Astonishing irregularities in Sokoban gameplay data

Adani Abutto
aabutto@stanford.edu
Stanford University
Stanford, CA, USA

Abstract

Game environments are fun and also highly controlled. This offers fertile ground for studying various aspects of human learning, motivation, play, exploration, planning, and many other aspects of cognition and behavior. Here, I draw on data from the popular puzzle game "Sokoban" to analyze trends in player-level improvement on various performance metrics (likelihood of solving puzzles, how far puzzles were solved, time until solution, and number of moves). Results were ambivalent: Players generally seemed to not improve or do worse in terms of likelihood of solving puzzles with an increasing number of puzzle trials played, but also showed some decrease in time spent until solution and number of moves attempted. Albeit inconclusive, this opens up a variety of follow-up analyses and directions for future research.

ACM Reference Format:
Adani Abutto. 2025. Astonishing irregularities in Sokoban gameplay data. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Sokoban is a popular puzzle game in which people pursue a well-defined goal within a well-defined, highly controlled world (see Figure 1). Importantly, game environments like Sokoban represent settings that are also *fun* for players, and are therefore uniquely conducive to the study of learning, motivation, play, exploration, planning, and many other aspects of human cognition and behavior (see Allen et al., 2023, for a comprehensive review). Here, I leverage data collected in such a setting to examine possible systematic variability in learning, and what may predict such variability in learning.

2 Methods

The data used here were collected by Chu et al. (2025) and sourced from https://github.com/cogtoolslab/fun-puzzles. The dataset encompasses demographic and gameplay information from N = 205 U.S. adult participants (42.4% female) recruited from Prolific. The subjects' average age was 36.74 years (SD = 12.31, range: 18-74); most of them had never played Sokoban before (n = 126), some had played a few times (n = 53), and a minority had either played several times (n = 18) or was very familiar with the game (n = 8).

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

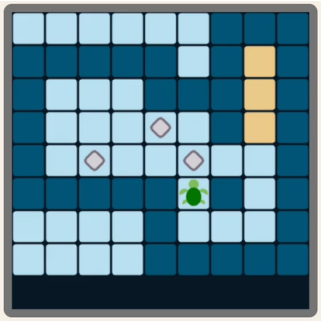

Figure 1: Example puzzle. Grey diamonds represent gems which need to be pushed to the yellow target tiles. Players control the turtle avatar which can move across one tile at a time (up, down, left, right) and push no more than one gem at a time. Dark tiles represent impassable terrain. A puzzle is solved when all three gems reach their target location within 5 minutes. Puzzle size ranged from 3x3 to 7x7 tiles.

2.1 Data and Measures

Subjects played a custom online version of Sokoban (see Figure 1). The original authors had identified and selected a set of 24 puzzles fit for novice players from a broader collection of popular Sokoban puzzles (cf. Chu et al., 2025 for details). These puzzles were then split into three sets of 8 puzzles whereby each of their subjects attempted just *one* (randomly drawn) set of 8 puzzles. Subjects were free to play puzzles in any order; after each puzzle, they were shown a screen displaying all eight puzzles, and they could pick which one they would like to play next. A puzzle was considered solved if all three gems were moved to their target location within the time limit of 5 minutes. Subjects also provided difficulty and enjoyment ratings for each puzzle.

The nature of the online setup allowed for extensive logging of every player's in-game actions (e.g., individual moves and move trajectories over time) and choices (e.g., which puzzle to play next). The original authors also collected and computed various puzzle metadata (e.g., puzzle size, puzzle complexity). In addition, all subjects responded to various standardized survey questions. In the present analyses, I focus on the following subject-level variables: Puzzle solved (yes or no), time until solution (in seconds), number of boxes solved (0 to 3), number of input events, self-reported effort (1 = very low to 5 = very high; reported across all puzzles), and age. All subject-level performance metrics were available for each of the 8 attempted puzzles. I also used the following puzzle-level variables: Puzzle trial number (1 through 8), puzzle set (A, B, or C), and two difficulty metrics: The number of iterations it took for the A* planning algorithm (Hart et al., 1968; Todd et al., 2023) to solve

each puzzle (max. 1 million), and the average solve rates among a broader human player base for each puzzle.

2.2 Inferential Statistical Analyses

For the purpose of this short report, I sought to answer just one overarching question: Did players improve in their performance as they played more? To this end, my main unit of analysis was subject-level performance as measured by four performance metrics: Whether they solved the puzzle in question, how many boxes they solved, time until solution, and number of moves performed.

My main inferential statistical analyses assessed (1) whether player performance increased on said performance metrics over time (i.e., with increasing number of puzzles attempted), and (2), if so, which (additional) factors predicted improving performance. Importantly, given that players completed puzzles in a non-standardized order, I controlled for puzzle-level variance (difficulty metrics, puzzle id) in all of my analyses.

To examine how much each source of variance was systematically related to variance in performance improvement over time, I used mixed-effects regression as my main modeling approach: I regressed the different performance metrics (binary: solved vs not solved; continuous: number of boxes solved, time until solution, and number of moves) on trial number and later added the additional predictors as fixed effects. I included random intercepts and slopes per subject to allow for individual variability in improvement.

For each model, I report Chi-squared (based on the comparison to an "empty" model), unstandardized regression coefficients, standard error (SE), and p-values. Additionally, I report model metrics AIC and BIC. All of the above was done using the R packages *lme4* (Bates et al., 2003), *ggeffects* (Lüdecke, 2020), *broom.mixed* (Bolker et al., 2019), and *emmeans* (Lenth, 2018).

3 Results

3.1 Descriptive Statistics

Across the 24 puzzles, solve rates ranged from 0% (Microco 28) to 81% (Dimitri 37); Figure 2 below shows solve rates across the full set of 24 puzzles. Further plots are available under https://github.com/adaniabutto/datasci294l_final-project/tree/main/figures.

There was also considerable puzzle- and subject-level variability on the other three performance metrics: On average, players solved 2.1 out of 3 boxes (SD = .83), spent 143.7 seconds until solution (SD = 76.1, range: 18-300), and performed 177 moves (SD = 112.0, range: 3-622; includes puzzle trials where players did not solve the puzzle at hand). Figure 3 gives a first visual indication that players improved on some (but not all) performance metrics when comparing the first 4 puzzles versus the last 4 puzzles they attempted. In the next step, I examined whether this trend was reflected in statistical effects.

3.2 Inferential Statistical Analyses

3.2.1 Model 1: Does success on a previous puzzle predict success on the concurrent puzzle? As a first pass, I examined whether success on a given puzzle was predicted by success on the preceding puzzle. That is, did players who tended to do well on a puzzle tend to also have done well on the one right before? I predicted that, even when controlling for effort and holding constant puzzle difficulty, players who solved a previous puzzle would be more likely to also solve the current puzzle. The results of the mixed-effects logistic regression supported this, with the predicted likelihood of puzzle-solving being greater given preceding success: χ^2 = 14.39, df = 1, unstandardized β = .335, z = 3.84, SE = .088, p < .001; AIC = 2282.5, BIC = 2309.5.

3.2.2 Model 2: (How) does performance improve over time? Next, at a more fine-grained level, I examined whether the number of puzzle trials (trial order) predicted player improvement on the four performance metrics. For our first metric, solved vs not solved, the results of the mixed-effects logistic regression suggested that the predicted likelihood of puzzle-solving *decreased* with an increasing number of puzzle trials: χ^2 = 12.69, df = 1, unstandardized β = -.09, z = -3.54, SE = .028, p < .001; AIC = 1840.4, BIC = 1878.2. The same negative effect held true when predicting the number of boxes solved: χ^2 = 5.44, df = 1, unstandardized β = -.02, z = -2.33, SE = .008, p = .020; AIC = 4820.1, BIC = 4857.9. The average predicted duration until solution also increased with an increasing number of puzzle trials: χ^2 = 23.49, df = 1, unstandardized β = .01, z = 4.85, SE = .002, p < .001; AIC = 19297, BIC = 19328. The predicted number of moves also tended to increase with increasing number of puzzle trials: χ^2 = 620.3, df = 1, unstandardized β = -.02, z = -24.90, SE = .001, p < .001; AIC = 61033, BIC = 61071, though this model was nearly unidentifiable and its results should be interpreted with caution.

When re-running models including a) a random slope for puzzle to account for puzzle-level variability, and b) random intercepts and slopes per subject to allow for individual variability in improvement, the former two effects either continued holding true or diminished in statistical significance. Not the latter two, however: The models now predicted that number of moves as well as time until solution decreased with increasing puzzle trials. Figure 4 visualizes these growth model results.

4 Discussion

In the present study, I analyzed data from subjects solving Sokoban puzzles to investigate learning trajectories and its predictors. In sum, the model results yielded mixed evidence, suggesting that independent of self-rated effort and puzzle-level variability, players did *not* consistently tend to improve as they played more puzzles; in two cases (likelihood of solving puzzles, number of boxes solved), they performed worse. I also found some evidence that while the actual solve rate may not have improved, players may have performed better in terms of time until solution and the number of moves made.

4.1 Limitations and Future Directions

The results here should be interpreted with caution. For one, it is crucial to note that puzzles were not presented to subjects in a standardized order, and while including puzzle features as well as puzzle ID in the models technically helps account for puzzle-level effects, the number of data points for any given particular order of solving puzzles was highly limited. This is reflected both in the failures to converge and the highly volatile model predictions as seen in Figure 4. However, the results captured here may capture fatigue effects. Subjects played for a total of 50 minutes, which is a significant amount; it is plausible that, as time went on, this indeed

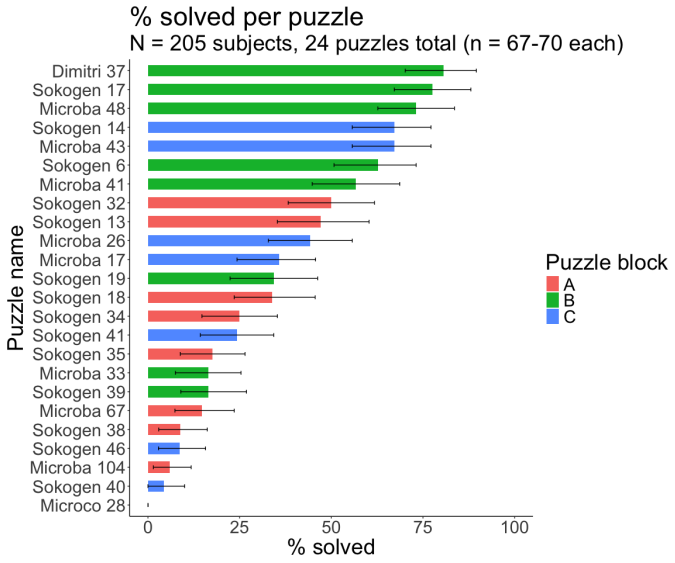

Figure 2: Solve Rates; error bars denote bootstrapped 95% CIs.

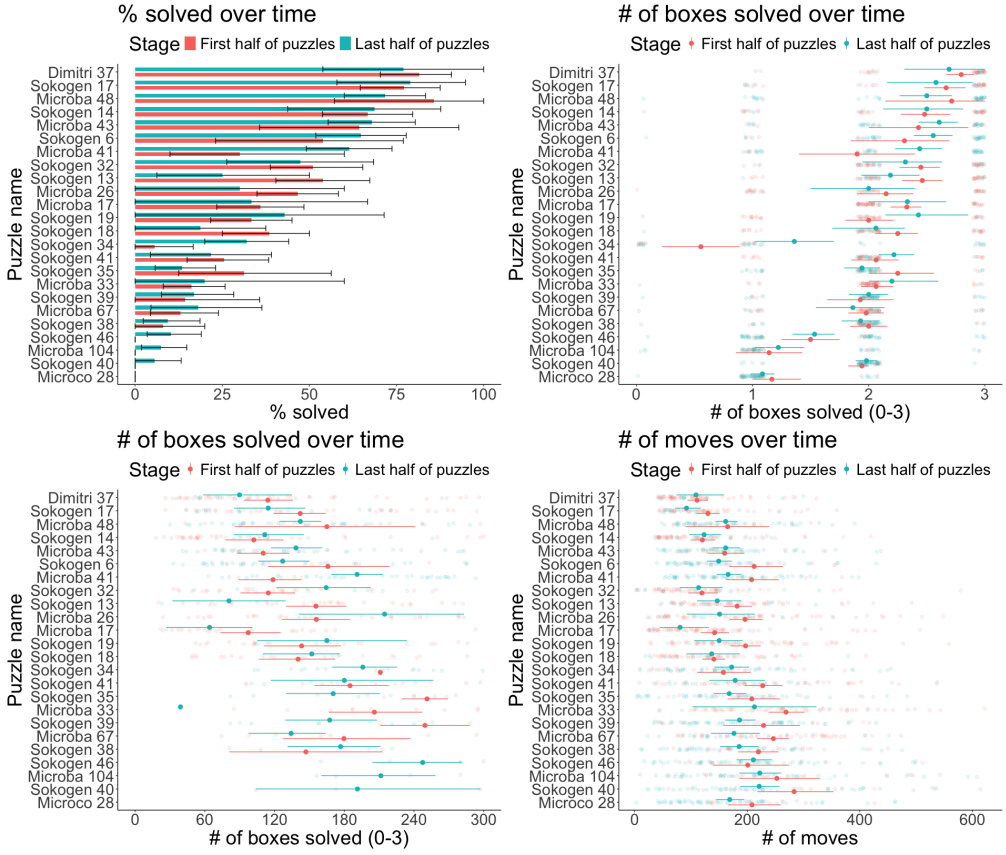

Figure 3: Performance metrics split by first and last half of puzzles

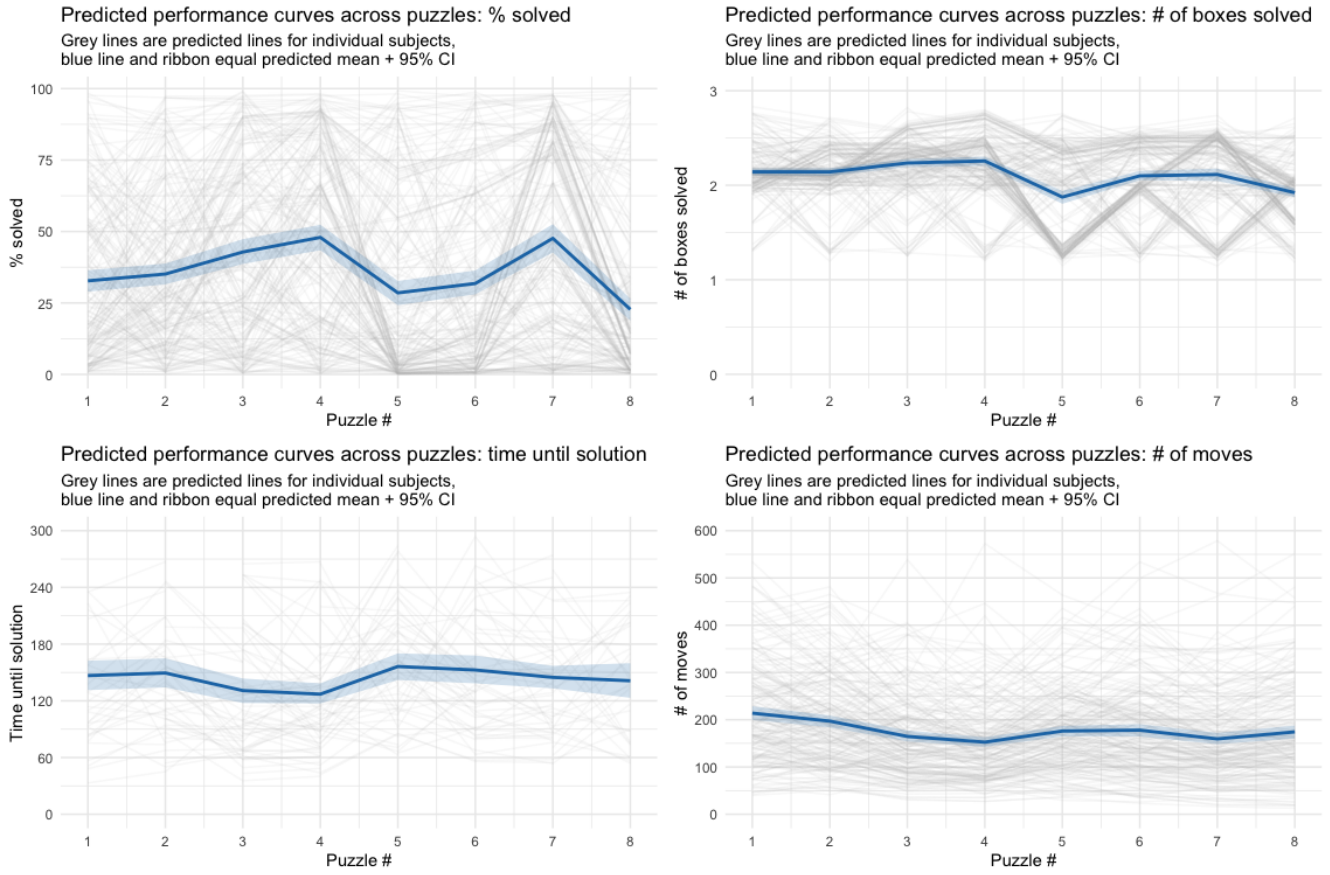

Figure 4: Growth model results

led them to tend to solve fewer puzzles, fewer boxes, spending less time thinking through moves, and making fewer moves overall.

Future research can further look into learning trajectories in multiple ways. For one, among players who *did* improve, what were they actually learning? Did they learn to avoid bad moves by thinking longer? Did they learn "super sequences", getting quicker at spotting and executing move chains that get them to a critical point for moving gems to target locations? Secondly, may they have learned about their own puzzle-solving abilities, for instance being more likely to try another hard puzzle after succeeding at one they thought they could not solve? Is this reflected in some systematicity in puzzle choices?

Beyond analyzing learning, there is a host of exciting analyses around (1) player profiles, (2) puzzle-level variability, and (3) players' exploration behavior that were not included here. For (1), future research may look into whether and how players who succeeded on a given puzzle differed from players who did not succeed. For instance, when controlling for puzzle difficulty, were succeeding players generally more familiar with Sokoban? Did they perform more restart and undo actions, and at which points? Did they generally think for longer, and at which points?

Secondly, are more or less competent participants generally indicated by a combination of how well they perform and how they evaluate the puzzles? (e.g., there may be capable players who enjoy easy wins versus hard-won solutions). For (2), puzzles such as Microba 41, Sokogen 32, Sokogen 13, and Microba 26 were solved by just about half the players. These puzzles could provide interesting insights into how players who solved versus did not solve the puzzle differed. Did they solve the puzzles around the same time, and how much do they vary in their traces? Do these puzzles generally vary from other puzzles in that there are more possible ways to solve them, or fewer ways to solve them but requiring more time?

Finally, for (3), future research may want to compare how player traces compare to A* traces both for a) paths to solution, and b) paths from specific bottleneck points. For example, when people get stuck, do they approach their next steps in a way similar to A*? And when systematically tweaking the heuristics A* has access to, do player traces become more or less similar to A* traces?

While the present analyses yielded conflicting evidence about players' learning trajectories, they do corroborate that game environments offer fertile ground for studying various aspects of human learning, motivation, play, exploration, planning, and many other aspects of cognition and behavior. Given further research, they may

take us from astonishing contradictions to astonishing regularities in learning rates and trajectories.

5 Code and Data Availability

All analysis scripts are publicly available in the following GitHub repository: https://github.com/adaniabutto/datasci294l_final-project.

References

[1] Kelsey Allen, Franziska Brändle, Matthew Botvinick, Judith E. Fan, Samuel J. Gershman, Alison Gopnik, Thomas L. Griffiths, Joshua K. Hartshorne, Tobias U. Hauser, Mark K. Ho, Joshua R. de Leeuw, Wei Ji Ma, Kou Murayama, Jonathan D. Nelson, Bas van Opheusden, Thomas Pouncy, Janet Rafner, Iyad Rahwan, Robb B. Rutledge, Jacob Sherson, Özgür Şimşek, Hugo Spiers, Christopher Summerfield, Mirko Thalmann, Natalia Vélez, Andrew J. Watrous, Joshua B. Tenenbaum, and Eric Schulz. 2024. Using games to understand the mind. *Nature Human Behaviour* (June 2024), 1035–1043. doi:10.1038/s41562-024-01878-9

[2] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2003. lme4: Linear Mixed-Effects Models using 'Eigen' and S4. doi:10.32614/CRAN.package.lme4

[3] Ben Bolker, David Robinson, Dieter Menne, Jonah Gabry, and Paul Buerkner. 2019. Package "broom. mixed". (2019).

[4] Junyi Chu, Kristine Zheng, and Judith E Fan. 2025. What makes people think a puzzle is fun to solve? *Proceedings of the Cognitive Science Society Conference* (2025).

[5] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* (July 1968), 100–107. doi:10.1109/TSSC.1968.300136

[6] Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level Generation Through Large Language Models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games (FDG '23)*. New York, NY, USA, 1–8. doi:10.1145/3582437.3587211