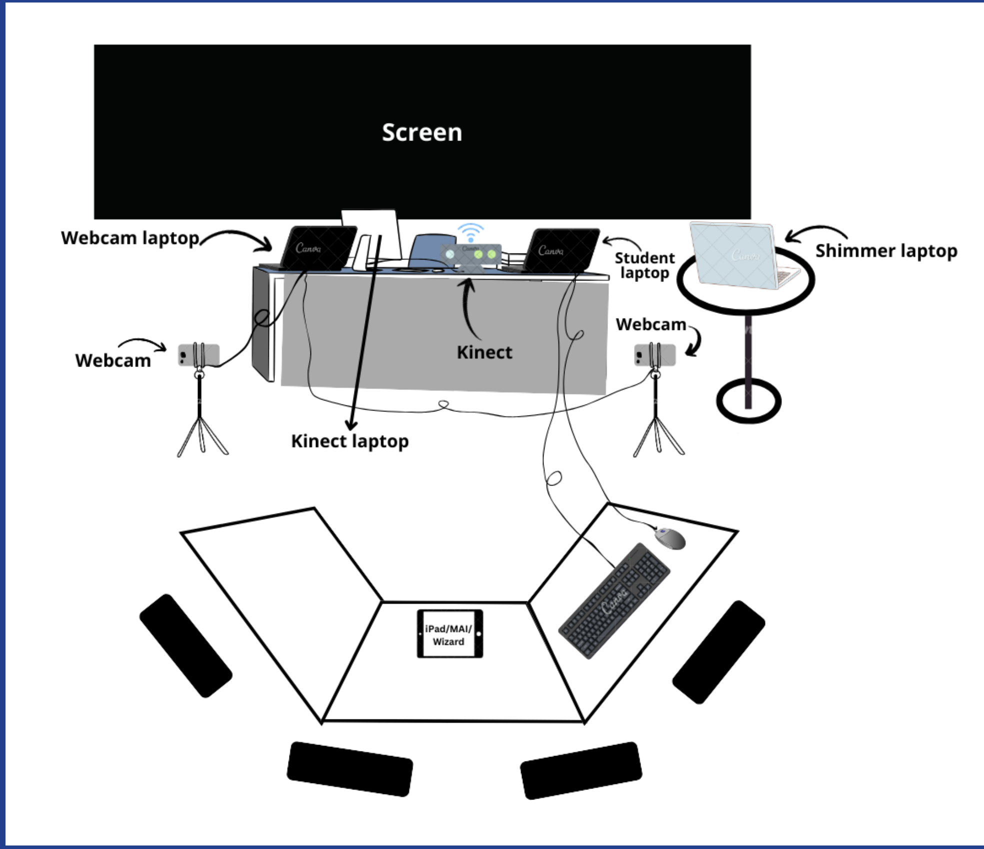
3-4 member groups

60 min collaborative inquiry learning task on physics

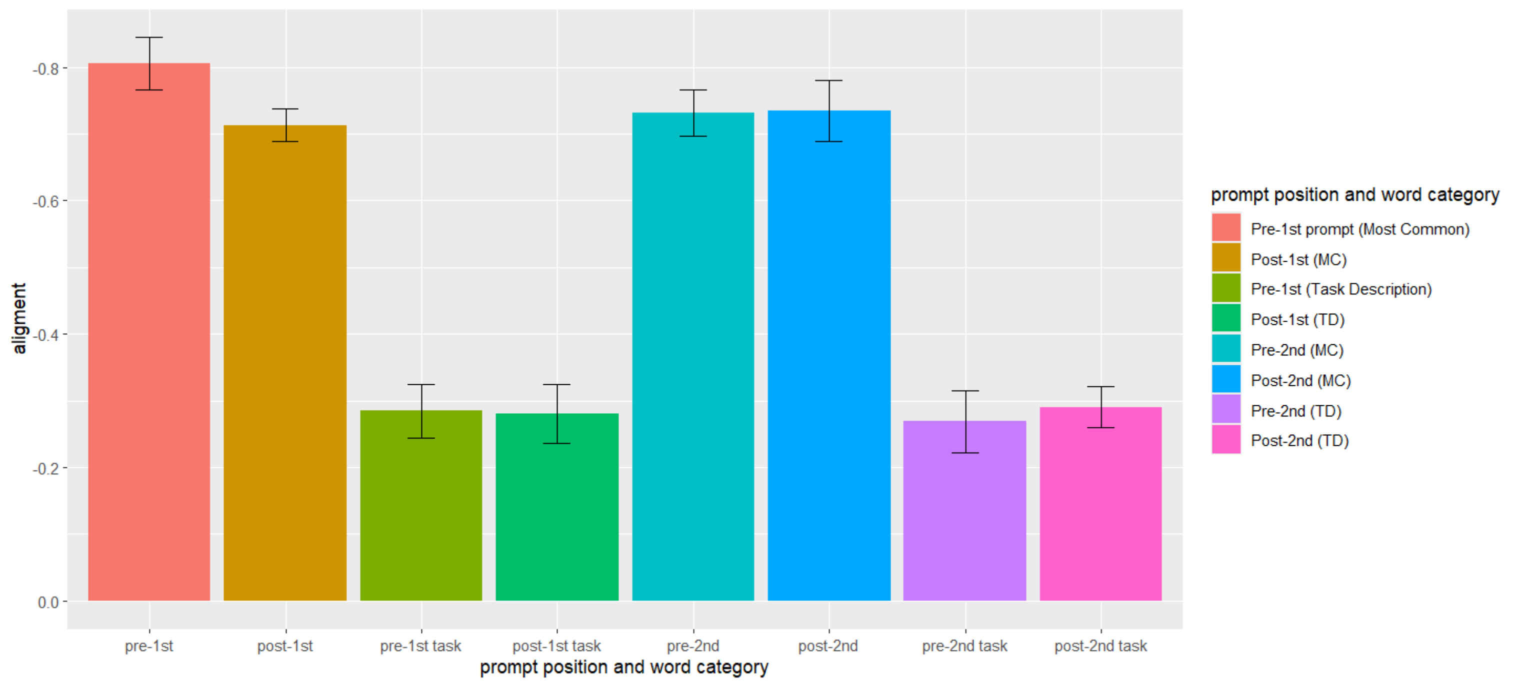
- Collaborative learning script

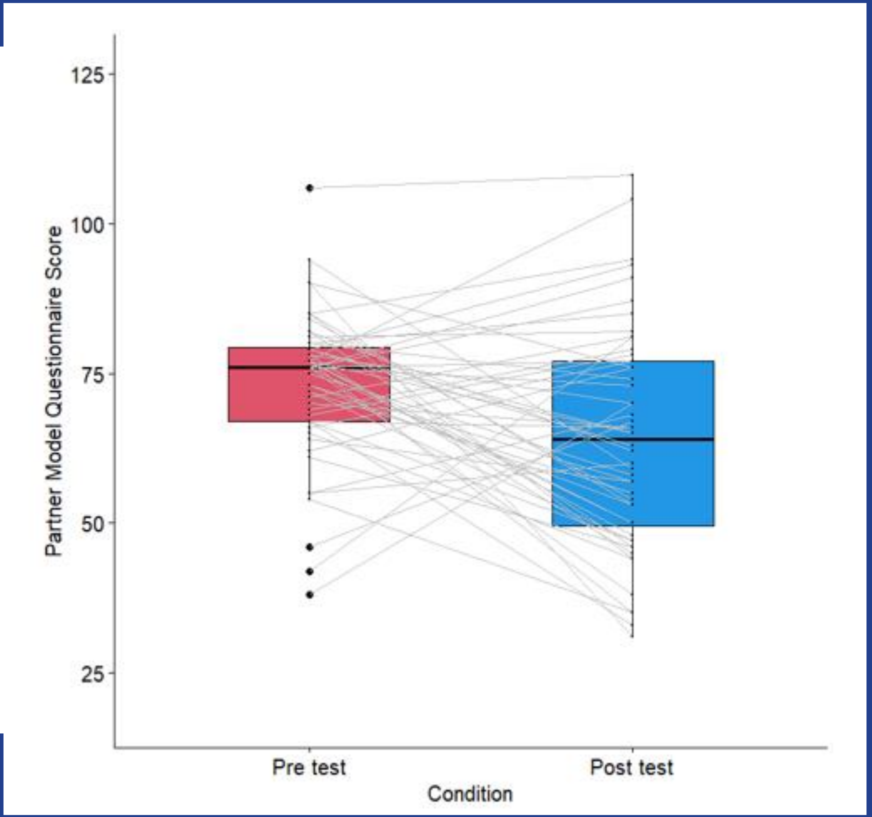
  1. introductory video to the topic (5 min)
  2. simulation with a set of guiding questions and answer sheet (University of Colorado Boulder, 2023; 20 min)
  3. an experimental, hands-on task (25 min).
  4. reflective discussion on students' ideas and thoughts on using these kinds of learning activities in their teaching practices (10 min)

All the learning activities and their instructions in GoLab learning environment (University of Twente, 2023).



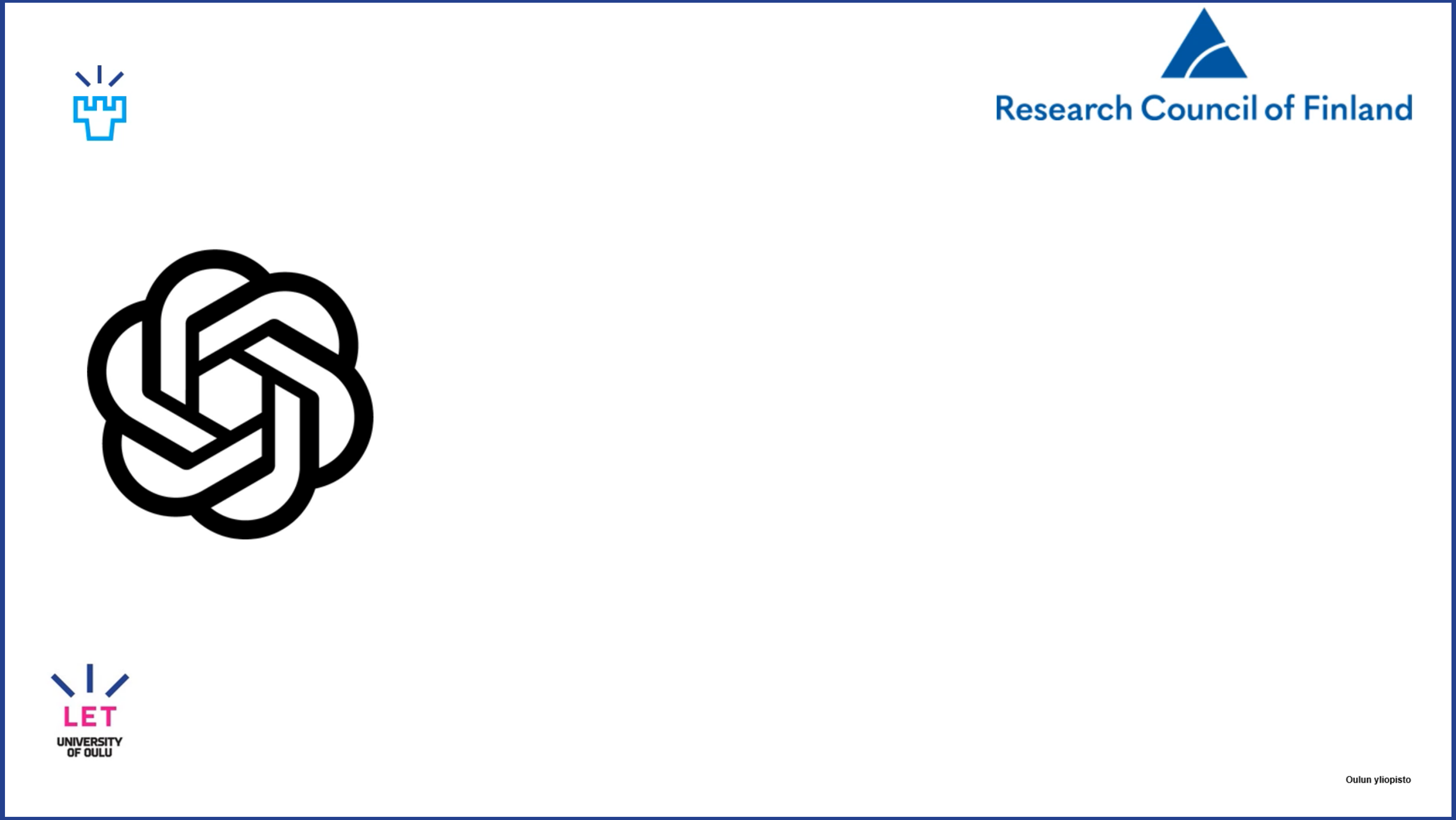
MAI Study 1 findings





Edwards, J., Nguyen, A., Lämsä, J., Sobocinski, M., Whitehead, R., Dang, B., Roberts, A.-S., & Järvelä, S. (2024). Human-AI collaboration: Designing artificial agents to facilitate socially shared regulation among learners. *British Journal of Educational Technology*, 00, 1–22. https://doi.org/10.1111/bjet.13534

MAI Study 2



Research Council of Finland

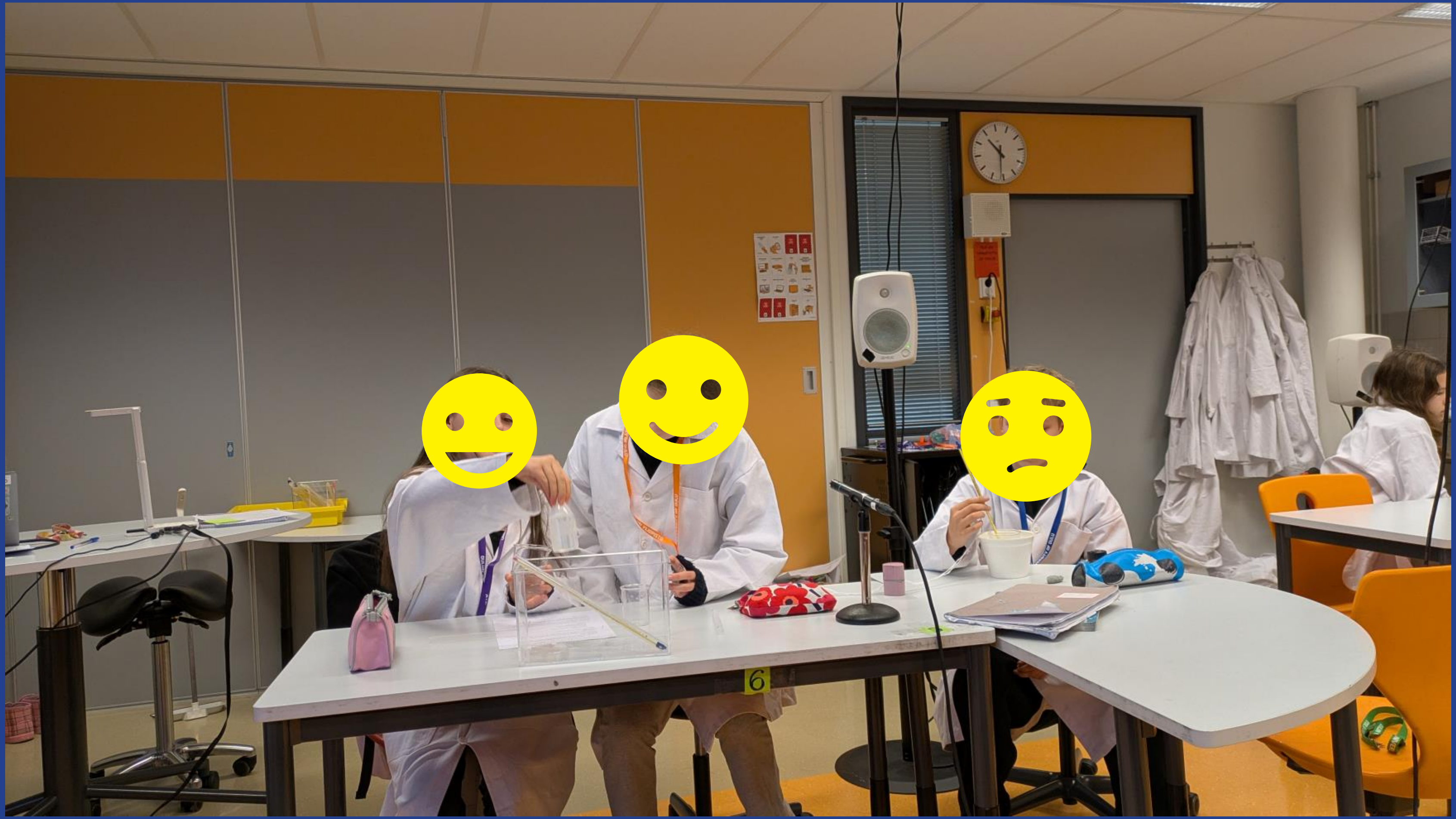
LET
UNIVERSITY
OF OULU

Oulun yliopisto

MAI’s new prompting rules

Metacognitve trigger Rule: IF what group members express is semantically similar to confusion, related to the task instructions, tools, plans in one episode, over 3 turns of speaking, THEN prompt: e.g. • Hmm, are you having trouble? Describe to each other how you understand the problem at the moment. Do you all agree on how to take this task forward? • Discuss together how your current activities are helping you achieve your goal. This will enable you to move forward in a difficult situation. • Talk together about how what you are doing now will help you in this task. This will enable you to move on from a difficult situation.	Cognitive trigger Rule: IF what group members express is semantically similar to confusion, related to the content understanding, over 3 turns of speaking THEN prompt: • There seems to be confusion in the air. Explain to each other what you are trying to understand at the moment. • As I listen to you, I wonder if you all understand the task in the same way. Consider together what the key concepts of the task are. Good luck with reflection! • Discuss again how confident you are that you have everything necessary to accomplish the task. Share thoughts about what could help you move forward
. Behavioural triggers Rule: IF the distribution of conversational turns is above a given inequality threshold • Wait! I noticed that not everyone necessarily participates in the assignment. How can you ensure that everyone can participate? • Hmm – I wonder if everyone has had a chance to make their voices heard? • Let's think for a moment – what does everyone think of their own participation so far? • How could you change the way you do things today to ensure that everyone is involved?	Socio-emotional triggers Rule: IF on group member has not participated in discussion over the last 5 minutes • It seems that emotions are running a little hot. How could you have communicated in a more friendly conversation? • Cooperation is sometimes challenging! How could you ensure that everyone is still in good spirits? • Think for a moment—how you might express your thoughts without offending anyone else.
Lack of a common perspective: Rule: IF group members use of first or second person singular (”I”/”You”) repeatedly in on-task communication: • Great progress! However, are you still thinking from your own point of view? Consider how you might progress as a group. • I noticed that you may still think about it from your own perspective. I would encourage you to discuss what your common plan is for solving this task.	

MAI Study 3 – Conversational AI in school



Preliminary results

Thanks to…

Thanks to…

Thanks to…

Thank you for listening!

Contact me!

Justin Edwards

justin.edwards@oulu.fi

justinhci.github.io

slido

Please download and install the
Slido app on all computers you use

Audience Q&A

ⓘ Start presenting to display the audience questions on this slide.