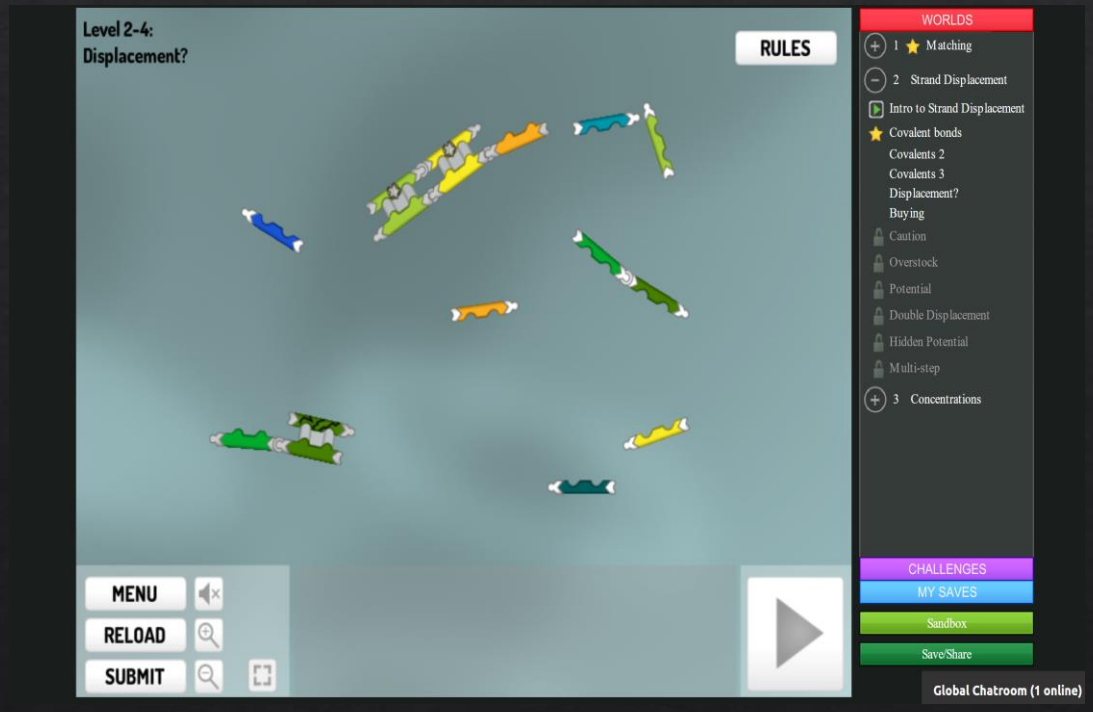
Engagement Effects of Player Rating System-Based Matchmaking for Level Ordering in Human Computation Games

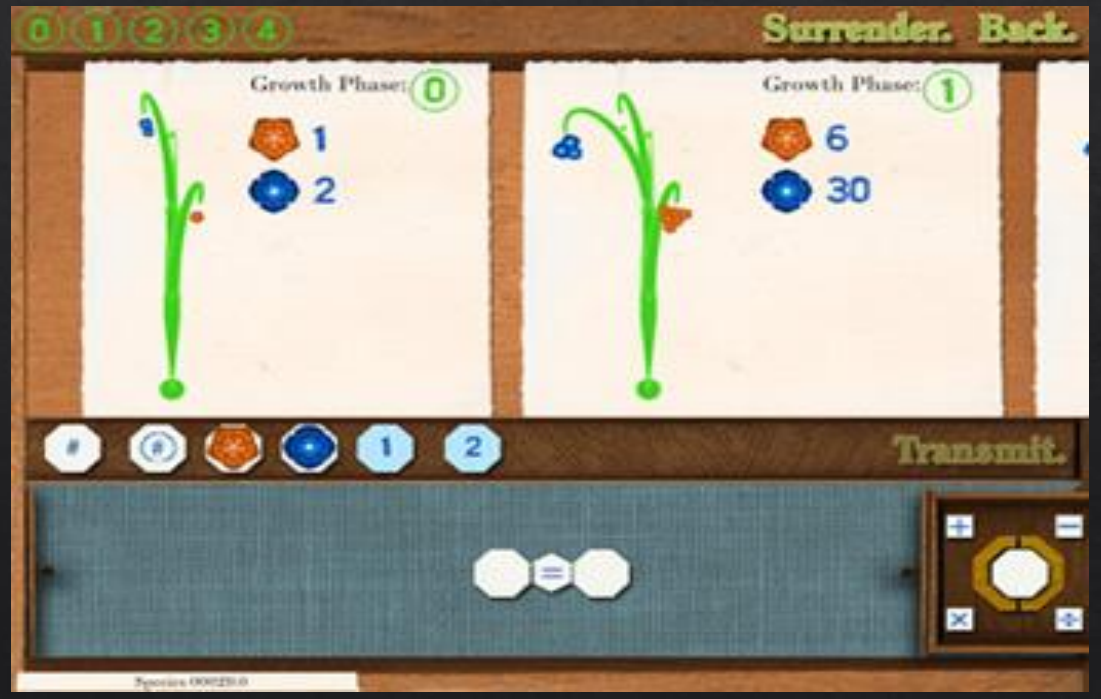
Anurag Sarkar, Michael Williams, Sebastian Deterding, Seth Cooper

Human Computation Games (HCGs)

Games that motivate large numbers of people to solve tasks that are hard to automate



Nanocrafter



Xylem

Engagement

- PROBLEMS
  - Poor engagement
  - Poor retention

Engagement

- PROBLEMS
  - Poor engagement
  - Poor retention

- ENGAGEMENT
  - Degree and quality of a person's involvement in a task
  - Theory of Flow
    - *Flow State* –when one is motivated and deeply engrossed in an activity
  - Games engage players by having challenges be balanced relative to player skill

Difficulty Balancing

- REASON - Lack of difficulty balancing in HCGs

Difficulty Balancing

- REASON - Lack of difficulty balancing in HCGs
  - No *a priori* knowledge of difficulty of tasks to be solved

Difficulty Balancing

- REASON - Lack of difficulty balancing in HCGs
  - No *a priori* knowledge of difficulty of tasks to be solved
  - Not possible to modify tasks without compromising validity of solutions

Difficulty Balancing

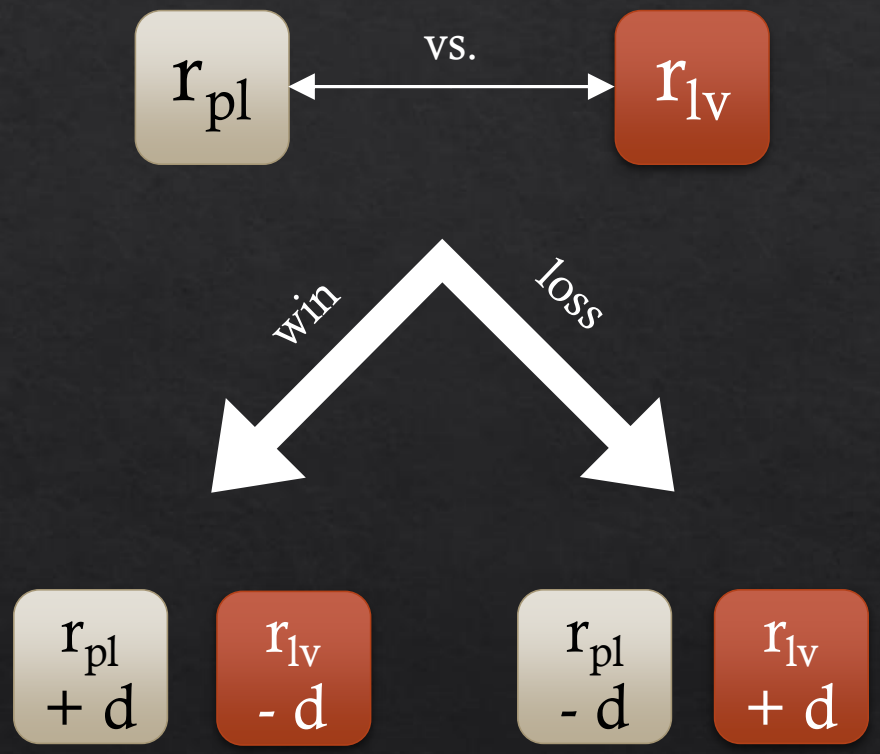
- REASON - Lack of difficulty balancing in HCGs
  - No *a priori* knowledge of difficulty of tasks to be solved
  - Not possible to modify tasks without compromising validity of solutions

- POSSIBLE SOLUTION – Player Rating Systems

Player Rating Systems

Examples: Elo, Glicko, TrueSkill

Player Rating Systems



Examples: Elo, Glicko, TrueSkill

Research Questions/Hypotheses

- RQ1 – How does difficulty balancing affect engagement in HCGs?
- RQ2 – How does rating-based matchmaking affect engagement in HCGs?

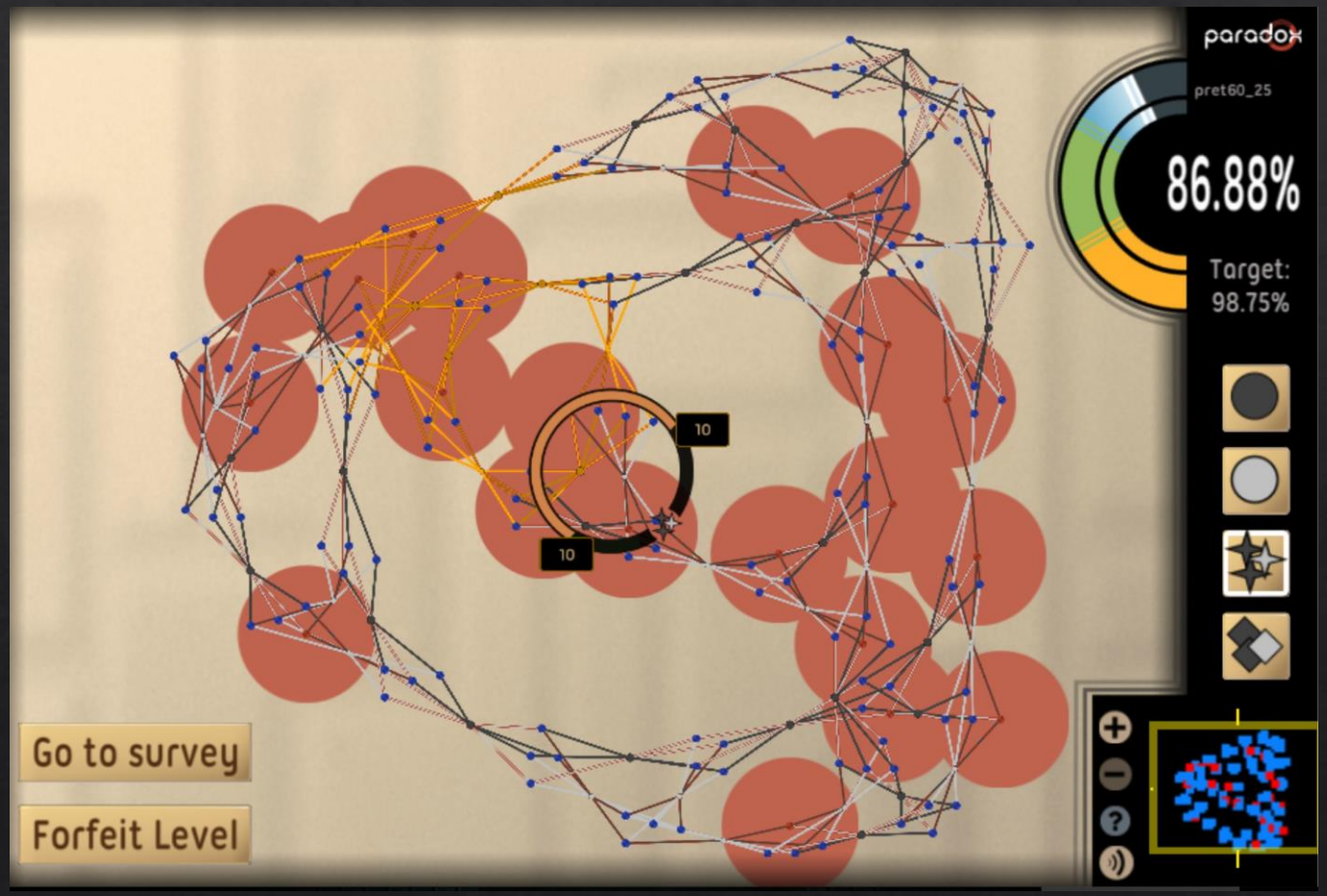
Research Questions/Hypotheses

- RQ1 – How does difficulty balancing affect engagement in HCGs?
- RQ2 – How does rating-based matchmaking affect engagement in HCGs?

- H1 – Serving levels in strictly increasing order of difficulty leads to higher engagement than serving levels randomly
- H2 – Serving levels in order defined by matchmaking system leads to highest engagement

Paradox

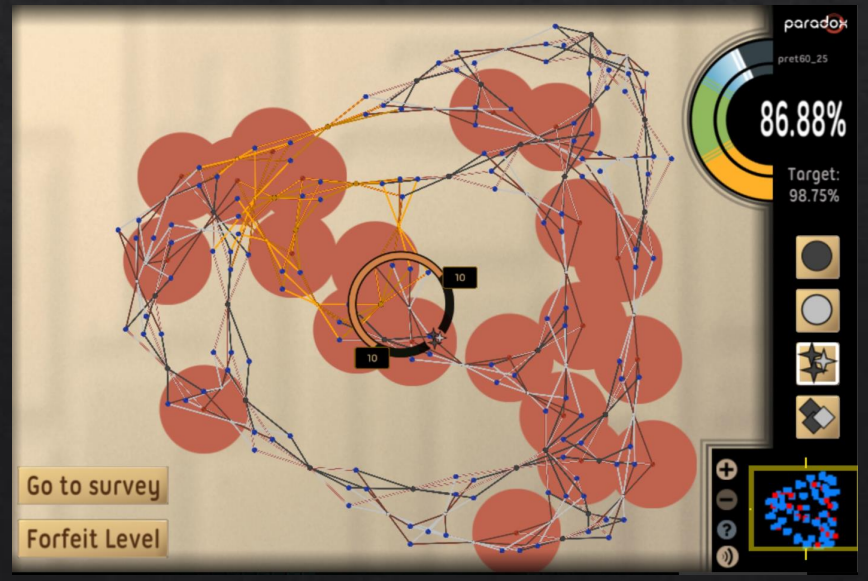
- 2D puzzle game for crowdsourced formal verification of software
- Each level represents a MAX-SAT problem
- Players assign values to variables, schedule optimizations
- Player completes level by reaching target score



Participant Recruitment and Study

- Players recruited using Amazon Mechanical Turk

- Two phase study
  - Initial Level Rating Generation
  - Matchmaking using generated level ratings



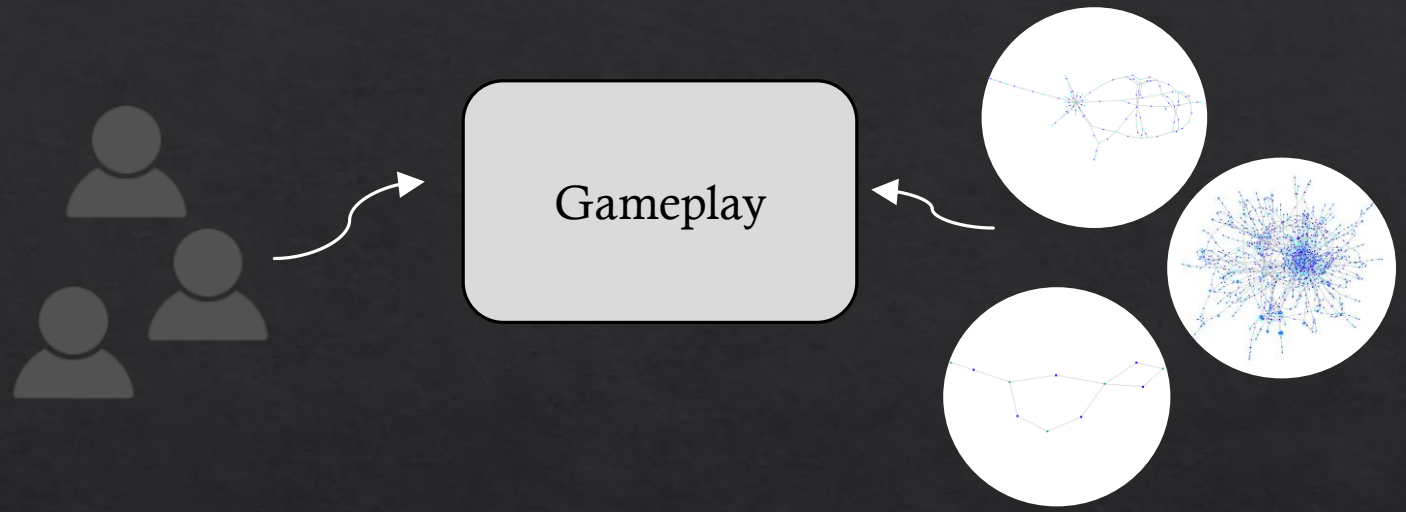
- Glicko-2 Rating System

- 9 tutorial levels, 33 challenge levels

Phase 1: Initial Level Rating Generation

- 98 players
- Player-level pairings considered as matches
- Match outcomes:
  - Level Completed => Player wins
  - Level Forfeited => Level wins
  - Level Skipped => Ignore
- Default Glicko-2 Parameter Values

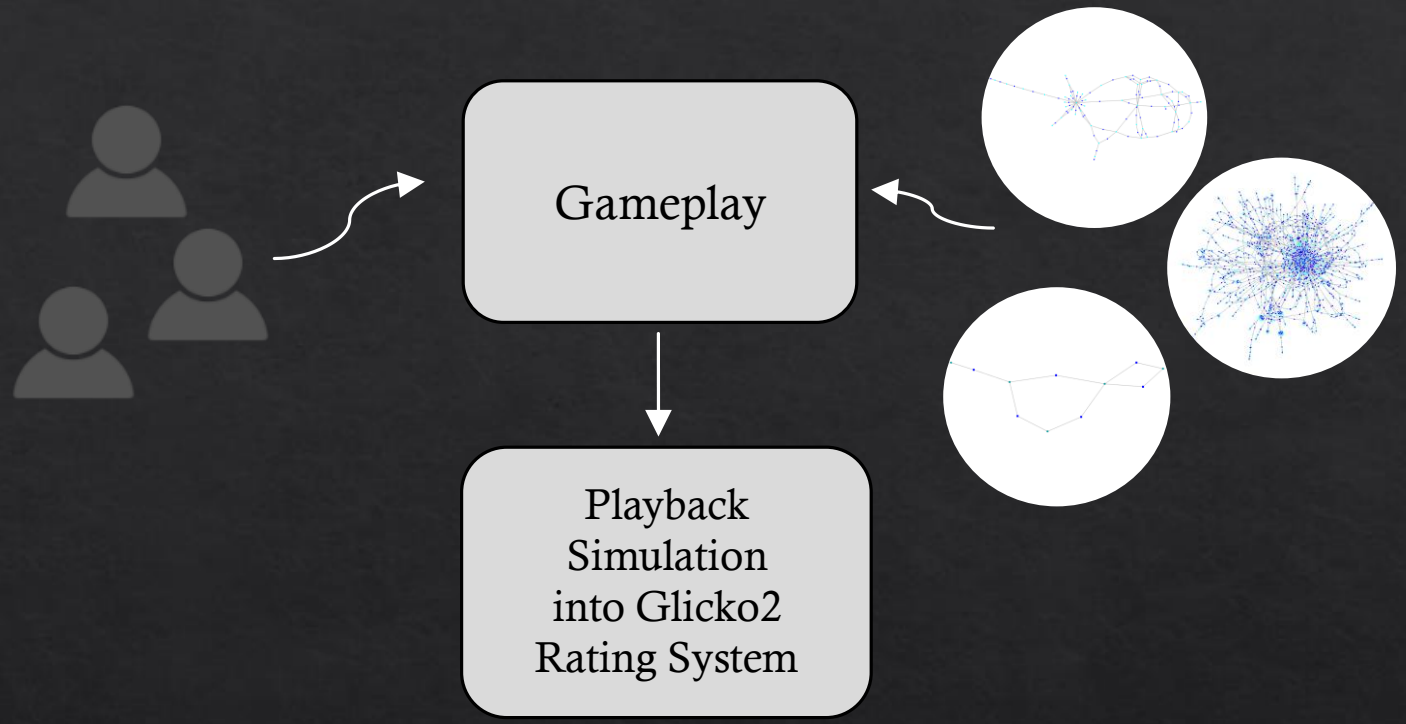
  (Rating – 1500, Deviation – 350, Volatility – 0.06)



Phase 1: Initial Level Rating Generation

- 98 players
- Player-level pairings considered as matches
- Match outcomes:
  - Level Completed => Player wins
  - Level Forfeited => Level wins
  - Level Skipped => Ignore
- Default Glicko-2 Parameter Values

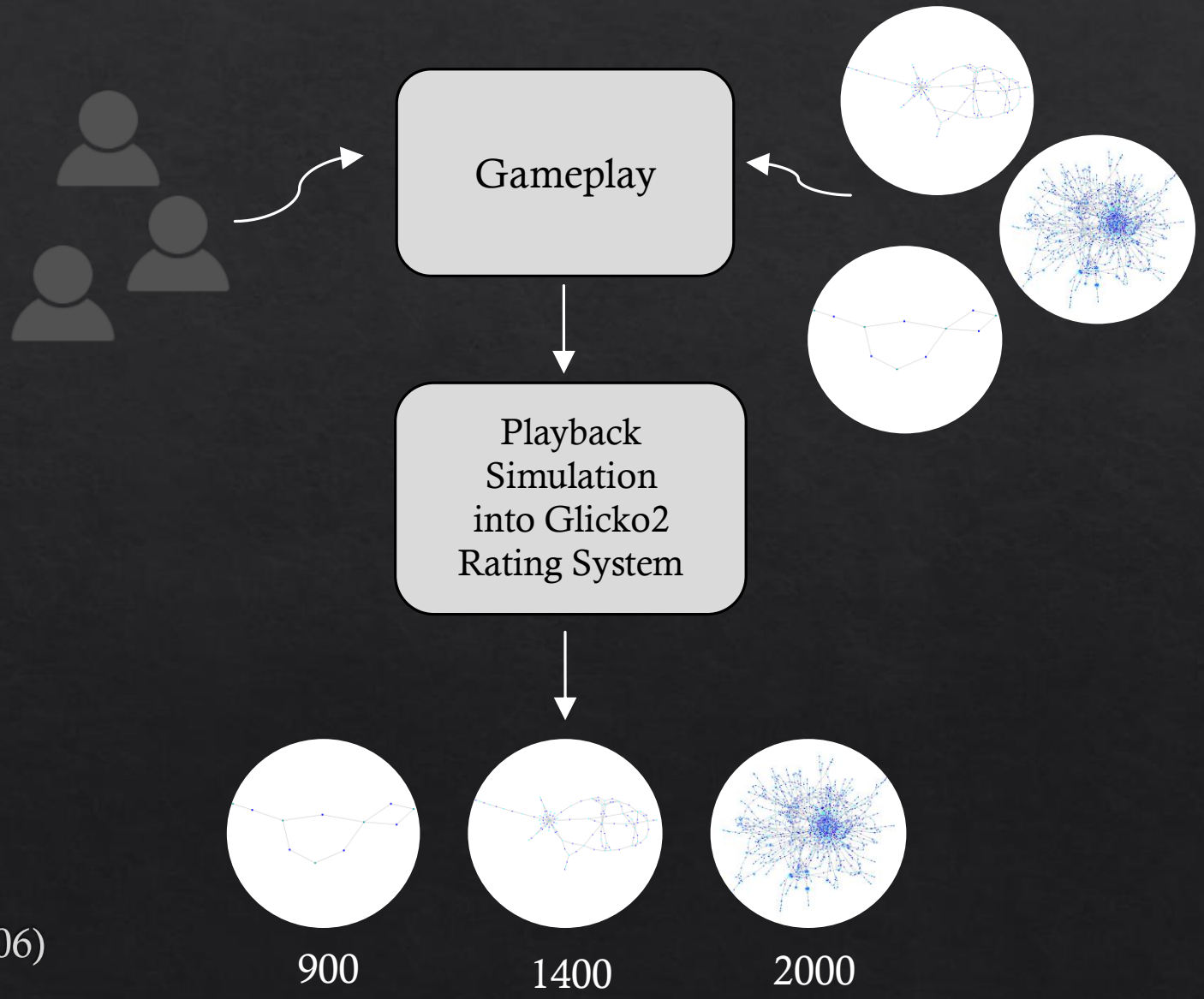
  (Rating – 1500, Deviation – 350, Volatility – 0.06)



Phase 1: Initial Level Rating Generation

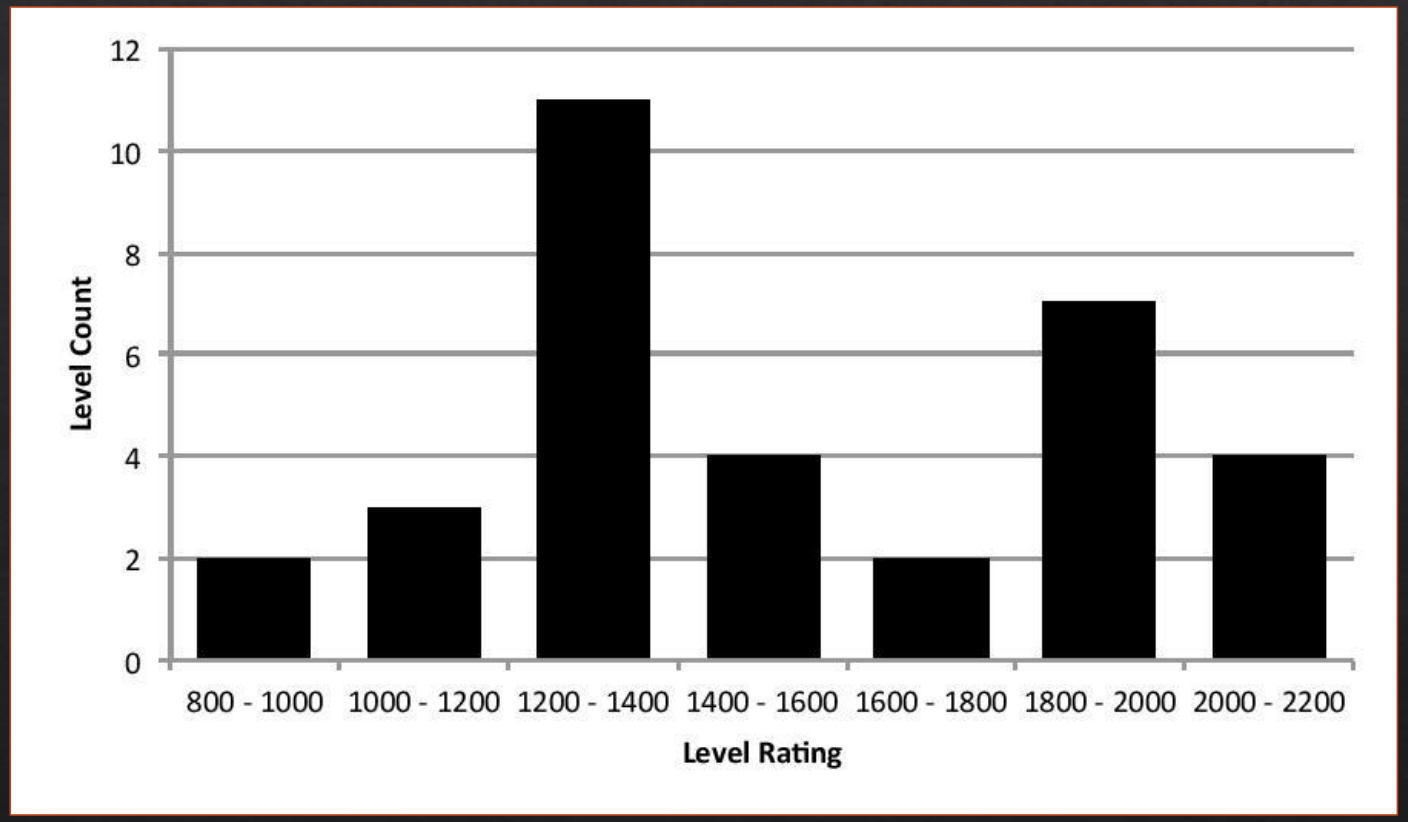
- 98 players
- Player-level pairings considered as matches
- Match outcomes:
  - Level Completed => Player wins
  - Level Forfeited => Level wins
  - Level Skipped => Ignore
- Default Glicko-2 Parameter Values

  (Rating – 1500, Deviation – 350, Volatility – 0.06)

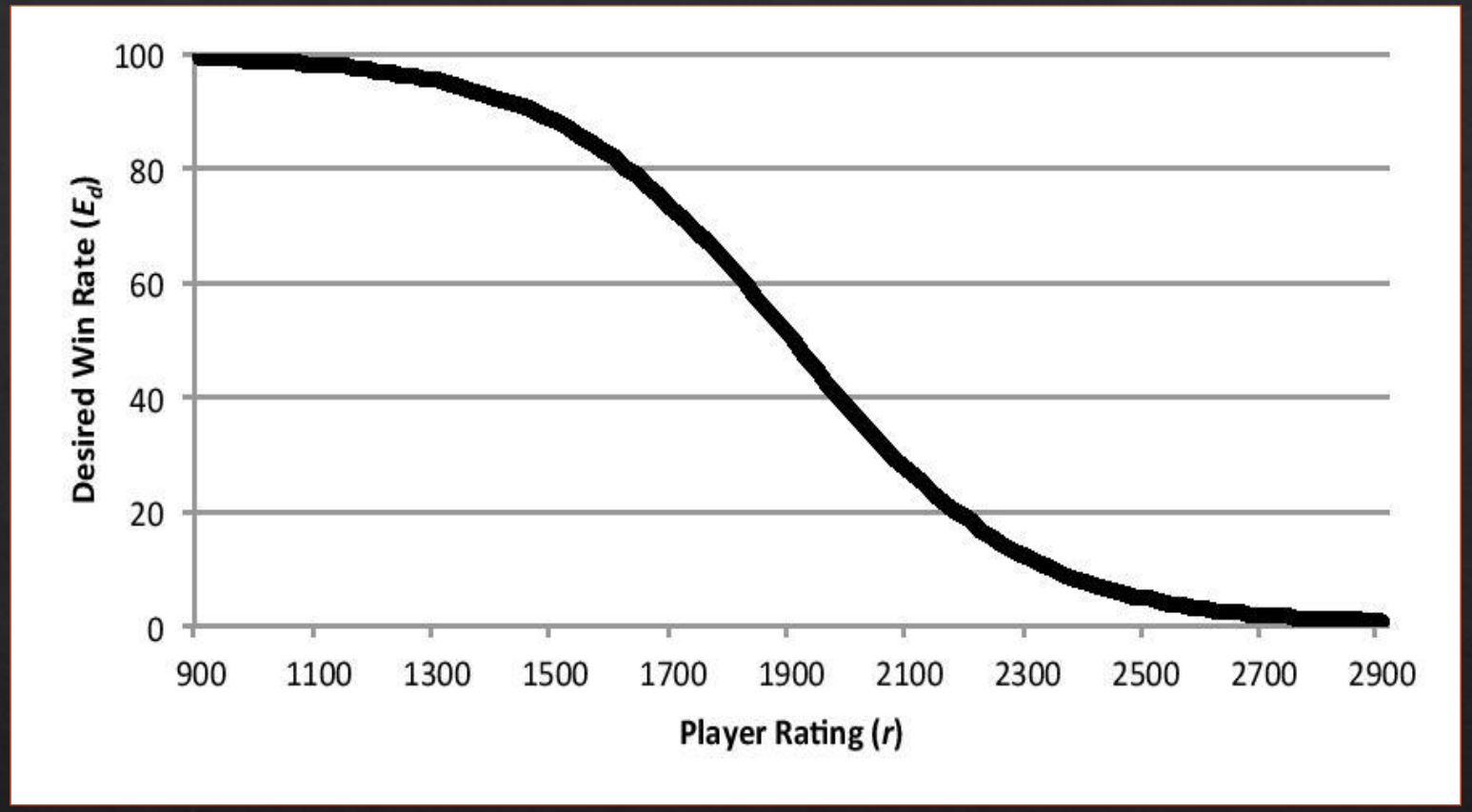


Phase 1: Initial Level Rating Generation

- 98 players
- Player-level pairings considered as matches
- Match outcomes:
  - Level Completed => Player wins
  - Level Forfeited => Level wins
  - Level Skipped => Ignore

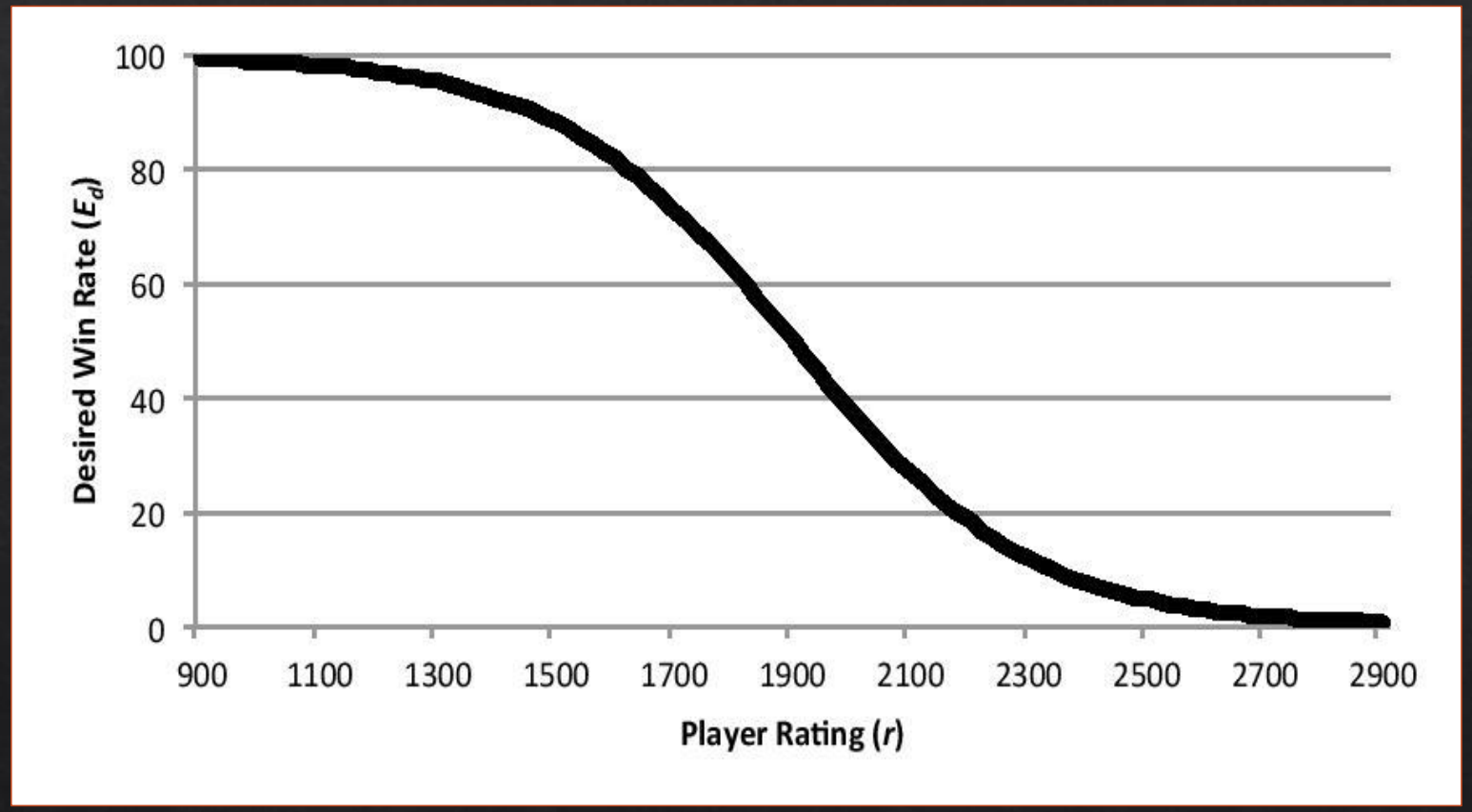


Phase 2: Ordering Experiment



- Low rating → High Desired Win Rate → Easy levels served
- High rating → Low Desired Win Rate → Hard levels served

Phase 2: Ordering Experiment

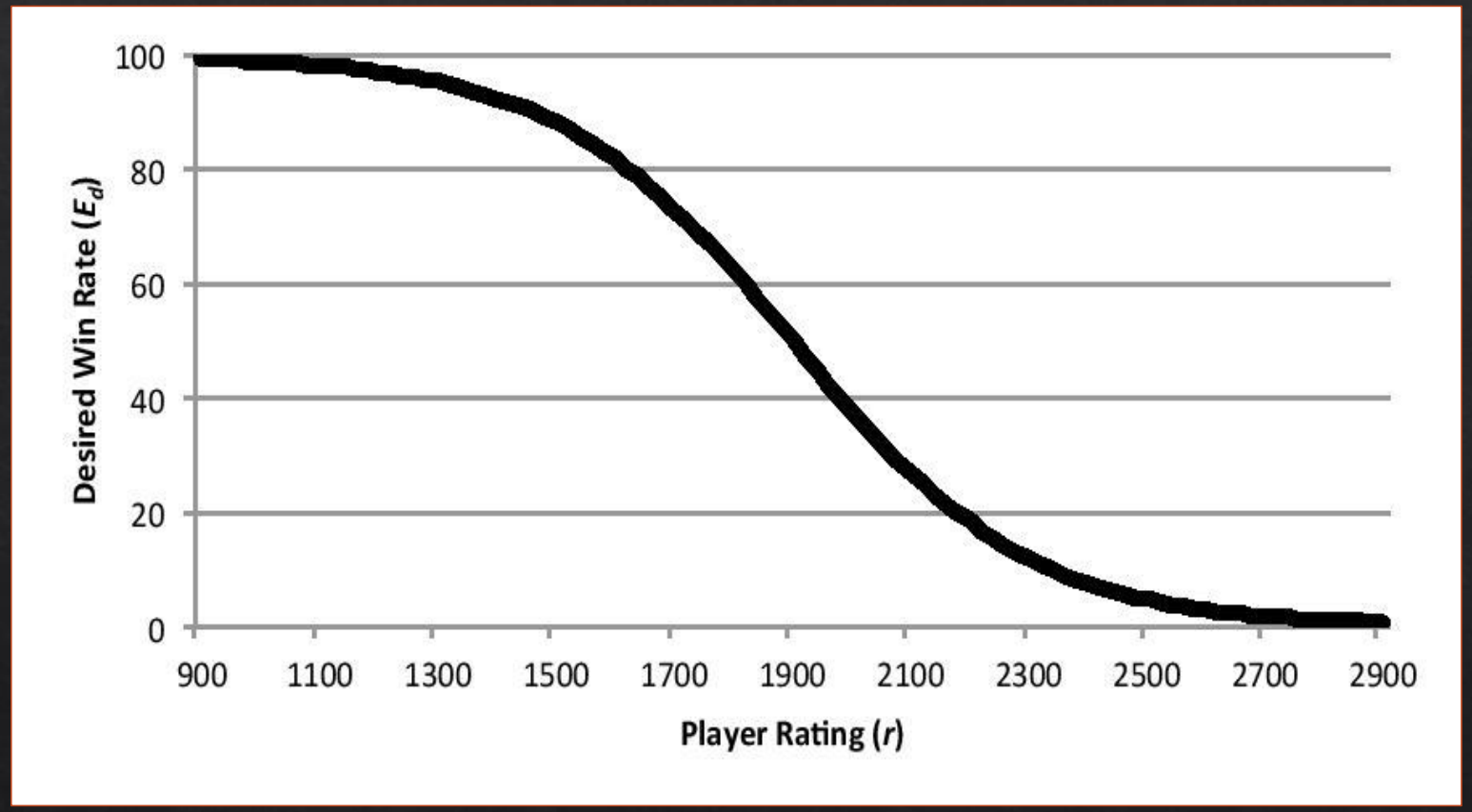


- Desired Win Rate:

  $E_d(r) = 1 - 1/(1 + e^{-k(r-r0)})$

  $k = 0.005$, $r0 = 1900$

Phase 2: Ordering Experiment



- Desired Win Rate:

  $E_d(r) = 1 - 1/(1 + e^{-k(r-r0)})$

  $k = 0.005$, $r0 = 1900$

- Win Expectancy Formula:

  $E_p(r, v) = 1/(1 + 10^{(v-r)/400})$

  r – player's current rating

  v – level rating

Phase 2: Ordering Experiment

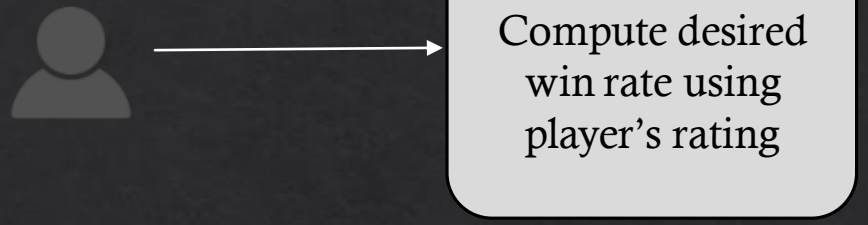
- 393 workers accepted HIT
- 294 completed HIT (75% completion rate)

- Ordering:
  - MATCHMAKING – 79
  - INCREASING – 99
  - RANDOM – 116

- Levels and players initialized with default Glicko2 parameters except levels were initialized with ratings from phase 1

Phase 2: Ordering Experiment

- 393 workers accepted HIT
- 294 completed HIT (75% completion rate)



- Ordering:
  - MATCHMAKING – 79
  - INCREASING – 99
  - RANDOM – 116

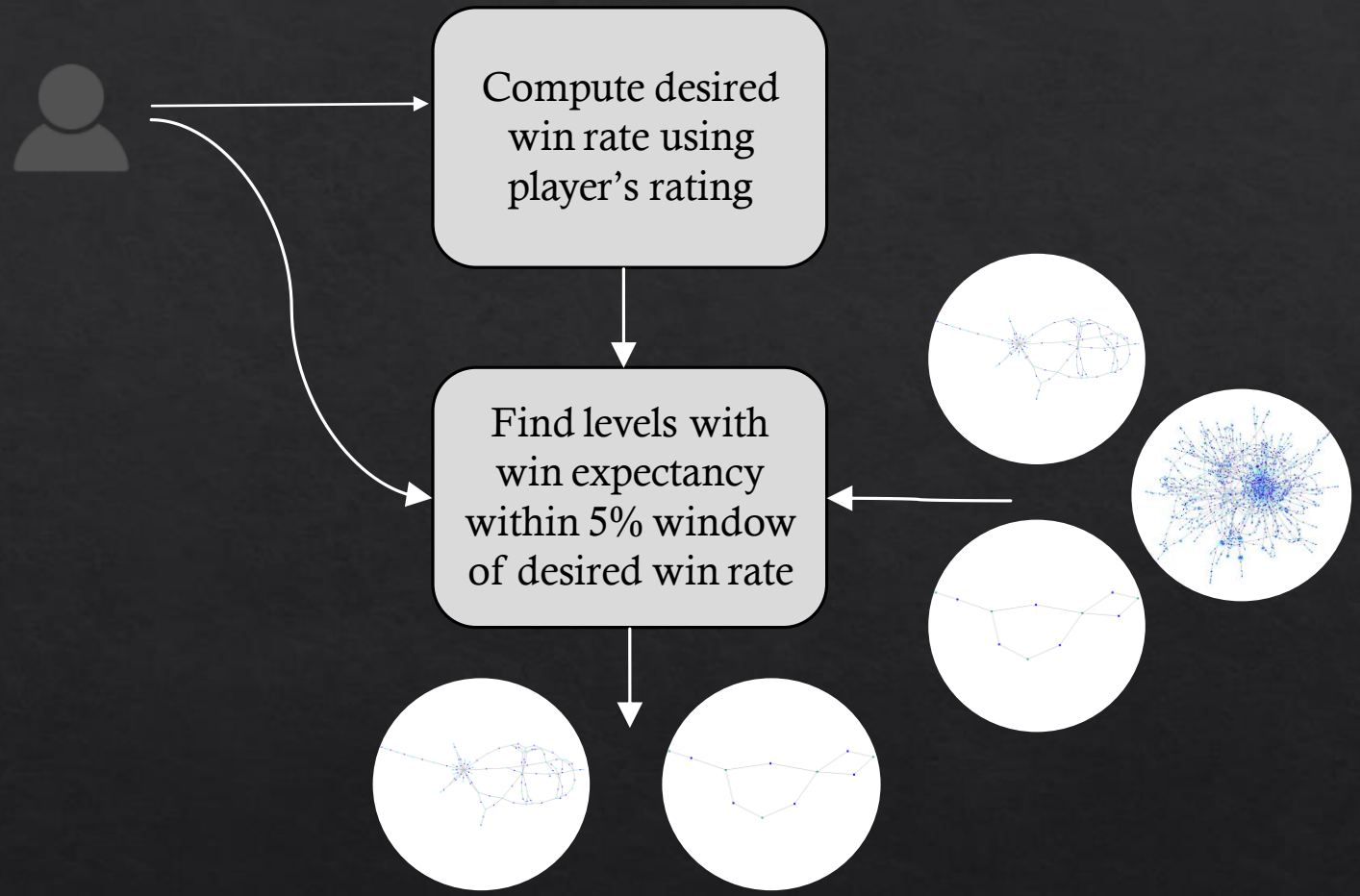
- Levels and players initialized with default Glicko2 parameters except levels were initialized with ratings from phase 1

Phase 2: Ordering Experiment

- 393 workers accepted HIT
- 294 completed HIT (75% completion rate)

- Ordering:
  - MATCHMAKING – 79
  - INCREASING – 99
  - RANDOM – 116

- Levels and players initialized with default Glicko2 parameters except levels were initialized with ratings from phase 1

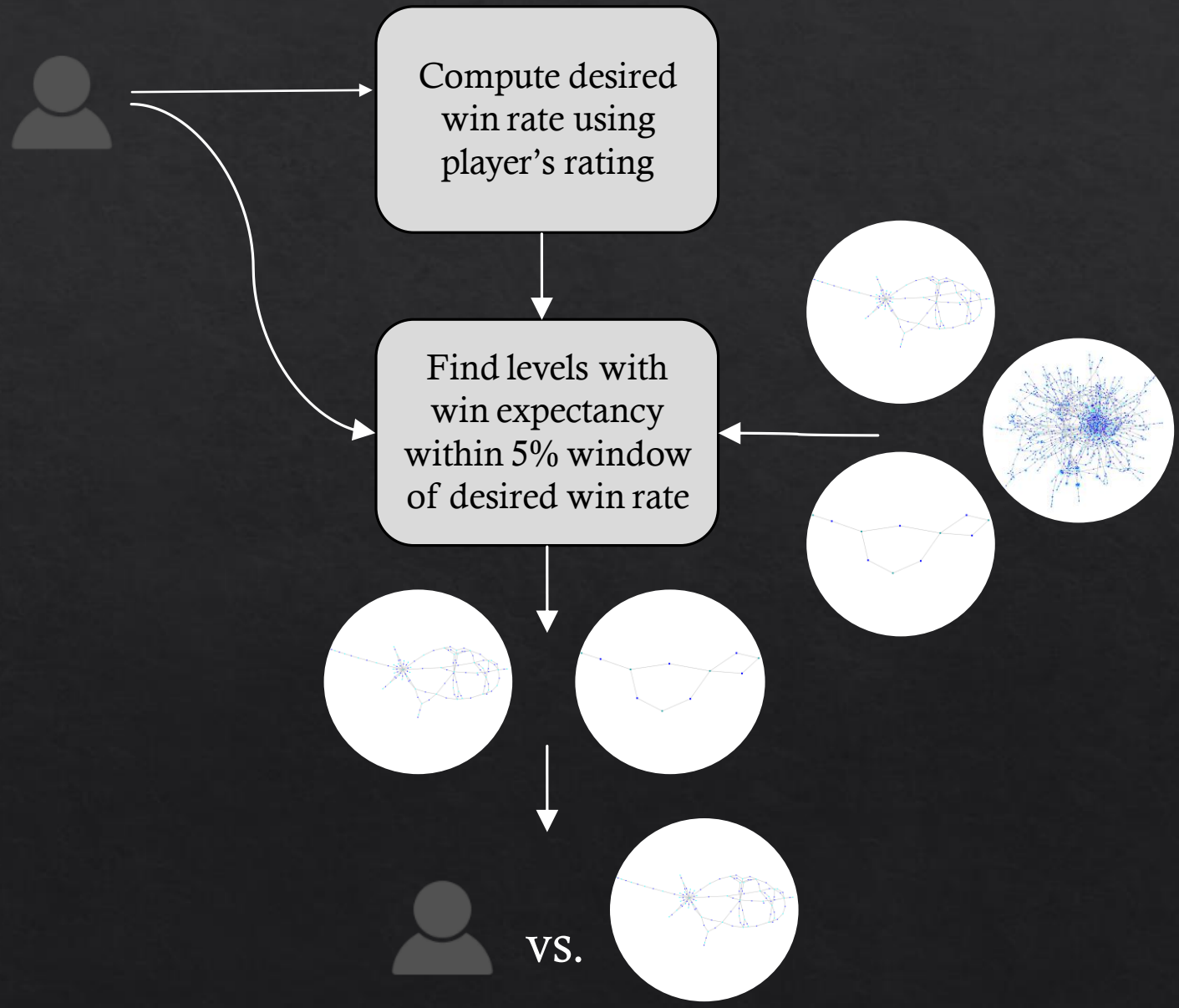


Phase 2: Ordering Experiment

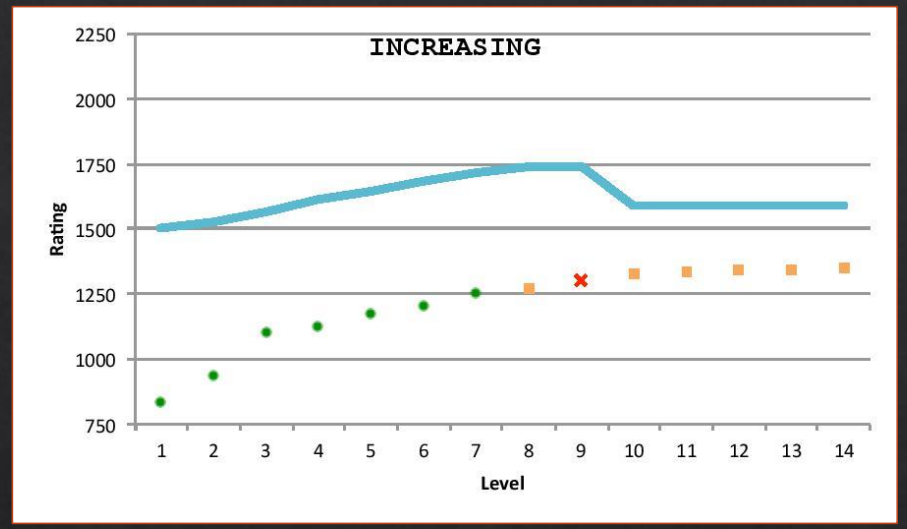
- 393 workers accepted HIT
- 294 completed HIT (75% completion rate)

- Ordering:
  - MATCHMAKING – 79
  - INCREASING – 99
  - RANDOM – 116

- Levels and players initialized with default Glicko2 parameters except levels were initialized with ratings from phase 1



Example Player Trajectories

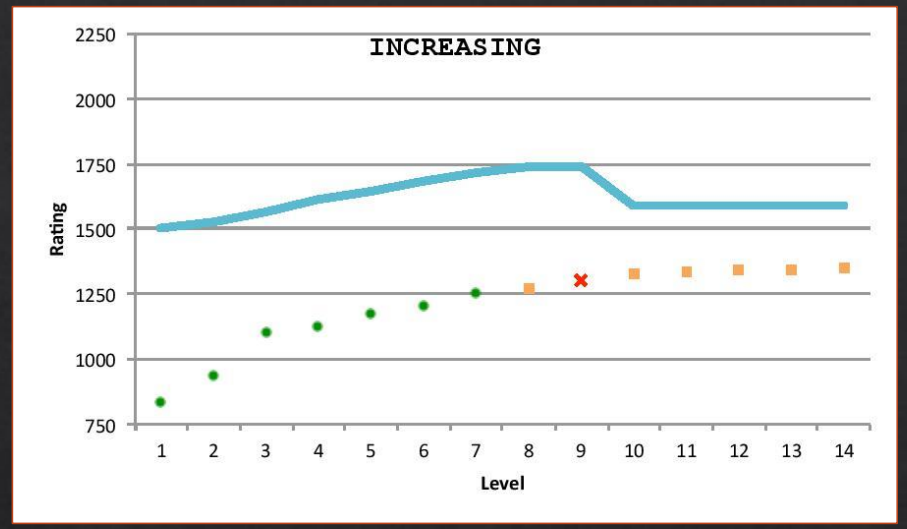


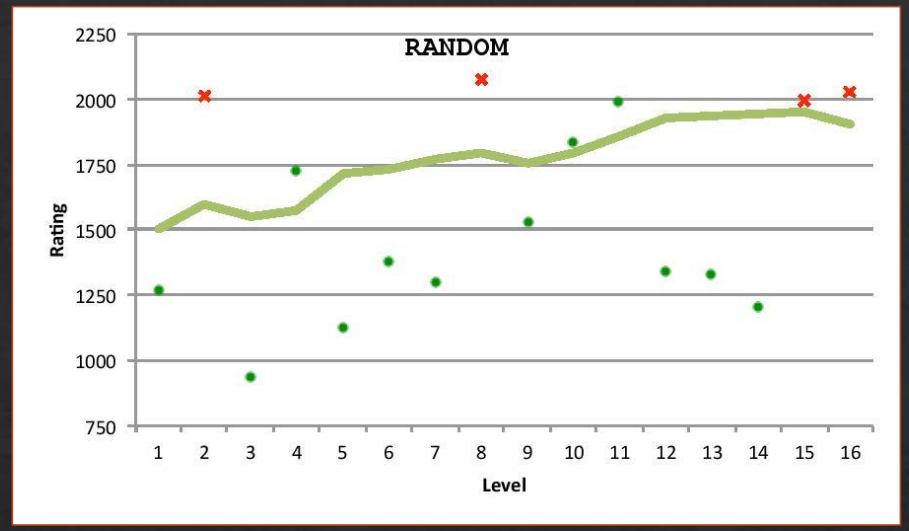
Complete (Win)

Forfeit (Loss)

Skip

Example Player Trajectories



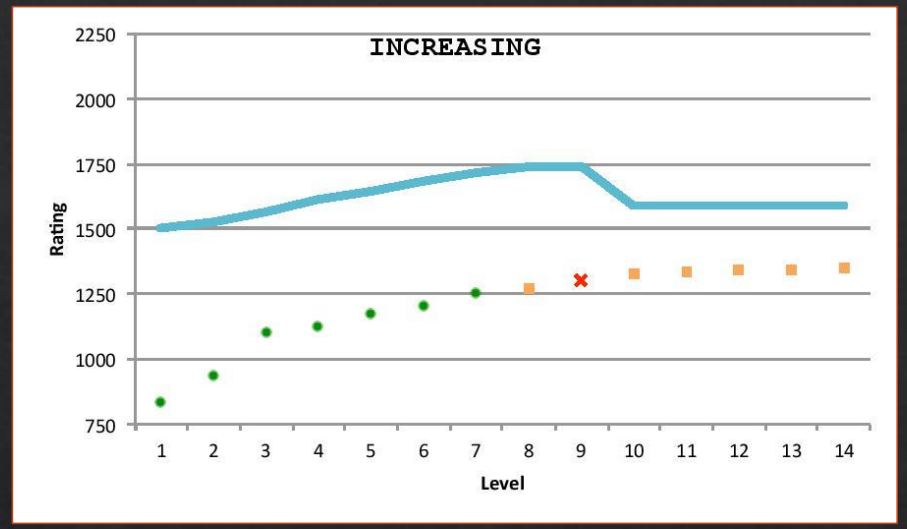


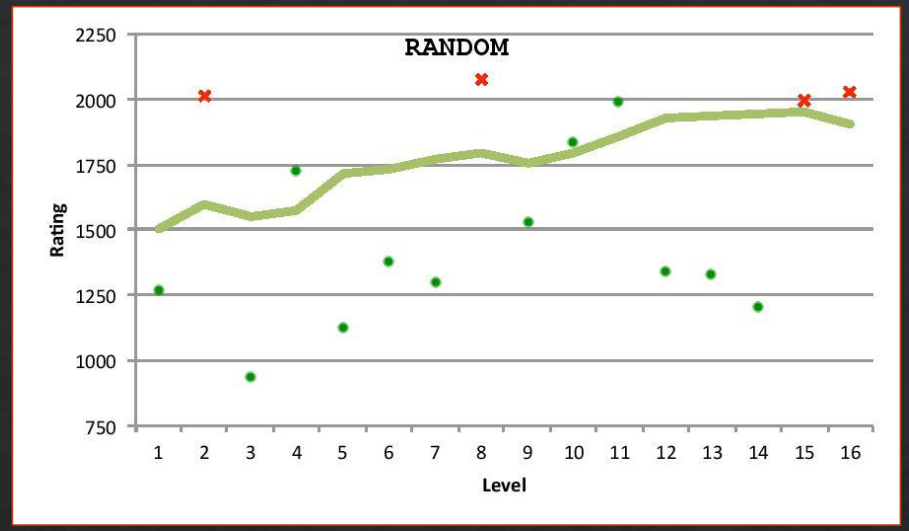
Complete (Win)

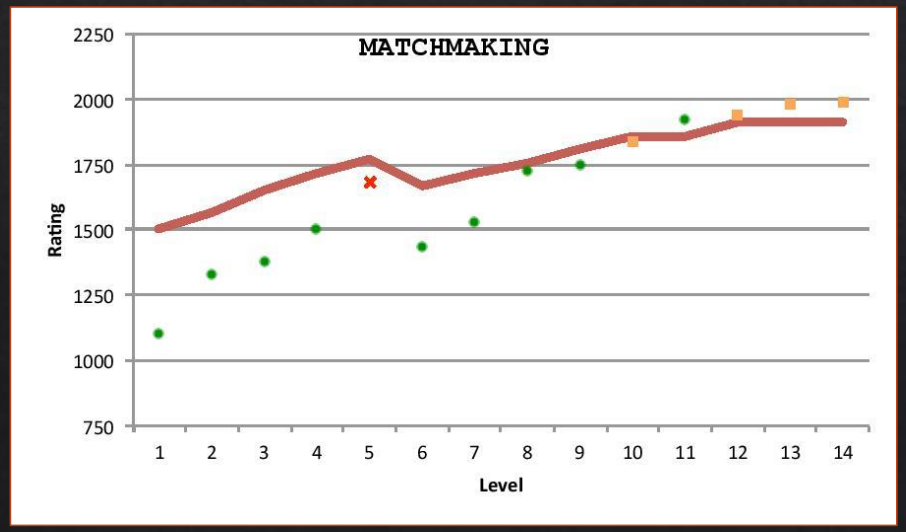
Forfeit (Loss)

Skip

Example Player Trajectories







Complete (Win)

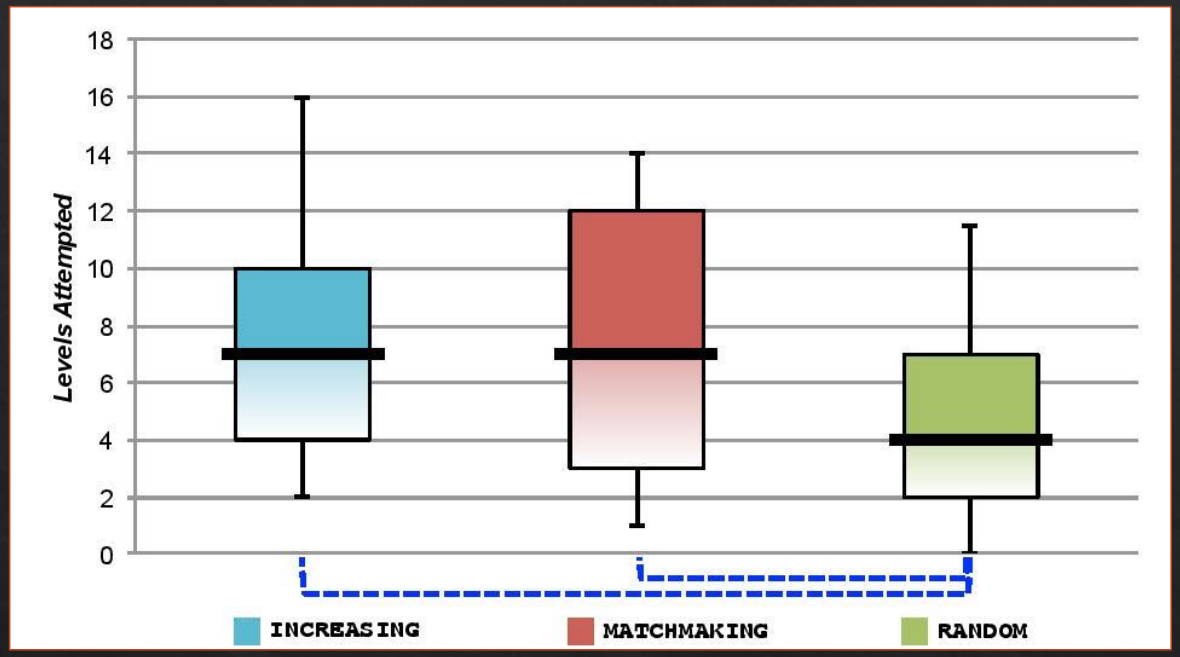
Forfeit (Loss)

Skip

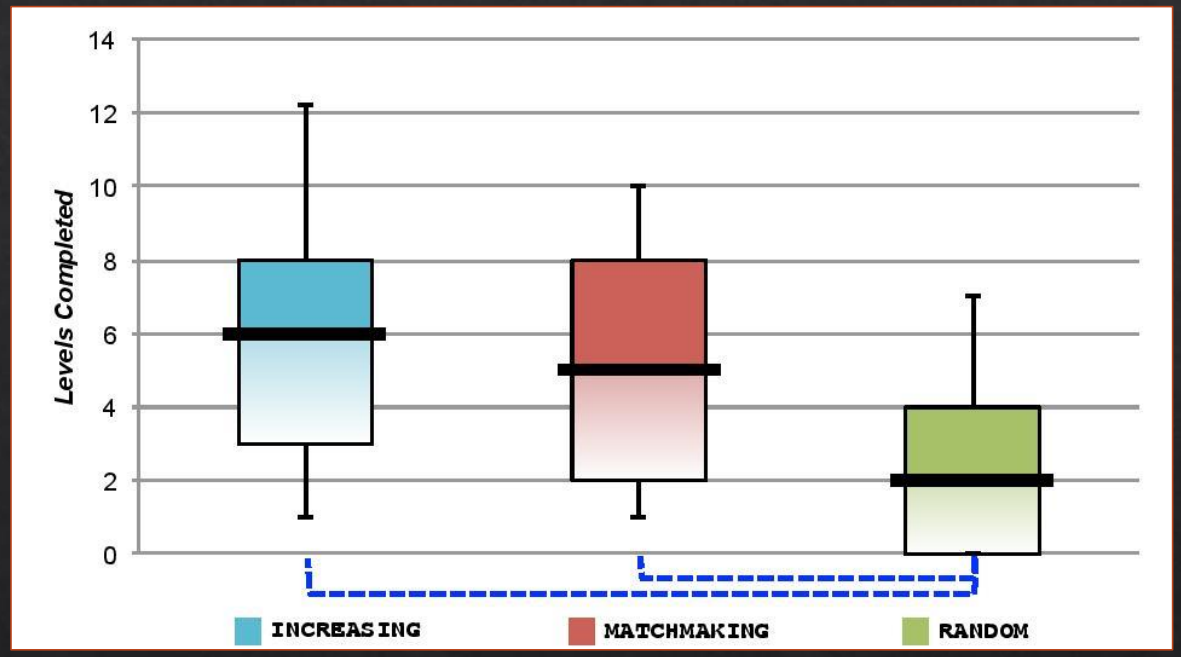
Measures of Engagement

- *Quantitative Engagement:* The *amount* of work done by players
  - Challenge Time
  - Levels Attempted
  - Levels Completed
- *Qualitative Engagement:* The aggregate *difficulty* of work done by players
  - Highest Rating (of any level completed by a player)
  - Per-Level Rating (avg. difficulty/rating of completed levels)
- Statistical Tests: Omnibus Kruskal-Wallis Test, post-hoc Wilcoxon Rank-Sum Test

Quantitative Engagement



Levels Attempted

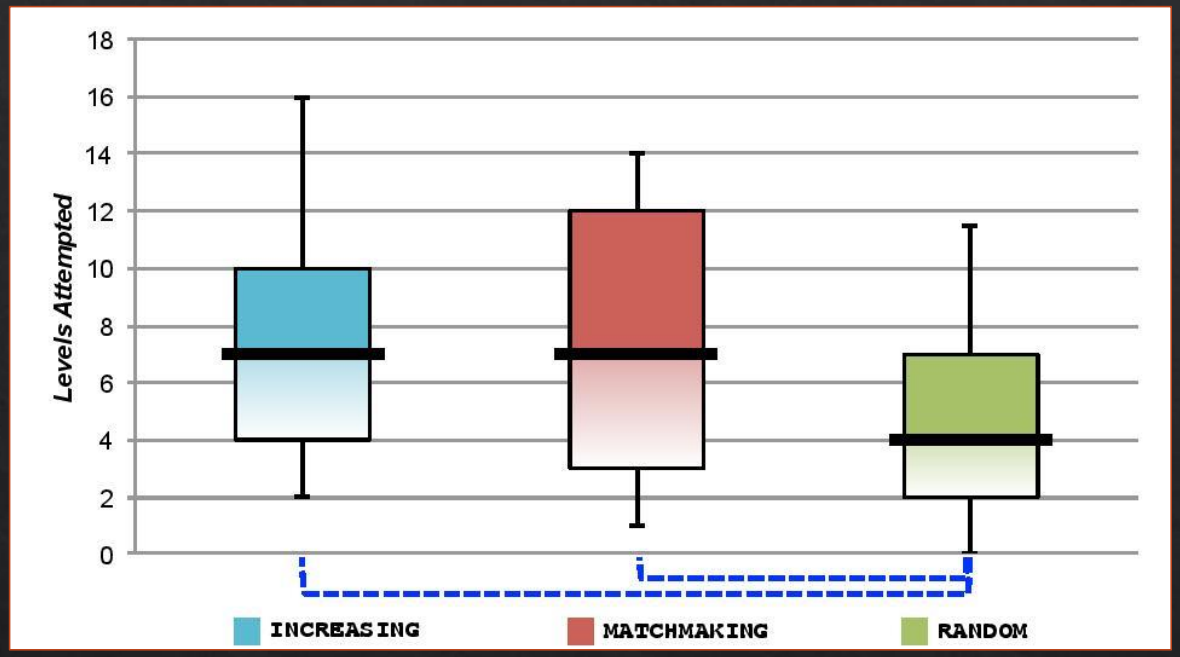


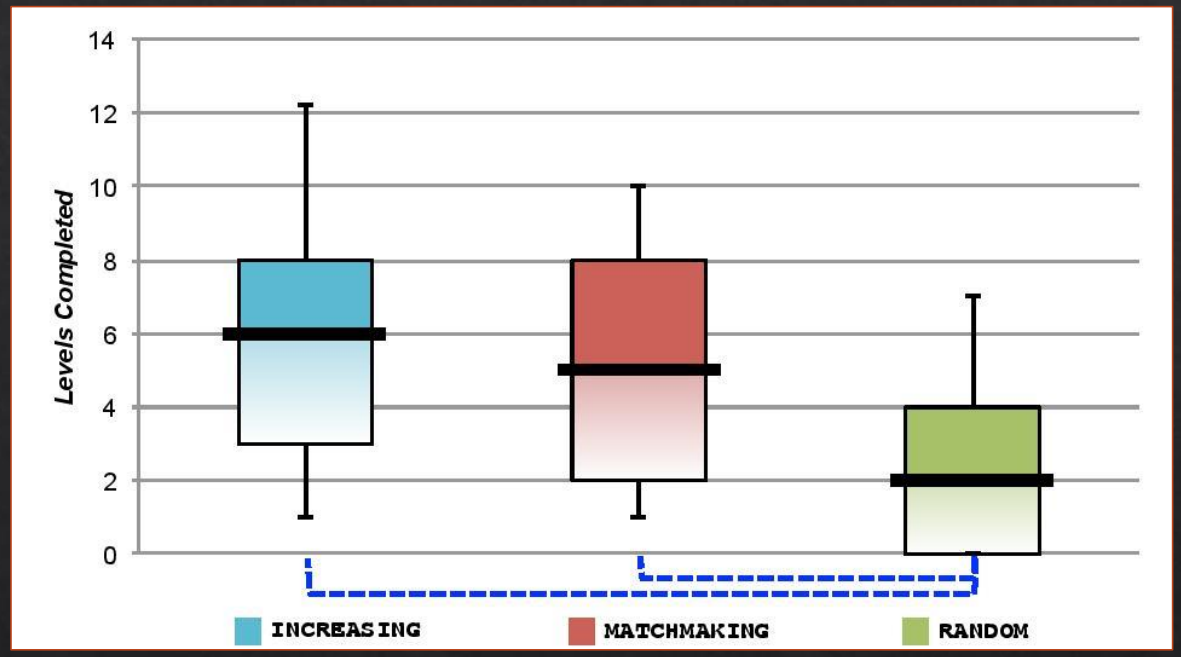
Levels Completed

Box Plots: 10th, 25th, 50th, 75th, 90th percentiles

– – – Significant differences

Quantitative Engagement





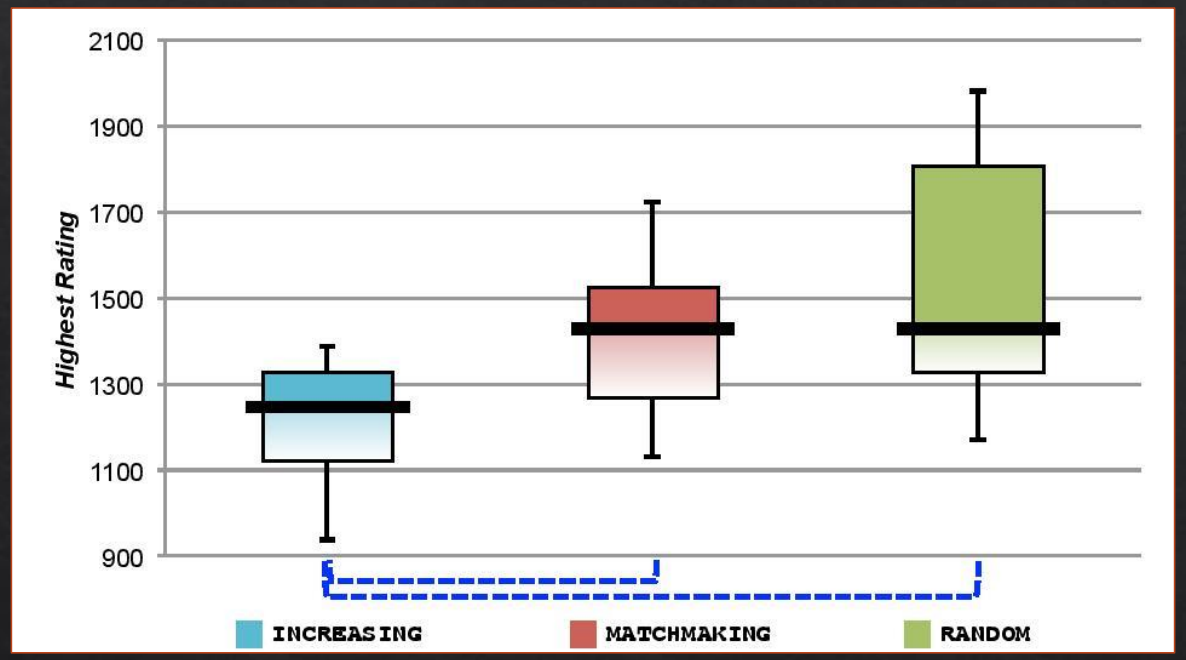
Levels Attempted ***Levels Completed***

No differences among conditions for Challenge Time

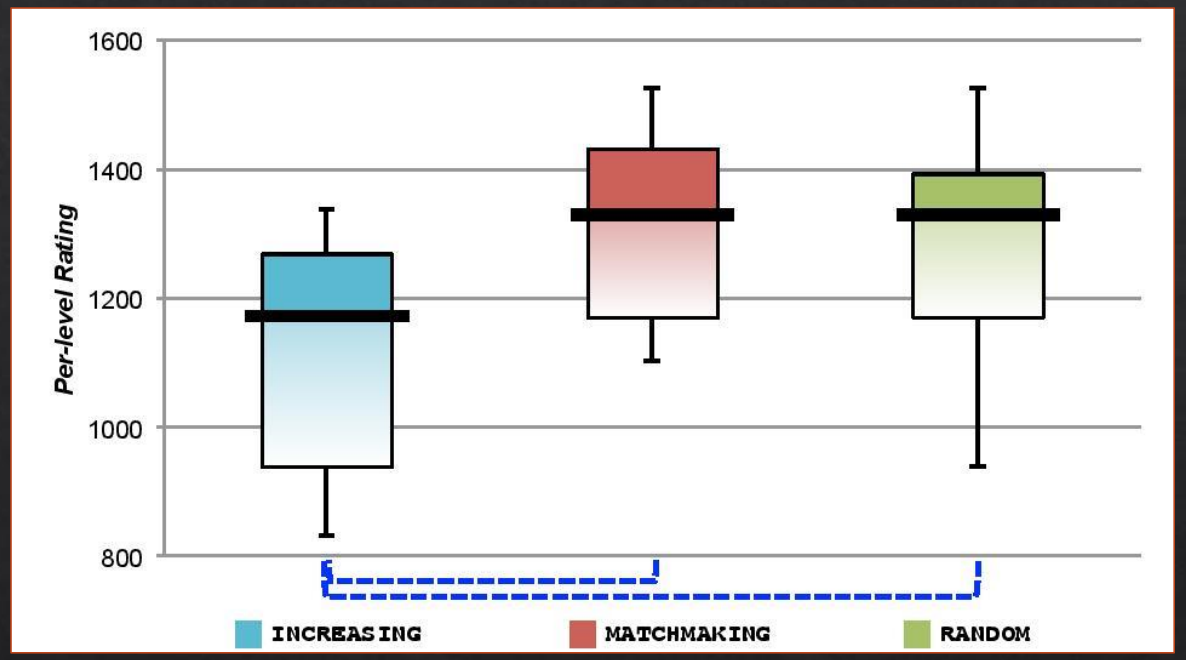
Box Plots: 10th, 25th, 50th, 75th, 90th percentiles

– – – Significant differences

Qualitative Engagement



Highest Rating



Per-Level Rating

Box Plots: 10th, 25th, 50th, 75th, 90th percentiles

– – – Significant differences

Discussion

- H1 is partially supported
  - Quantitatively, INCREASING does better
  - Qualitatively, RANDOM does better

Discussion

- H1 is partially supported
  - Quantitatively, INCREASING does better
  - Qualitatively, RANDOM does better
- H2 is rejected
  - Quantitatively, MATCHMAKING performed better than RANDOM but on par with INCREASING
  - Qualitatively, MATCHMAKING performed better than INCREASING but on par with RANDOM

Discussion

- MATCHMAKING and RANDOM engage players to do equivalently difficult work --- but MATCHMAKING engages them to do so for a greater number of levels

Discussion

- MATCHMAKING and RANDOM engage players to do equivalently difficult work
  --- but MATCHMAKING engages them to do so for a greater number of levels

- MATCHMAKING and INCREASING engage players to do more than RANDOM
  --- but MATCHMAKING engages them to do more *difficult* work

Conclusion

- MATCHMAKING is thus a ‘best of both worlds’ approach
  - Outperforms RANDOM in terms of *quantity* of work done
  - Outperforms INCREASING in terms of *quality* of work done

Future Work

- Effects of exposing players to rating system

Future Work

- Effects of exposing players to rating system
- Online (one-phase) system

Future Work

- Effects of exposing players to rating system
- Online (one-phase) system
- Other games with unknown difficulties

Future Work

- Effects of exposing players to rating system
- Online (one-phase) system
- Other games with unknown difficulties
- Generating levels to fill in gaps

Acknowledgments

This work was supported by a **Northeastern University** TIER 1 grant and partly conducted in the **Digital Creativity Labs** (digitalcreativity.ac.uk), jointly funded by **EPSRC/AHRC/InnovateUK** under grant no. EP/M023265/1. 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.

Contact

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

Variable	Omnibus	MATCHMAKING / INCREASING	INCREASING / RANDOM	RANDOM / MATCHMAKING
Challenge Time (s)*	$n.s., H(2) = 1.62$	395 / 329	329 / 386	386 / 395
*Levels Attempted**	$p < .001, H(2) = 14.91$	7 / 7 $n.s., U = 3869$	7 / 4 $p < .001, U = 4143$ $r_{rb} = 0.28$	4 / 7 $p = .003, U = 3441$ $r_{rb} = 0.25$
*Levels Completed**	$p < .001, H(2) = 45.80$	5 / 6 $n.s., U = 3536$	6 / 2 $p < .001, U = 2911.5$ $r_{rb} = 0.49$	2 / 5 $p < .001, U = 2672$ $r_{rb} = 0.42$
*Highest Rating***	$p < .001, H(2) = 55.67$	1431 / 1249 $p < .001, U = 1631$ $r_{rb} = 0.52$	1249 / 1431 $p < .001, U = 1436$ $r_{rb} = 0.60$	1431 / 1431 $n.s., U = 2581$
Per-level Rating†	$p < .001, H(2) = 224.41$	1328 / 1171 $p < .001, U = 88440$ $r_{rb} = 0.45$	1171 / 1328 $p < .001, U = 84872$ $r_{rb} = 0.43$	1328 / 1328 $n.s., U = 102830$

Table 1: Summary table of variable analysis. Variables analyzed using *all players, **players who completed at least one level, and †all completed levels. Shaded cells show significant post-hoc comparisons. Medians are given.