Comparing Paid and Volunteer Recruitment in Human Computation Games

Anurag Sarkar and **Seth Cooper**

College of Computer and Information Science

Northeastern University

Crowdsourcing

- Paid crowdsourcing platforms like Amazon Mechanical Turk are popular for recruiting participants

Crowdsourcing

- Paid crowdsourcing platforms like Amazon Mechanical Turk are popular for recruiting participants

- In games, used for recruiting participants for playtesting, design experiments, user research *(Khajah et al., 2016; Sarkar et al., 2017; Sharek and Weibe, 2014; Birk and Mandryk, 2016; Weibe et al. 2014; Birk et al., 2017; Williams et al., 2017)*

amazon mechanical turk™
Artificial Artificial Intelligence

Recruitment Strategy

- Behaviors and motivations of paid participants may differ from those who play voluntarily (i.e. through banner ads, web search, social media posts etc.) *(Cooper and Farid, 2016; Crump et al., 2013; Paolacci et al., 2010; Sprouse, 2011; Krause and Kizilcec, 2015; Mao et al., 2013)*

Recruitment Strategy

- Behaviors and motivations of paid participants may differ from those who play voluntarily (i.e. through banner ads, web search, social media posts etc.) *(Cooper and Farid, 2016; Crump et al., 2013; Paolacci et al., 2010; Sprouse, 2011; Krause and Kizilcec, 2015; Mao et al., 2013)*

- Often, we wish to understand volunteers but end up studying paid participants

Recruitment Strategy

- Behaviors and motivations of paid participants may differ from those who play voluntarily (i.e. through banner ads, web search, social media posts etc.) *(Cooper and Farid, 2016; Crump et al., 2013; Paolacci et al., 2010; Sprouse, 2011; Krause and Kizilcec, 2015; Mao et al., 2013)*

- Often, we wish to understand volunteers but end up studying paid participants

- Wanted to compare the impact of recruitment strategy (i.e. paid vs volunteer) on player's engagement and subjective experience in the context of human computation games (HCGs)

Research Questions

- *RQ1 – Does recruitment strategy impact participant behavior and experience in HCGs?*

Research Questions

- *RQ1 – Does recruitment strategy impact participant behavior and experience in HCGs?*

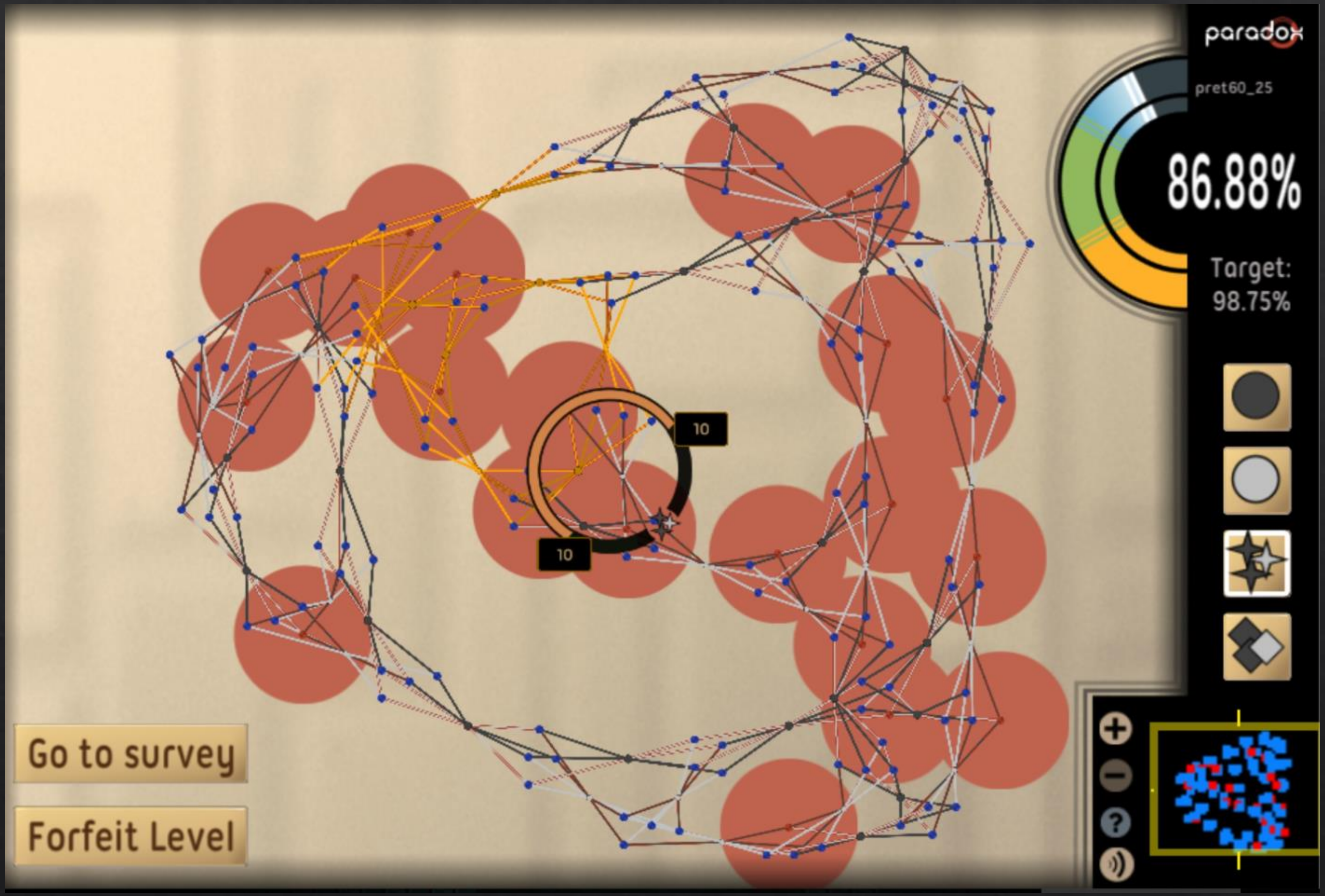
- *RQ2 – Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

Paradox

- 2D puzzle game for crowdsourced formal verification of software

- Each level represents a MAX-SAT problem

- Used same matchmaking system



Participant Recruitment and Study

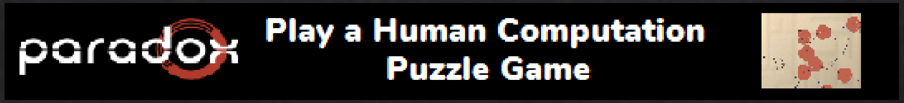
- Paid players recruited using Amazon Mechanical Turk

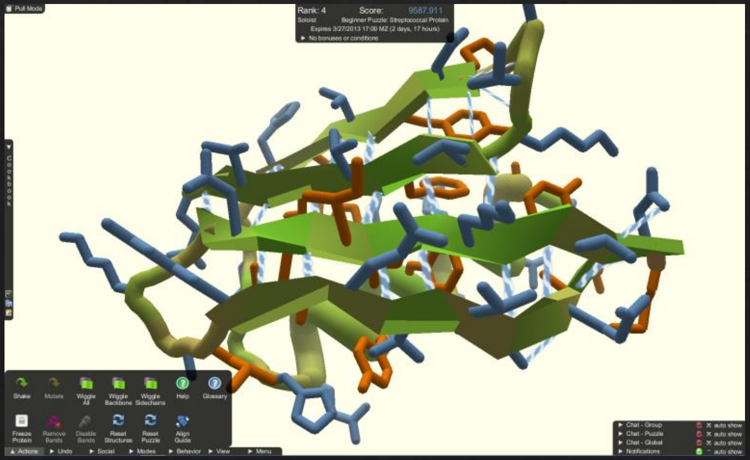
amazon mechanical turk™
Artificial Artificial Intelligence

Participant Recruitment and Study

- Paid players recruited using Amazon Mechanical Turk
- Volunteers recruited using banner ad on the website for the HCG Foldit (*http://fold.it)*

amazon mechanicalturk™
Artificial Artificial Intelligence




Rank: 4 Score: 9587.911
Soloist Beginner Puzzle: Streptococcal Protein
Expires 3/27/2013 17:00 MZ (2 days, 17 hours)
▶ No bonuses or conditions

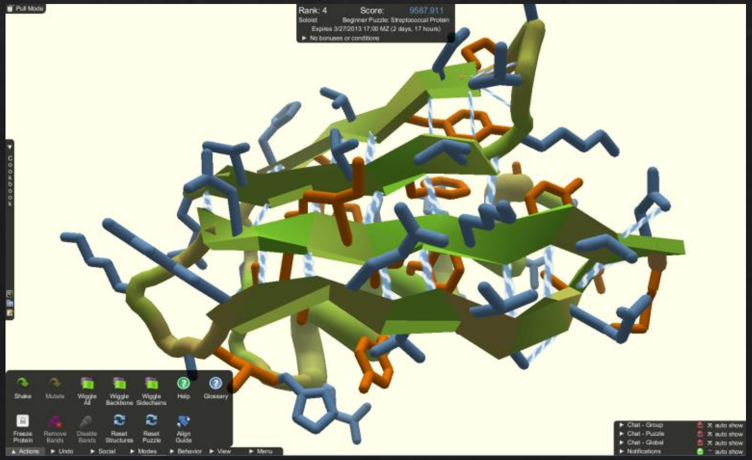
Participant Recruitment and Study

- Paid players recruited using Amazon Mechanical Turk
- Volunteers recruited using banner ad on the website for the HCG Foldit (*http://fold.it)*

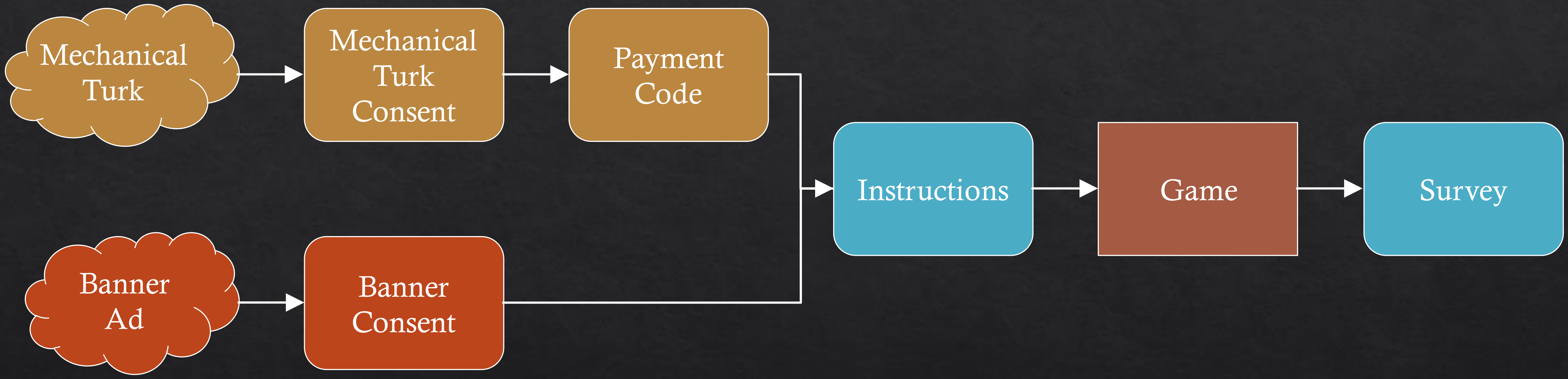
amazon mechanicalturk™
Artificial Artificial Intelligence

- Two experiments
  - RQ1: Effect of volunteer vs paid recruitment on engagement
  - RQ2: Effect of change in design on paid vs voluntary players

paradox Play a Human Computation Puzzle Game



Experiment Flow



Recruitment vs Participation

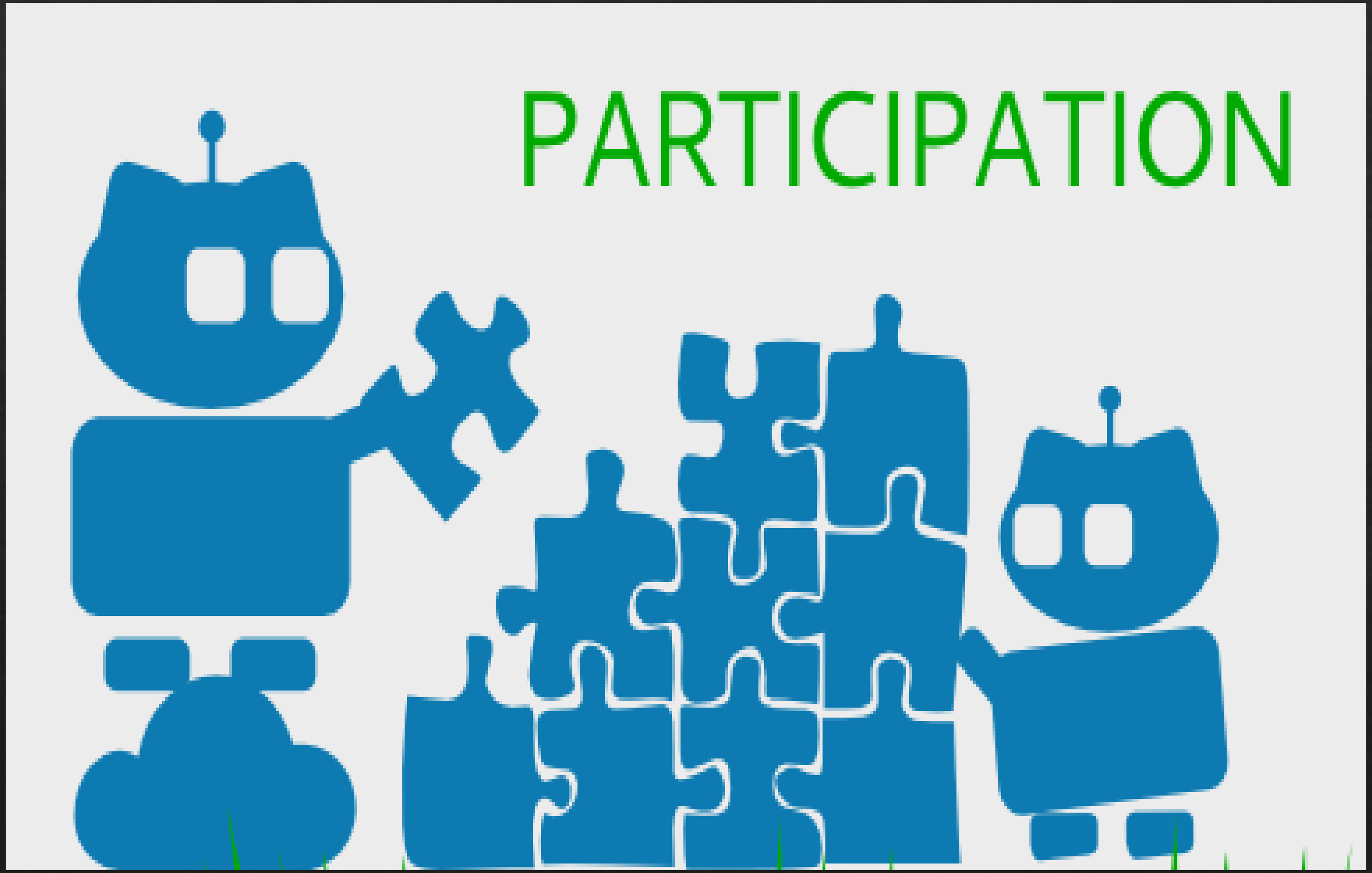
amazon mechanical turk™
Artificial Artificial Intelligence

Paid

paradox Play a Human Computation Puzzle Game

Voluntary

Recruitment vs Participation



amazon mechanical turk™
Artificial Artificial Intelligence

Voluntary

paradox Play a Human Computation Puzzle Game

Voluntary

Measures

- Behavioral Engagement
  - *Play Time*
  - *Levels Attempted*
  - *Levels Completed*
  - *Player Rating (Player's Glicko-2 rating after completing the game)*
  - *Highest Level Rating (Highest Glicko-2 rating of any level completed by the player)*

Measures

- Behavioral Engagement
  - *Play Time*
  - *Levels Attempted*
  - *Levels Completed*
  - *Player Rating (Player's Glicko-2 rating after completing the game)*
  - *Highest Level Rating (Highest Glicko-2 rating of any level completed by the player)*

- Intrinsic Motivation Inventory
  - *Interest/Enjoyment*
  - *Perceived Competence*
  - *Perceived Choice*
  - *Effort/Importance*

Experiment One: Recruitment Strategy

- *Does recruitment strategy impact participant behavior and experience in HCGs?*

Experiment One: Recruitment Strategy

- *Does recruitment strategy impact participant behavior and experience in HCGs?*

- Three conditions
  - BANNER

Experiment One: Recruitment Strategy

- *Does recruitment strategy impact participant behavior and experience in HCGs?*

- Three conditions
  - BANNER
  - MTURK-SM ($0.10)

Experiment One: Recruitment Strategy

- *Does recruitment strategy impact participant behavior and experience in HCGs?*

- Three conditions
  - BANNER
  - MTURK-SM ($0.10)
  - MTURK-LG ($1.00)

Experiment One: Recruitment Strategy

- *Does recruitment strategy impact participant behavior and experience in HCGs?*

- Three conditions
  - BANNER
  - MTURK-SM ($0.10)
  - MTURK-LG ($1.00)

- 177 players recruited through the banner
- 225 players recruited through each MTurk condition
  - 162 (72%) played in MTURK-SM after being paid
  - 194 (86%) played in MTURK-LG after being paid

Experiment One: Recruitment Strategy

- *Does recruitment strategy impact participant behavior and experience in HCGs?*

- Three conditions
  - BANNER
  - MTURK-SM ($0.10)
  - MTURK-LG ($1.00)

- 177 players recruited through the banner
- 225 players recruited through each MTurk condition
  - 162 (72%) played in MTURK-SM after being paid
  - 194 (86%) played in MTURK-LG after being paid

> *"A requester on mturk giving away free money… that truly is a paradox!"*

Experiment One Results

Variable	BANNER	MTURK-SM	MTURK-LG
Player Rating	**1808**	1509	1636
Highest Level Rating	**1625**	1222	1367
Levels Attempted	3	3	**4**
Levels Completed	3	3	**4**

Statistical Tests: Omnibus Kruskal-Wallis Test, post-hoc Wilcoxon Rank-Sum Test

- No significant differences across conditions for *Play Time*

Experiment One Results

Variable	BANNER	MTURK-SM	MTURK-LG
Effort/Importance	46%	63%	**74%**
Interest/Enjoyment	53%	56%	**65%**
Perceived Competence	43%	48%	**60%**

Statistical Tests: Omnibus Kruskal-Wallis Test, post-hoc Wilcoxon Rank-Sum Test

- No significant differences across conditions for *Perceived Choice*

Experiment One Results

Variable	BANNER	MTURK-SM	MTURK-LG
Effort/Importance	46%	63%	**74%**
Interest/Enjoyment	53%	56%	**65%**
Perceived Competence	43%	48%	**60%**

Statistical Tests: Omnibus Kruskal-Wallis Test, post-hoc Wilcoxon Rank-Sum Test

- No significant differences across conditions for *Perceived Choice*
- Only 6% of BANNER completed the survey compared to 70% of MTURK-SM and 82% of MTURK-LG

Experiment One Discussion

- If goal is to maximize *task volume*, then paid recruitment may be preferred

Experiment One Discussion

- If goal is to maximize *task volume*, then paid recruitment may be preferred

- If goal is to maximize *task quality*, then volunteer recruitment may be preferred

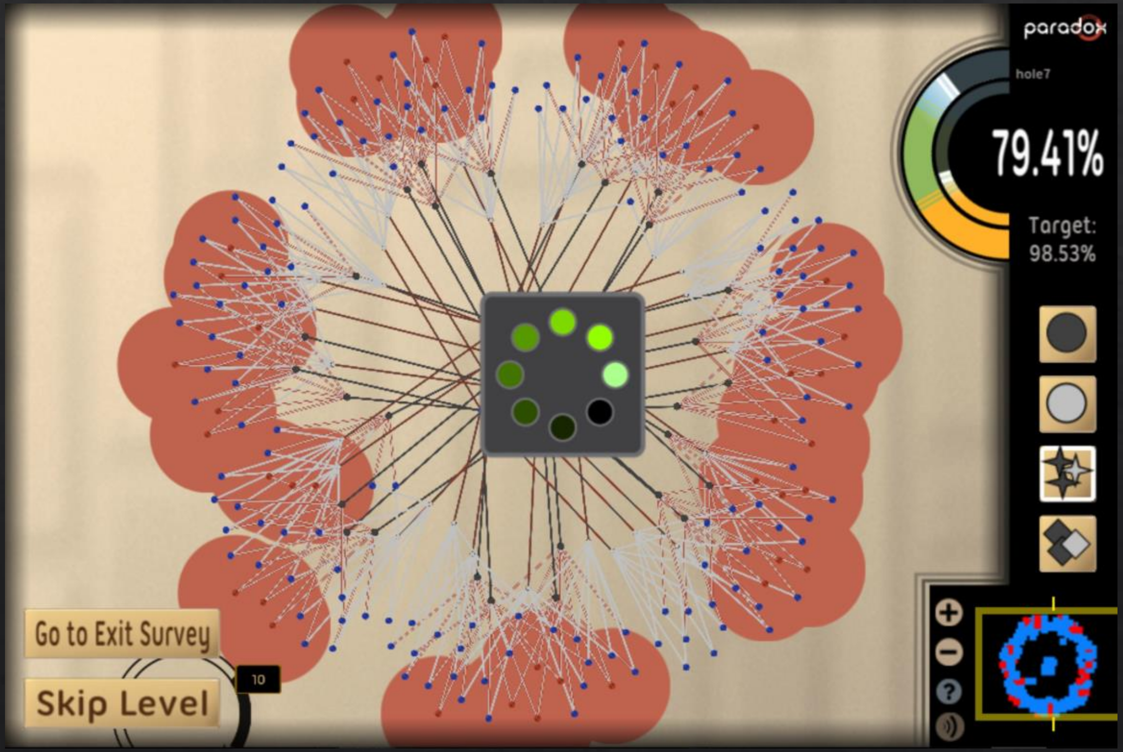
Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

- Added an artificial loading delay of 20-seconds between levels (*Card et al., 1991; Miller, 1968; Sharek and Wiebe, 2014*)

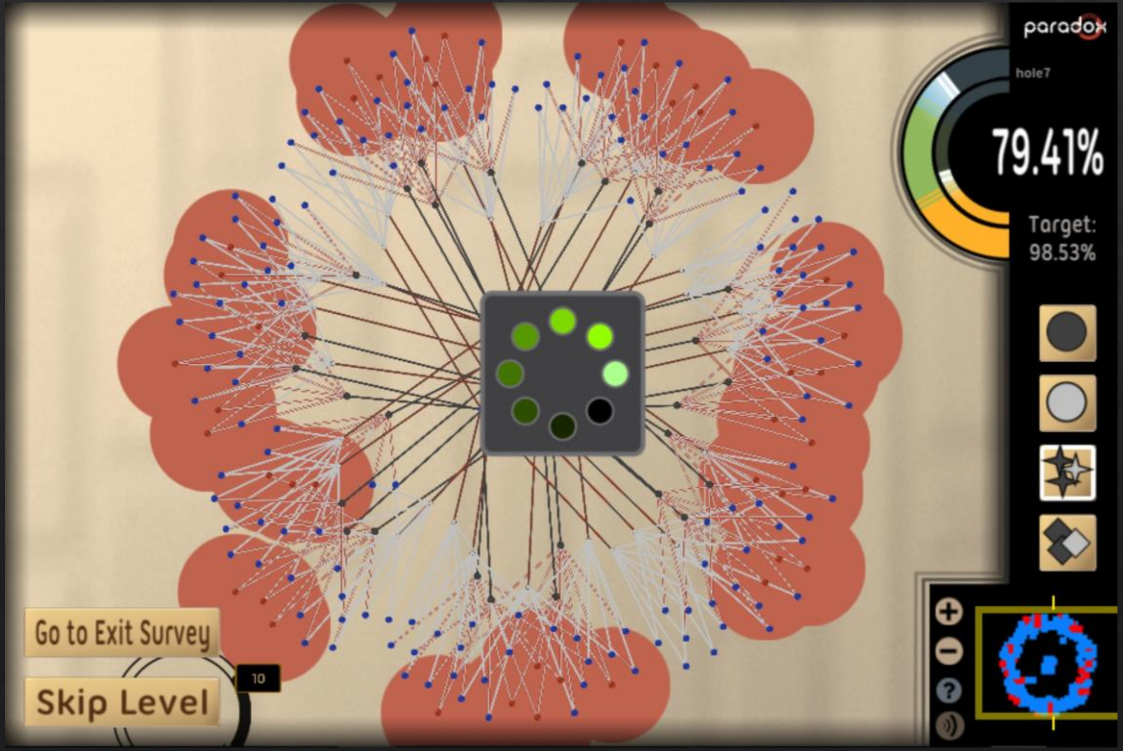


Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

- Added an artificial loading delay of 20-seconds between levels

- 2x2 between-subjects design with four conditions
  - RECRUITMENT
    - BANNER

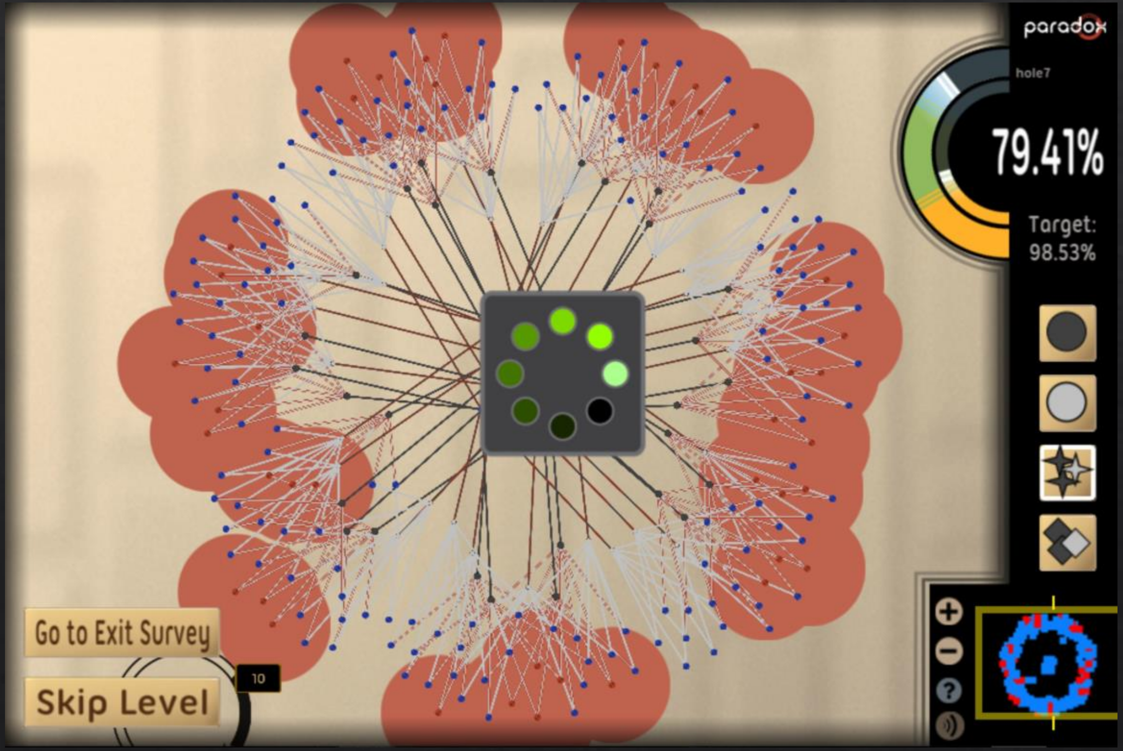


Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

- Added an artificial loading delay of 20-seconds between levels

- 2x2 between-subjects design with four conditions
  - RECRUITMENT
    - BANNER
    - MTURK-LG

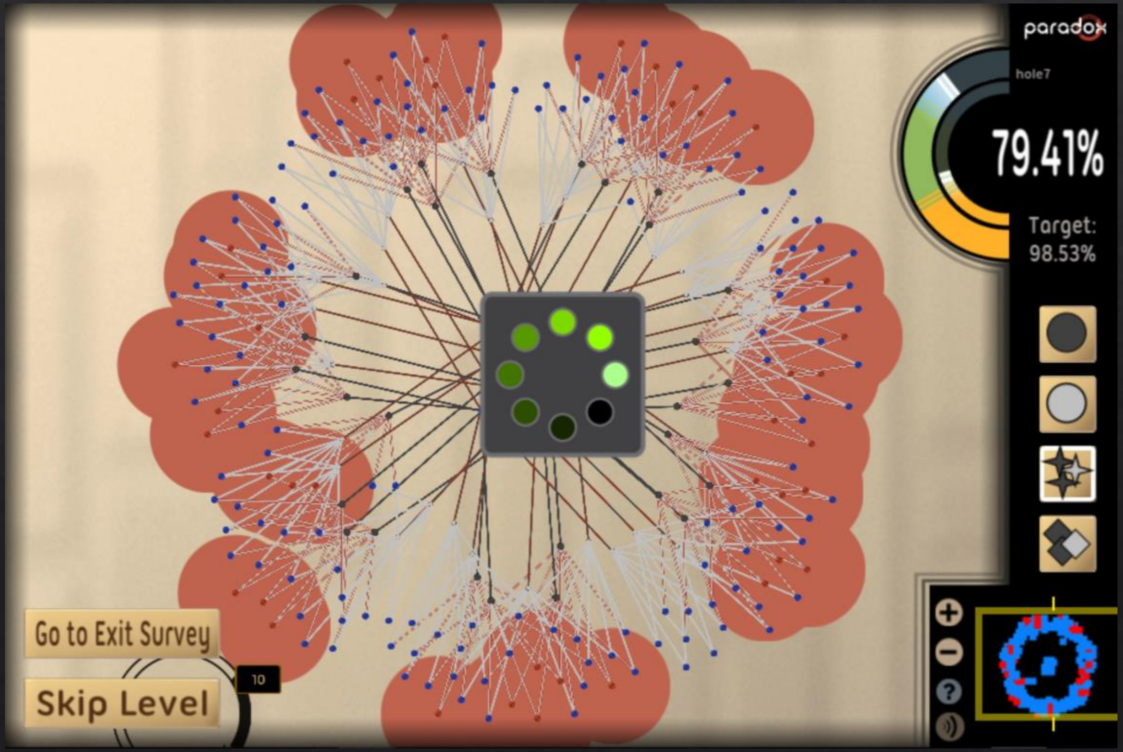


Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

- Added an artificial loading delay of 20-seconds between levels

- 2x2 between-subjects design with four conditions
  - RECRUITMENT
    - BANNER
    - MTURK-LG
  - DESIGN
    - DELAY

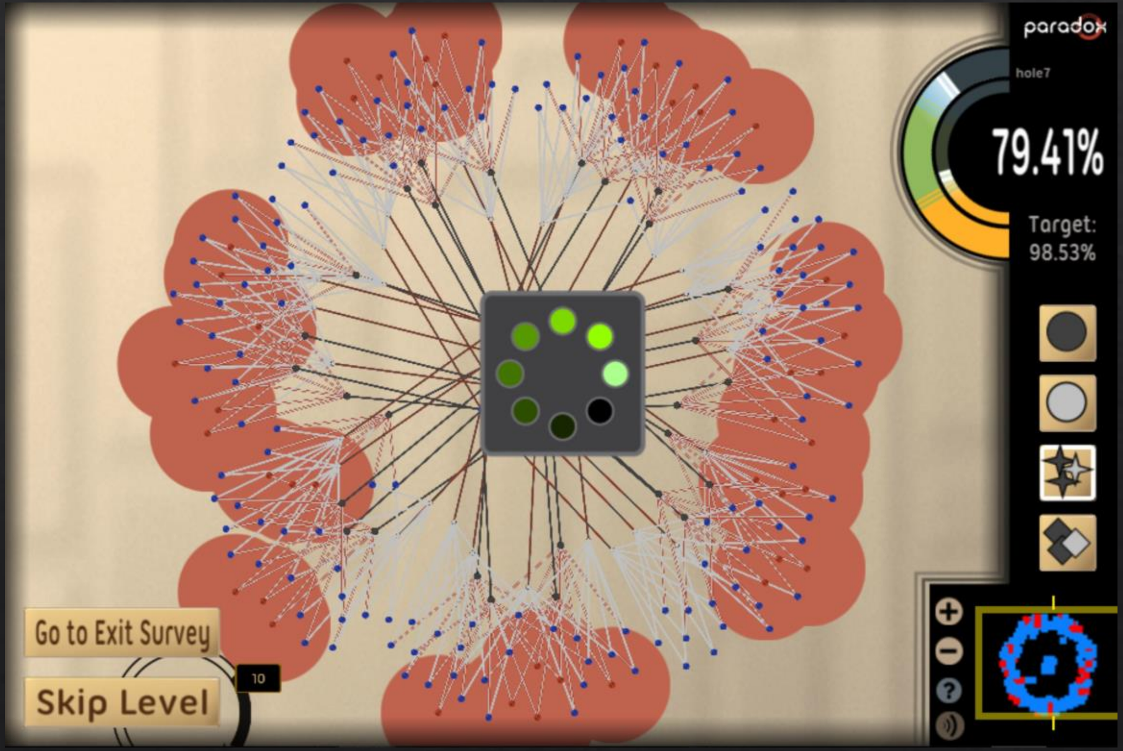


Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

- Added an artificial loading delay of 20-seconds between levels

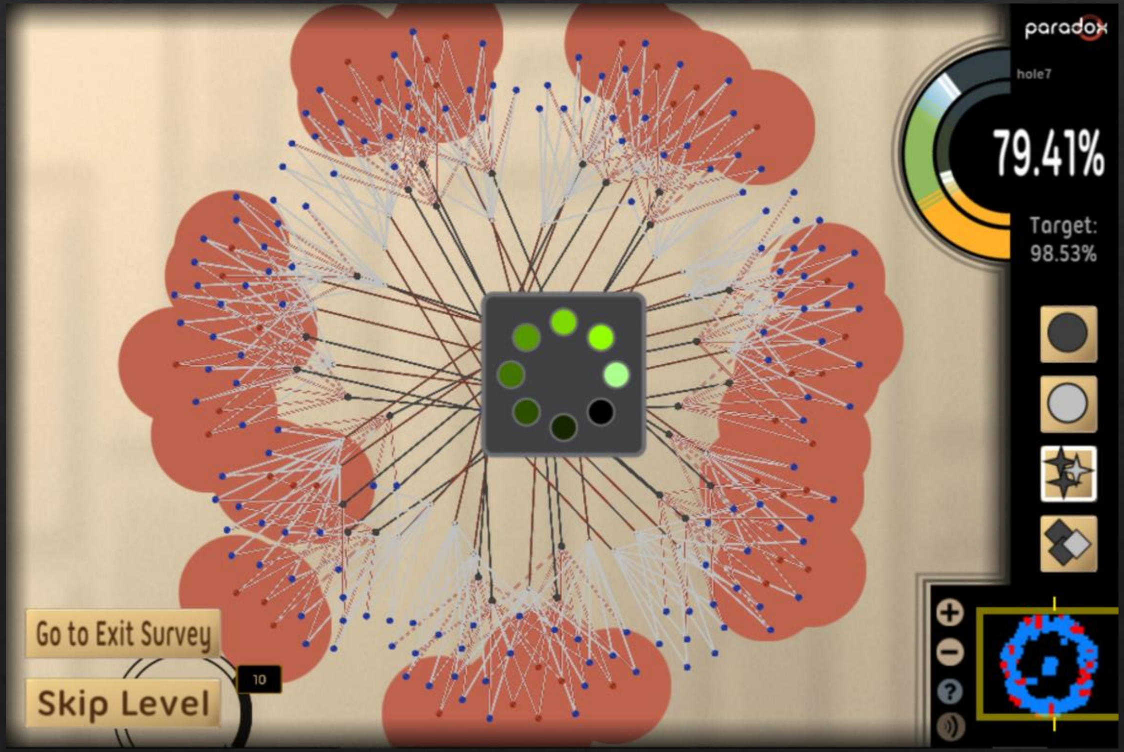
- 2x2 between-subjects design with four conditions
  - RECRUITMENT
    - BANNER
    - MTURK-LG
  - DESIGN
    - DELAY
    - NO-DELAY



Experiment Two: Recruitment Strategy vs Delay

- *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

- Added an artificial loading delay of 20-seconds between levels
- 2x2 between-subjects design with four conditions
  - RECRUITMENT
    - BANNER
    - MTURK-LG
  - DESIGN
    - DELAY
    - NO-DELAY
- 260 players were recruited through the banner
- 300 players were recruited through MTurk with 244 (81.3%) proceeding to play the game



Experiment Two Results

Variable	BANNER	MTURK-LG	DELAY	NO-DELAY
Play Time	119s	**206.5s**	129s	**162s**
Levels Attempted	3	**4**	2	**4**
Levels Completed	3	**4**	2	**4**
Player Rating	**1657**	1627	1636	1646
Effort/Importance	57%	**71%**	66%	71%

Statistical Test: Aligned Rank Transform (ART)

- No interaction effects for any response variable
- No significant differences across conditions for *Highest Level Rating, Interest/Enjoyment, Perceived Competence* and *Perceived Choice*

Experiment Two Results

Variable	**BANNER**	**MTURK-LG**	**DELAY**	**NO-DELAY**
Play Time	119s	**206.5s**	129s	**162s**
Levels Attempted	3	**4**	2	**4**
Levels Completed	3	**4**	2	**4**
Player Rating	**1657**	1627	1636	1646
Effort/Importance	57%	**71%**	66%	71%

Statistical Test: Aligned Rank Transform (ART)

- Main effect of recruitment and delay

Experiment Two Results

Variable	BANNER	MTURK-LG	DELAY	NO-DELAY
Play Time	119s	**206.5s**	129s	**162s**
Levels Attempted	3	**4**	2	**4**
Levels Completed	3	**4**	2	**4**
Player Rating	**1657**	1627	1636	1646
Effort/Importance	57%	**71%**	66%	71%

Statistical Test: Aligned Rank Transform (ART)

- Main effect of only recruitment

Experiment Two Discussion

- Different recruitment strategies did not observably impact the effects of changing the game’s design

Experiment Two Discussion

- Different recruitment strategies did not observably impact the effects of changing the game's design

- As in experiment one, a measure of *task volume* was higher for paid recruitment and a measure of *task quality* was higher for volunteer recruitment

Conclusion

- Paid player recruitment results in a higher *volume* of tasks being completed

Conclusion

- Paid player recruitment results in a higher *volume* of tasks being completed

- Volunteer player recruitment results in a higher *quality* of completed tasks

Conclusion

- Paid player recruitment results in a higher *volume* of tasks being completed
- Volunteer player recruitment results in a higher *quality* of completed tasks
- Effects of recruitment strategies remain consistent with changes to the game’s design

Future Work

- Interaction effects of other changes to the game’s design

Future Work

- Interaction effects of other changes to the game's design

- Alternate methods of gathering self-reported experience metrics from more players without compromising voluntary nature of participation

Contact

Anurag Sarkar
Northeastern University
sarkar.an@husky.neu.edu

Acknowledgments

This material is based upon work supported by the **National Science Foundation** under grant no. 1652537. We would like to thank the **University of Washington**'s **Center for Game Science** for initial *Paradox* development.