A Study in Extensible Algorithmic Thinking

Andrew Nam
Department of Psychology
Stanford University
ajhnam@stanford.edu

James McClelland
Department of Psychology
Stanford University
jlmcc@stanford.edu

1 Introduction

In this study, we explore how well humans can extend a generalizable principle from limited exemplars to novel instances. We introduce a task called the Hidden Singles task, named after the Sudoku technique *Hidden Singles*, where given a configuration of numerals on a 9x9 Sudoku, subjects are asked to solve for the content of an empty target cell using the existing hints on the grid. We will recruit participants with no prior exposure to Sudoku, and then teach them the basic rules of Sudoku and the Hidden Singles technique through a series of self-paced instructions and exercises. After the tutorial and practice phase using similarly constructed puzzles, the participants will be tested on new puzzles with features varied from the tutorial and practice puzzles to measure changes in accuracy and time taken to solve the puzzles.

This study has been iteratively refined through multiple pilot experiments. Here, we specify the main analyses we plan to conduct for the final version of the study. The details of the methods will be available in the full publication; here we offer descriptions of the experimental design, exclusion criteria, and statistical analyses sufficient to understand the nature of the study. We also provide sample outputs and figures of our planned analyses using our pilot data.

2 Experiment Design

The experiment begins with a brief diagnostic test and survey described in **Section 3** to screen out participants who have prior experience with Sudoku. The experiment is divided into 3 phases: (1) Tutorial Phase where participants learn about Sudoku and the Hidden Singles technique, (2) Practice Phase where participants practice the technique, and (3) Test Phase where they are tested on their ability to apply the technique in generalized settings.

2.1 Sudoku Tutorial

The experiment begins with a series of tutorials using text instructions accompanied by relevant figures and exercises. First, the participants are told the following about Sudoku: "Sudoku is a puzzle with a 9x9 grid of numbers where each row, column, and 3x3 box must contain exactly one of each number from 1 to 9." Next, to ground their understanding of the rules, participants are given two exercises. The first exercise requires participants to identify and select the two numbers that form a contradiction (**Figure 1a** [1]). The second exercise removes one of the contradicting numbers and asks the participants to solve for the missing number (**Figure 1b**).

The experiment then proceeds to teach subjects the Hidden Singles technique where one searches for a candidate digit for a cell by eliminating other cells in a house (row, column, or box) as possible locations the digit could be in. First, each participant is shown a randomly generated puzzle and walked through the process of elimination to solve for the green cell by selecting the hints that constrain specific parts of the highlighted house (**Figure 2a**). Next, the participant is shown the same

[1] All examples shown in this document are examples of what one participant might see. Actual stimuli would be randomly generated for each participant.

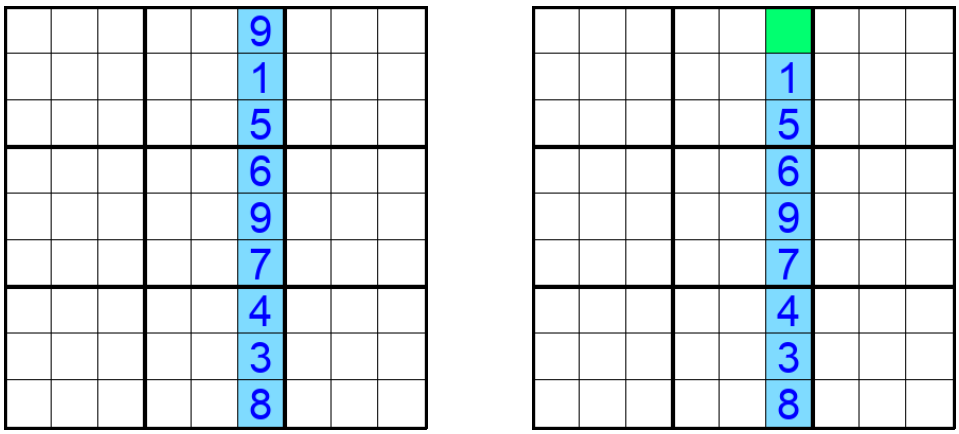

Figure 1: Exercises to ground the understanding of basic principles in Sudoku. The following instructions would be provided: (a) The column in the grid below does not contain every number between 1 and 9, but rather contains two copies of the digit 9, forming a contradiction. Select the two cells that create this contradiction. (b) One of the 9s forming the contradiction has been removed. Fill in the missing number in the green cell so that the column contains every number between 1 and 9.

puzzle augmented with distractor hints (a digit that also appears three times in the grid but has two possible cells, including the goal cell, that it could go in the highlighted house; see **Figure 2b**) and walked through a similar process of elimination, this time selecting the blue cells constrained by specific hints. The tutorial concludes by showing summaries of both the target and distractor digits as candidates (**Figures 2c & 2d**) simultaneously to visualize how the target digit successfully eliminates every blue cell as a candidate but the distractor does not. Across both sequences, participants receive two methods of approaching the Hidden Singles technique (from hints to empty cells vs from empty cells to hints), both a positive and negative example of applying the Hidden Single technique, and an implicit caution against using the most common digit as a heuristic.

Other than the initial one sentence description of Sudoku, we take particular care to avoid statements that describe the abstract formulation of the Hidden Single technique using role-filler words, making no suggestions that this technique should work in any other context than in the example provided. For instance, rather than a generalized statement such as "the Hidden Single technique identifies the solution digit in a cell by eliminating the digit as a valid candidate in the other 8 cells of a house containing the cell", we would refer to a particular state such as **Figure 2c** and use statements such as "the purple cells in the column cannot be a 2 because they share the same box with the 2 in the purple cell". We note, however, that some of the words in the tutorial may have role-filler effects such as "empty/filled cells" or "digit of interest." These phrases were used primarily to motivate the actions in the tutorial, such as "a 5 cannot be in any of the red cells because they already contain digits, so we shift our attention to the empty cells" when referring to a state similar to **Figure 2c**.

We used stimuli that shared features across the entire tutorial sequence. The Hidden Single tutorial and both exercises all use the same house (column 6 in **Figures 1 & 2**), location of the goal cell, and the solution digit. Moreover, any hints that appear in multiple puzzles have common digits and locations. For example, the puzzles in **Figure 2** have 1, 5, and 7 in the 2nd, 3rd, and 6th cells of column 6. Likewise, the exercises in **Figure 1** have the same numbers in the same cells. This restricts the sample space used during the tutorial and practice to allow testing of generalization outside of the training example space.

2.2 Puzzle Features

To test for extensibility of applying the Hidden Single technique, puzzles have 4 variable features:

- **House Type (HT):** the type of house to apply the Hidden Single technique to, indicated by whether a row or a column is highlighted blue. Can be *row* or *column*.

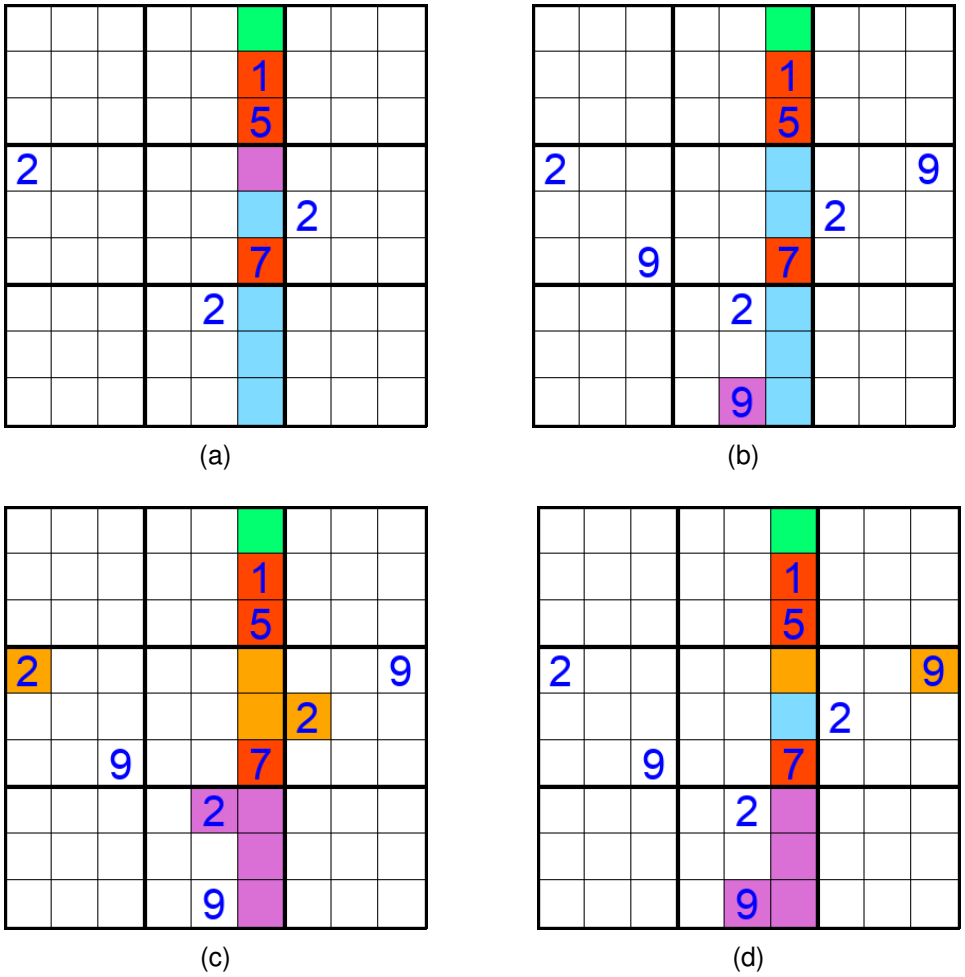

Figure 2: Hidden Single tutorial where 2 is the target and 9 is the distractor. (a) Positive example by solving for 2; task requires participant to select the 2 that shares a row with the purple cell. (b) Negative example by attempting to solve for 7. Task requires participant to select the three empty blue cells that share a box with the purple 9. (c) Color-coded visualization of why 2 is correct. (d) Color-coded visualization of why 9 is incorrect.

- **House Index (HI):** which row or column to apply the Hidden Single technique to, indicated by the row or column highlighted in blue. For rows, refers to the *n*th row from the top. For columns, refers to the *n*th column from the left. Can be any integer between 1 and 9.
- **Cell Index (CI):** which cell to solve for within the house, indicated by the cell highlighted in green. For rows, refers to the *n*th cell from the left. For columns, refers to the *n*th cell from the top. Can be any integer between 1 and 9.
- **Digit Set (DS):** the set of 4 digits to randomly assign the target and distractor digits from. At the start of the experiment, each participant is randomly assigned 2 disjoint sets each containing 4 numbers between 1 and 9, e.g. {1, 2, 3, 4} and {5, 6, 7, 8}. Each puzzle is first assigned one of the two the digit sets, then selects a number from the set as the target digit and a different number from the set as the distractor digit. Other hints are randomly assigned from the remaining 7 digits.

For example,in **Figure 2**, the **house type** is column, **house index** is 6 (referring to the 6th column from the left), and the **cell index** is 1 (referring to the 1st cell from the top).

To make puzzles nontrivial and encourage deductive reasoning, every puzzle is generated with at least 1 box constraint (a hint that shares a box with 3 of the empty cells in the house, simultaneously

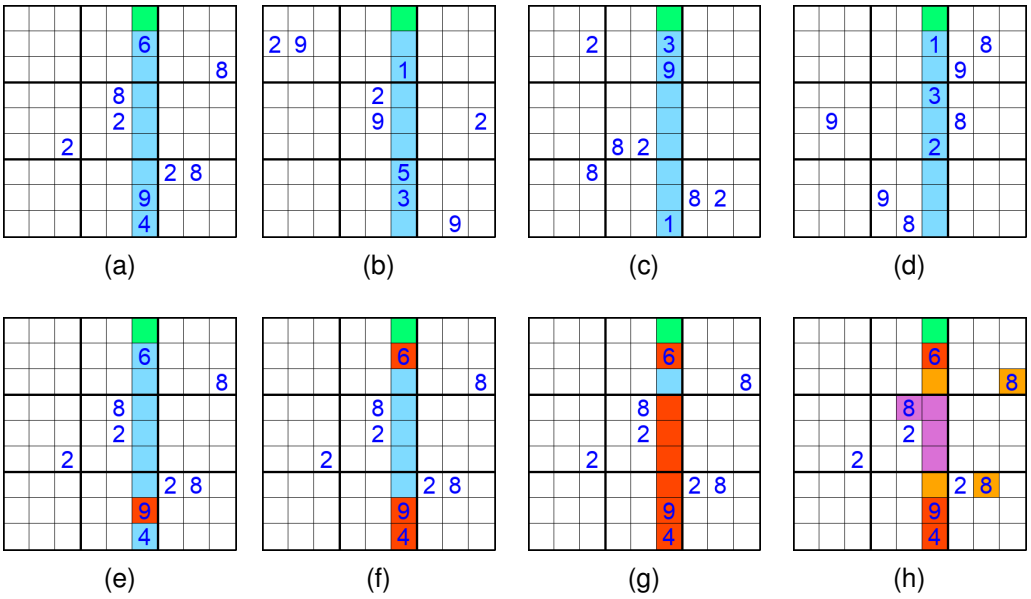

Figure 3: (a-d) First 4 puzzles of the Practice Phase as generated by the program for one participant. All puzzles highlight the 1th cell of the 6th column. Target and distractor digits are selected from {2, 4, 8, 9}. These would be the same features as the tutorial. (e-h) Feedback displays for incorrect responses to puzzle in *(a)* that would accompany text explanations. (e) Feedback display for a number already in the house, in this case 9. (f) Feedback display for a number not on the grid, such as 5. (g) Feedback display for the distractor, in this case 2. (h) Feedback display for correct response, in this case 8. Only shown if the participant first gave an incorrect response.

constraining all 3) and 2 orthogonal constraints (a hint that shares a column or row with an empty cell in the target house for puzzles with house types row or column respectively). Because this would require at least 3 hints that share digits with the target, we add a distractor digit with 3 hints that constrain the same box and one of the 2 other empty cells, making both the target and distractor digits salient as potential candidate target digits. **Figure 2c** shows box constraints in purple and orthogonal constraints in orange.

2.3 Practice Phase

In the Practice Phase, participants are given 25 puzzles to solve, each with unlimited attempts and time, with the goal of allowing them to saturate on their proficiency in using the Hidden Single technique. All puzzles in this phase use the same house type (HT), house index (HI), cell index (CI), and digit set (DS) as the tutorial, only varying the hint locations and the actual digits used as shown in **Figure 3**. During this phase, all incorrect attempts would provide visual feedback (**Figure 2b**) with text providing detailed explanations for why the given response was incorrect using specific references, preferring specific phrasing such as 'the empty orange cell constrained by the orange 8' over 'cells constrained by sharing a row with the target digit'. Participants would be rewarded only if the puzzles were solved on the first attempt.

2.4 Test Phase

In the Test Phase, participants are presented 8 sets of 8 puzzles to solve, each within 2 minutes using a single attempt, and are rewarded for correct responses. Unlike the Practice Phase, participants receive feedback only indicating whether or not their responses were correct. In this phase, puzzles could vary across the 4 feature dimensions: house type (HT), house index (HI), cell index (CI), and digit set (DS). We define test puzzles to have been conditioned with a feature if it differs from the tutorial for that feature, yielding 16 possible combinations of possible puzzle conditions (including control puzzles which share features with the tutorial and Practice Phase puzzles). For example, the puzzle in **Figure 4a** has the target cell on the same location as the puzzle in **Figure 4g** (the control)

at Row 1, Column 6, but highlights the row instead of the column, thus making it a HT puzzle. The puzzle in **Figure 4l** also adds the HI condition by changing the highlighted row to Row 2, but still keeps the target cell in Column 6.

To control for any ordering effects, each participant receives a constrained randomized 8 x 8 balanced Latin square (8 sets of 8 puzzles) using the 8 combinations of spatial conditions (HT, HI, CI) with the following properties:

- Each of the 8 combinations of the spatial conditions would appear in the k^{th} position of each set exactly once
- For each pair of spatial combination $A \neq B$, A would appear before B in 4 sets and B would appear before A in 4 sets
- 4 random puzzles in each set would be treated with the digit set condition (use a different digit set than the tutorial puzzle) while the remaining 4 puzzles would not
- In every two sets (16 puzzles), each of the 16 possible combinations of conditions would appear exactly once
- Each set would contain exactly 4 instances of each condition
- Each participant would have 4 instances of each of the 16 combinations across the entirety of the Test Phase

2.5 Exit Survey

2.5.1 Strategy Abstraction and Articulation

After the experiment, participants would be presented a series of questions to probe for any indication of their capacity to abstract and linguistically articulate the strategies they used to solve the puzzles. The questionnaire section would begin with brief instructions and 3 true-false attention check questions about the instructions. Next, subjects would be presented with a single puzzle to solve with the following characteristics:

- Same house type as tutorial
- Same digit set as tutorial
- Center box contains the goal cell
- Different house index and cell position than tutorial
- Single, untimed attempt.
- Not allowed to change response (after initial submission)
- No feedback whatsoever

The questions would then appear one at a time alongside the puzzle (filled in with the participant's response) and progressively increase in specificity towards identifying the participant's solution strategy.

1. Number response question: "How confident do you feel that your answer is correct, expressed as a percentage?" Participants could respond with numbers ranging from 0 to 100 in increments of 5.
2. Free response question: "Explain as clearly as possible the steps you went through to choose your answer. Please be as detailed as possible so that someone else could replicate your strategy by following your response."

   If the participant's response to the questionnaire puzzle was not the target or distractor, they would skip forward to question 10.
3. Single selection question: "There are two numbers in the puzzle that occur three times outside of the row/column containing the target cell. Which of the following best describes how you chose between the two candidate numbers to consider?"
   (a) I noticed something in the puzzle that initially made one candidate seem more likely to be correct than the other.

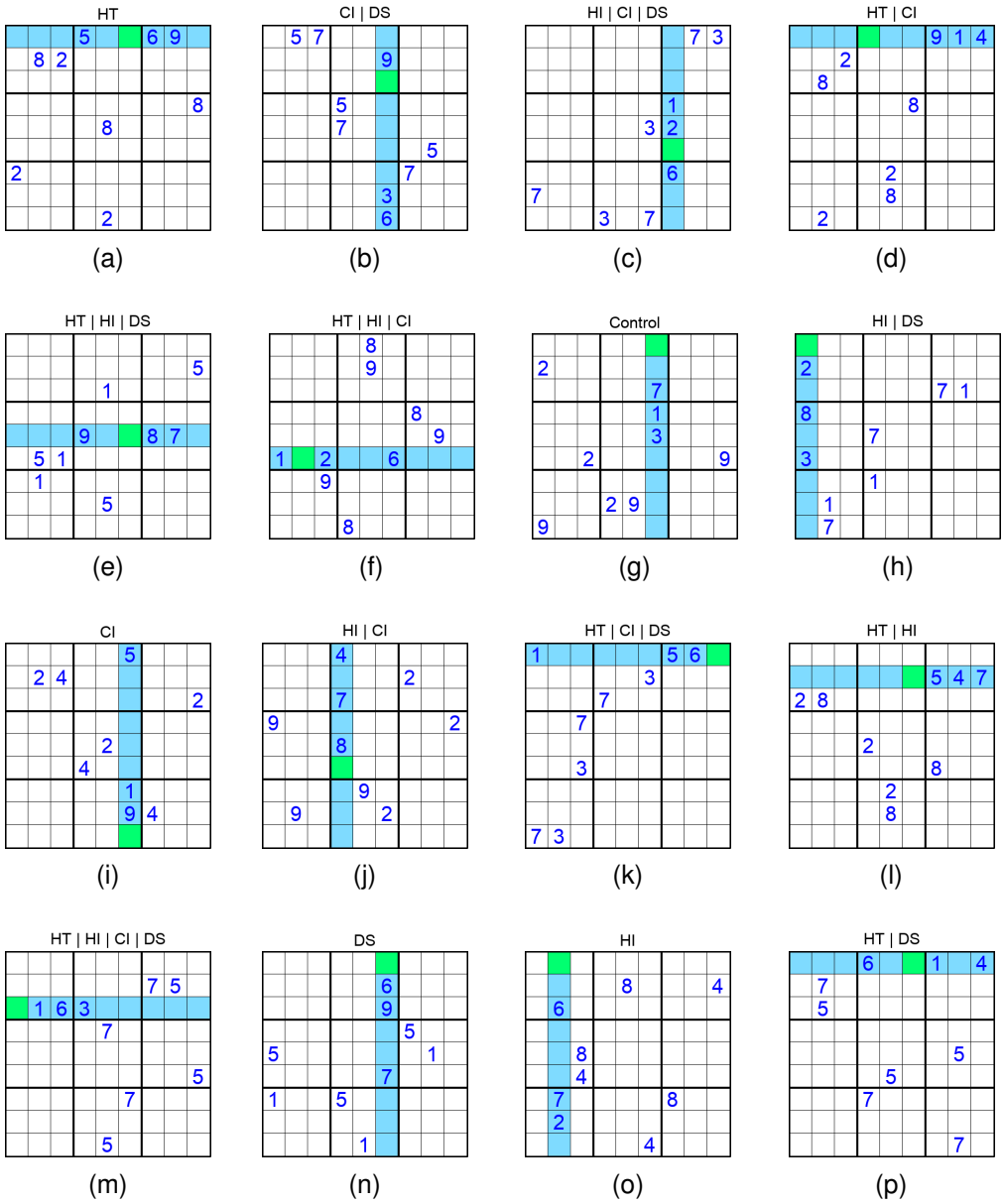

Figure 4: First 2 sets of Test Phase puzzles. ht, hi, ci, and ds indicate house type, house index, cell index, and digit set conditions respectively, signifying that the puzzle varies from the tutorial/Practice Phase/control puzzles in that feature dimension

(b) I arbitrarily chose between the two candidates because they seemed equally promising to consider.

If the participant selects the second option, they would skip forward to question 10.

4. Free response question: "What did you notice in the puzzle that initially made one candidate seem more likely to be correct than the other?"
5. Grid selection question: "Please select the cell(s) that initially made one candidate seem more likely to be correct than the other." Participants could respond by selecting one or more cells in the grid.
6. Free response question: "Please explain how the cell(s) you selected initially made one seem more likely to be correct than the other."
7. Single selection question: "After you selected a candidate to consider, did you check further to determine whether that candidate was actually correct or not?"
   (a) Yes, I checked to see whether the candidate was actually correct.
   (b) No, I just submitted my original guess without checking any further.

   If the participant selects the second option, they would skip forward to question 10.
8. Free response question: "What did you do to determine if that candidate was actually correct?"
9. Single selection question: "Which of the following best describes the way you determined whether or not the candidate was actually the correct answer?"
   (a) I looked at other numbers in the puzzle until I noticed something that helped me decide whether or not the candidate was correct.
   (b) I checked whether the candidate I chose could go in any of the empty blue cells in the row/column.
10. Free response question: "Please provide any additional information or clarifications to any of your previous responses so that we can most accurately understand as best we can how you solved this puzzle."

2.5.2 Demographics and Mathematical Background

We will also ask participants questions regarding their demographics and mathematical background.

1. Number response question: "What is your age?"
2. Single selection question: "What is your gender?"
   (a) Male
   (b) Female
   (c) Other
   (d) Prefer not to say
3. Single selection question: "What is your highest level of education (including currently pursuing)?"
   (a) Did not graduate high school
   (b) High school graduate, diploma or equivalent
   (c) Associate degree
   (d) Bachelor's degree
   (e) Master's degree
   (f) Professional degree (e.g. M.D., J.D.)
   (g) Doctoral degree
4. Single selection question: "Degree status"
   (a) Currently pursuing
   (b) Completed
5. Multiple selection question: "Which of the following mathematics topics have you taken a course in? Select all that apply."

(a) High school algebra
(b) High school geometry
(c) Trigonometric functions
(d) Single-variable calculus
(e) Multi-variable calculus
(f) Linear algebra
(g) Probability & statistics
(h) Discrete mathematics
(i) Formal logic

3 Subject Recruitment and Selection

All participants will be recruited through Amazon Mechanical Turk and must United States residents over 18 years old. We will use two criteria below to select participants to include in the experiment and analyses. We plan to run the experiment until we have data from at least 75 participants available for analyses.

3.1 Pre-Experiment Diagnostics

Seeking to only collect data from subjects who have never solved a Sudoku puzzle prior to the experiment, subjects would first be asked to solve a 4-by-4 Sudoku puzzle (**Figure 5**) without any description for the task except that it is a variant of Sudoku. Participants will be offered monetary incentives for correctness and be provided no additional information on the purpose of the puzzle. They will also be given the option to skip this puzzle and even encouraged to do so if they are not already familiar with Sudoku. Following the puzzle, subjects will be asked about their prior experience, including approximately how many Sudoku puzzles they have successfully completed before. Participants that either solved the diagnostic puzzle and/or responded that they have successfully solved at least one Sudoku puzzle in the past will be excluded from the remainder of the study.

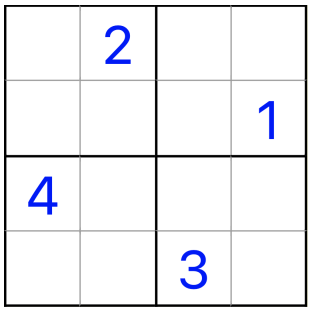

Figure 5: Diagnostic Test

3.2 Technique Proficiency

Once the data has been collected, we will partition the data into two groups based on the participants' proficiencies in using the Hidden Singles technique in Phase 1. First, we will create a mixed effects logistic regression model that predicts the likelihood that a participant correctly solved a puzzle. We specify this model as a logistic regression formulated as $P(\text{correct}_{t,s}) \sim \log t + (\log t|s)$ where t is the trial number, s is the subject, and $(\log t|s)$ is a random effect for each subject. If $P(correct_{T,s}) < 0.8$ where T is the last trial in Phase 1, we would assign the participant to the *non-solver* group. Those meeting or exceeding this threshold would be assigned to the *solver* group. Only the participants from the solver group would be used for the regression analyses.

Table 1: Statistics by group. Top 2 rows aggregate data across all participants in the solver and non-solver groups. Bottom 4 rows further split the groups by whether or not the participants solved at least 80% of the puzzles in Phase 2. "P1 Score" indicates average predicted accuracy for final trial of Phase 1.

Group	P2 Acc >80%	N	P1 Score	P1 Accuracy	P2 Accuracy
Solvers	N/A	63	0.909	0.880	0.924
Non-solvers	N/A	91	0.570	0.503	0.564
Solvers	Yes	57	0.914	0.889	0.954
Solvers	No	6	0.857	0.792	0.646
Non-solvers	Yes	12	0.716	0.629	0.896
Non-solvers	No	79	0.548	0.484	0.514

3.3 Dataset

In our pilot iterations, 821 people entered the study among which 645 were removed based on the diagnostic results, leaving 154 participants who completed the full course of the experiment. Among them, 63 were placed into the *solver* group and 92 were placed into the *non-solver* group. Based on these results, we will first deploy 800 instances of the experiment to generate the prediction model and partition the subjects. We will then iteratively deploy 50 instances of the experiment at a time, retrain the model, and repartition the dataset until we have at least 75 participants in the solver group.

See **Table 1** and **Figure 6** for group assignment and statistics for pilot data.

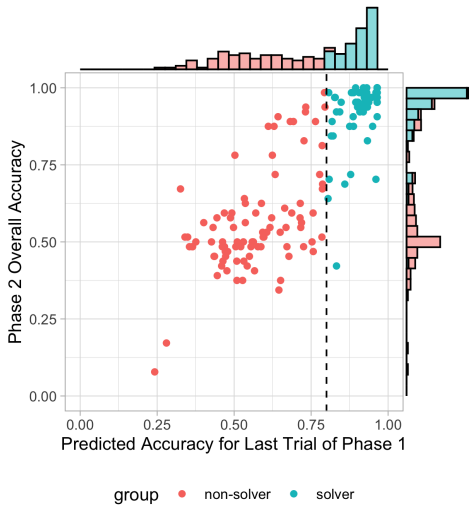

Figure 6: Exclusion criteria results.

4 Analyses

We will explore whether different features have an effect on accuracy and time to solve the puzzles through Bayesian regressions to compute 95% credible intervals for effect sizes. Again, we note that only participants in the *solver* group would be included in these analyses. Models for accuracy use logistic regressions and models for response time use log(duration). We only include trials that are correctly solved for response time models. For each regression, we include $(1|s)$ as a random intercept for each subject and $\log t + (\log t|s)$ as control measures for practice. We also use the following models as baseline comparisons for null effects:

$$P(\text{correct}_{t,s}) \sim \log t + (1 + \log t|s)$$

$$\log(\text{duration}_{t,s}) \sim \log t + (1 + \log t|s)$$

We hypothesize that even when controlling for the practice effect, there may be a clear initial main effect when the participants are first presented with the new conditions in Phase 2. Therefore, we plan to run 3 separate regressions for each analysis we conduct: 1. using data from the first 16 trials (2 sets) 2. using data from the last 48 trials (6 sets) 3. using the full data set. Only analyses of the full dataset will include interaction terms for practice.

4.1 Change in House Type

To investigate whether a change in house type has an effect, we add a main effect term HT and an interaction term with practice to the baseline models.

$$P(\text{correct}_{t,s}) \sim HT + \log t + HT * \log t + (1 + \log t|s)$$
$$\log(\text{duration}_{t,s}) \sim HT + \log t + HT * \log t + (1 + \log t|s)$$

4.1.1 Tutorial House Type

We ask whether learning and practicing the Hidden Singles technique using rows or columns makes a difference. Specifically, we add a main effect term C to indicate if the participant's tutorial used columns and interactions with existing fixed effects from the main HT regressions.

$$P(\text{correct}_{t,s}) \sim HT + C + \log t + HT * C + HT * \log t + C * \log t + HT * C * \log t + (1 + \log t|s)$$

$$P(\text{duration}_{t,s}) \sim HT + C + \log t + HT * C + HT * \log t + C * \log t + HT * C * \log t + (1 + \log t|s)$$

4.2 Change in House and Cell Indices

Here, we investigate whether a change in the position of the house and cell indices have any effects. We hypothesize that when the house type changes, the puzzle may appear sufficiently spatially distinct such that a change in the target cell position may not make a difference. To explore this possibility, we conduct two parallel sets of analyses, once using trials where the house type is the same as the tutorial and once using trials where the house type is different. We also augment the baseline model to include *HI*, *CI*, and their interactions with each other and with practice.

$$P(\text{correct}_{t,s}) \sim HI + CI + HI * CI + HI * \log t + CI * \log t + HI * CI * \log t + (1 + \log t|s)$$

$$P(\text{duration}_{t,s}) \sim HI + CI + HI * CI + HI * \log t + CI * \log t + HI * CI * \log t + (1 + \log t|s)$$

4.3 Change in Digit Set

To investigate whether a change in digit set for the target and distractor has an effect, we add a main effect term DS to indicate that the digit set has been swapped and an interaction term with practice to the baseline models:

$$P(\text{correct}_{t,s}) \sim DS + \log t + DS * \log t + (1 + \log t|s)$$
$$\log(\text{duration}_{t,s}) \sim DS + \log t + DS * \log t + (1 + \log t|s)$$

5 Pilot Results

We provide example results of our planned analyses using our pilot data. We note that parts of the tutorial in our pilot study differed from what is described in this document and Phase 1 had 20 puzzles instead of 25.

5.1 Solver vs. Non-Solver Comparison

After filtering out participants with prior exposure to Sudoku based on the diagnostic test and survey, our dataset contained data from 157 participants. Applying the exclusion criterion based on Phase 1 results, we narrowed our final pool of 65 participants for our pilot analyses. Figure 6 shows the relationship between the predicted accuracy at the end of Phase 1 to the overall accuracy in Phase 2.

5.2 Regressions

See **Section 7 Tables** for regression results. See **Section 8 Scatterplot Figures** for plotted pilot data.

6 Exploratory Analyses

6.1 Unified Model of Practice

Our pilot analyses suggest that effects of house type and position are strongest at the start of Phase 2, but steadily diminish until performances on control and treated trials are nearly indistinguishable. In our confirmatory analyses described above, we address this issue by splitting our data to early and late trials for improved statistical sensitivity to the effects of interest. Therefore, in our exploratory analyses, we aim to create a model that provides a unified framework for participants' accuracy and response times across all trials.

6.2 Error Analysis

All interactions with the application were recorded and incorrect responses are available for further analysis. However, the data may not be more informative than confirming that participants with $\sim 50\%$ accuracy selected either the target or distractor digit at random.

7 Tables

Table 2: Coefficients for effect of house type. Coefficients with credible intervals that exclude 0 in bold.

DV	Subset	Term	Estimate	CI Lower	CI Upper
accuracy	allsets	HT	-0.727	-1.494	0.037
accuracy	allsets	HT * log(trial)	0.089	-0.081	0.256
accuracy	**first2sets**	**HT**	**-0.742**	**-1.212**	**-0.278**
accuracy	last6sets	HT	-0.175	-0.499	0.147
duration	**allsets**	**HT**	**0.208**	**0.077**	**0.343**
duration	**allsets**	**HT * log(trial)**	**-0.031**	**-0.059**	**-0.003**
duration	**first2sets**	**HT**	**0.123**	**0.049**	**0.198**
duration	**last6sets**	**HT**	**0.044**	**0.005**	**0.084**

Table 3: Coefficients for effect of house type and starting house type. C indicates participants that were assigned columns in the tutorial and Phase 1. Coefficients with credible intervals that exclude 0 in bold.

DV	Subset	Term	Estimate	CI Lower	CI Upper
accuracy	allsets	C	0.447	-0.983	1.889
accuracy	allsets	HT	-0.640	-1.716	0.365
accuracy	allsets	HT * C	-0.249	-1.869	1.380
accuracy	allsets	C * log(trial)	-0.040	-0.340	0.275
accuracy	allsets	HT * log(trial)	0.068	-0.159	0.306
accuracy	allsets	HT * C * log(trial)	0.058	-0.303	0.415
accuracy	first2sets	C	0.253	-0.798	1.308
accuracy	**first2sets**	**HT**	**-0.827**	**-1.450**	**-0.246**
accuracy	first2sets	HT * C	0.193	-0.726	1.112
accuracy	last6sets	C	0.174	-0.814	1.136
accuracy	last6sets	HT	-0.141	-0.577	0.288
accuracy	last6sets	HT * C	-0.084	-0.752	0.577
duration	allsets	C	0.007	-0.365	0.385
duration	**allsets**	**HT**	**0.292**	**0.116**	**0.480**
duration	allsets	HT * C	-0.188	-0.458	0.084
duration	allsets	C * log(trial)	-5.00E-04	-0.067	0.065
duration	**allsets**	**HT * log(trial)**	**-0.043**	**-0.082**	**-0.006**
duration	allsets	HT * C * log(trial)	0.027	-0.029	0.082
duration	first2sets	C	-0.076	-0.393	0.247
duration	**first2sets**	**HT**	**0.169**	**0.070**	**0.271**
duration	first2sets	HT * C	-0.101	-0.247	0.051
duration	last6sets	C	0.006	-0.289	0.293
duration	**last6sets**	**HT**	**0.064**	**0.013**	**0.114**
duration	last6sets	HT * C	-0.041	-0.114	0.032

Table 4: Coefficients for effect of target cell position (house type unchanged). Coefficients with credible intervals that exclude 0 in bold.

DV	Subset	Term	Estimate	CI Lower	CI Upper
accuracy	**allsets**	**HI**	**-1.913**	**-3.623**	**-0.247**
accuracy	allsets	CI	0.301	-1.703	2.325
accuracy	allsets	HI * CI	0.615	-2.084	3.263
accuracy	**allsets**	**HI * log(trial)**	**0.425**	**0.054**	**0.796**
accuracy	allsets	CI * log(trial)	-0.020	-0.455	0.405
accuracy	allsets	HI * CI * log(trial)	-0.212	-0.793	0.372
accuracy	first2sets	HI	-0.027	-0.212	0.161
accuracy	first2sets	CI	-0.075	-0.262	0.109
accuracy	**first2sets**	**HI * CI**	**0.316**	**0.045**	**0.593**
accuracy	last6sets	HI	0.022	-0.017	0.061
accuracy	last6sets	CI	0.024	-0.014	0.063
accuracy	**last6sets**	**HI * CI**	**-0.065**	**-0.121**	**-0.009**
duration	**allsets**	**HI**	**-1.560**	**-2.835**	**-0.412**
duration	allsets	CI	-0.106	-1.473	1.257
duration	allsets	HI * CI	0.338	-1.348	2.015
duration	allsets	HI * log(trial)	0.065	-0.038	0.169
duration	allsets	CI * log(trial)	-0.018	-0.120	0.079
duration	allsets	HI * CI * log(trial)	0.124	-0.023	0.271
duration	first2sets	HI	0.405	-0.227	1.066
duration	first2sets	CI	0.214	-0.430	0.841
duration	first2sets	HI * CI	-0.421	-1.410	0.578
duration	**last6sets**	**HI**	**0.085**	**0.035**	**0.135**
duration	**last6sets**	**CI**	**0.058**	**0.007**	**0.108**
duration	last6sets	HI * CI	-0.022	-0.099	0.054

Table 5: Coefficients for effect of target cell position (house type changed). Coefficients with credible intervals that exclude 0 in bold.

DV	Subset	Term	Estimate	CI Lower	CI Upper
accuracy	allsets	HI	0.727	-0.608	2.040
accuracy	allsets	CI	0.888	-0.466	2.264
accuracy	allsets	HI * CI	-0.253	-2.328	1.854
accuracy	allsets	HI * log(trial)	-0.126	-0.422	0.176
accuracy	allsets	CI * log(trial)	-0.124	-0.434	0.178
accuracy	allsets	HI * CI * log(trial)	-0.025	-0.489	0.443
accuracy	first2sets	HI	0.403	-0.385	1.224
accuracy	first2sets	CI	0.501	-0.294	1.308
accuracy	first2sets	HI * CI	-0.320	-1.548	0.907
accuracy	last6sets	HI	0.085	-0.488	0.662
accuracy	last6sets	CI	0.270	-0.350	0.892
accuracy	last6sets	HI * CI	-0.380	-1.264	0.516
duration	allsets	HI	-0.142	-0.399	0.145
duration	allsets	CI	-0.152	-0.412	0.118
duration	allsets	HI * CI	0.244	-0.162	0.642
duration	allsets	HI * log(trial)	0.040	-0.018	0.093
duration	allsets	CI * log(trial)	0.033	-0.022	0.086
duration	allsets	HI * CI * log(trial)	-0.040	-0.120	0.044
duration	first2sets	HI	0.042	-0.098	0.183
duration	first2sets	CI	-0.043	-0.179	0.097
duration	first2sets	HI * CI	0.072	-0.141	0.278
duration	last6sets	HI	0.052	-0.022	0.122
duration	last6sets	CI	0.024	-0.048	0.096
duration	last6sets	HI * CI	0.040	-0.069	0.151

Table 6: Coefficients for effect of digit set. Coefficients with credible intervals that exclude 0 in bold.

DV	Subset	Term	Estimate	CI Lower	CI Upper
accuracy	allsets	DS	-0.303	-1.158	0.559
accuracy	allsets	DS * log(trial)	0.109	-0.075	0.296
accuracy	first2sets	DS	0.050	-0.382	0.487
accuracy	last6sets	DS	0.231	-0.075	0.539
duration	allsets	DS	-0.040	-0.186	0.113
duration	allsets	DS * log(trial)	0.012	-0.018	0.041
duration	first2sets	DS	-0.026	-0.102	0.047
duration	last6sets	DS	0.033	-0.004	0.071

8 Scatterplot Figures

8.1 Change in House Type

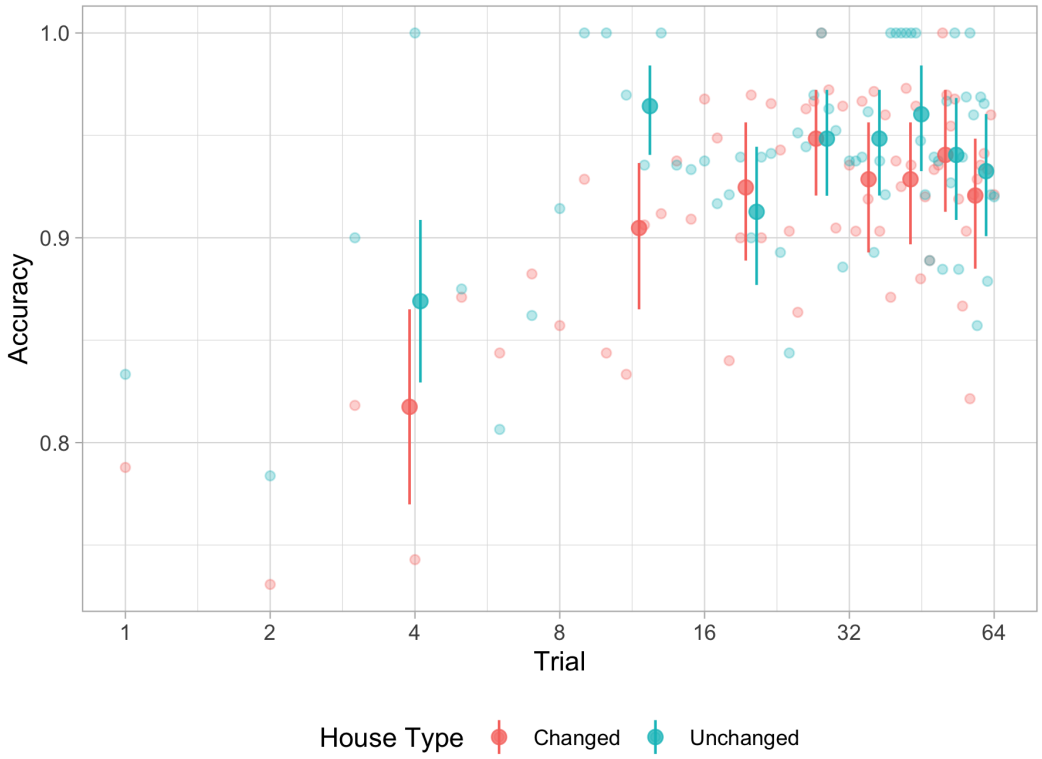

Figure 7: Effect of change in house type on accuracy. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

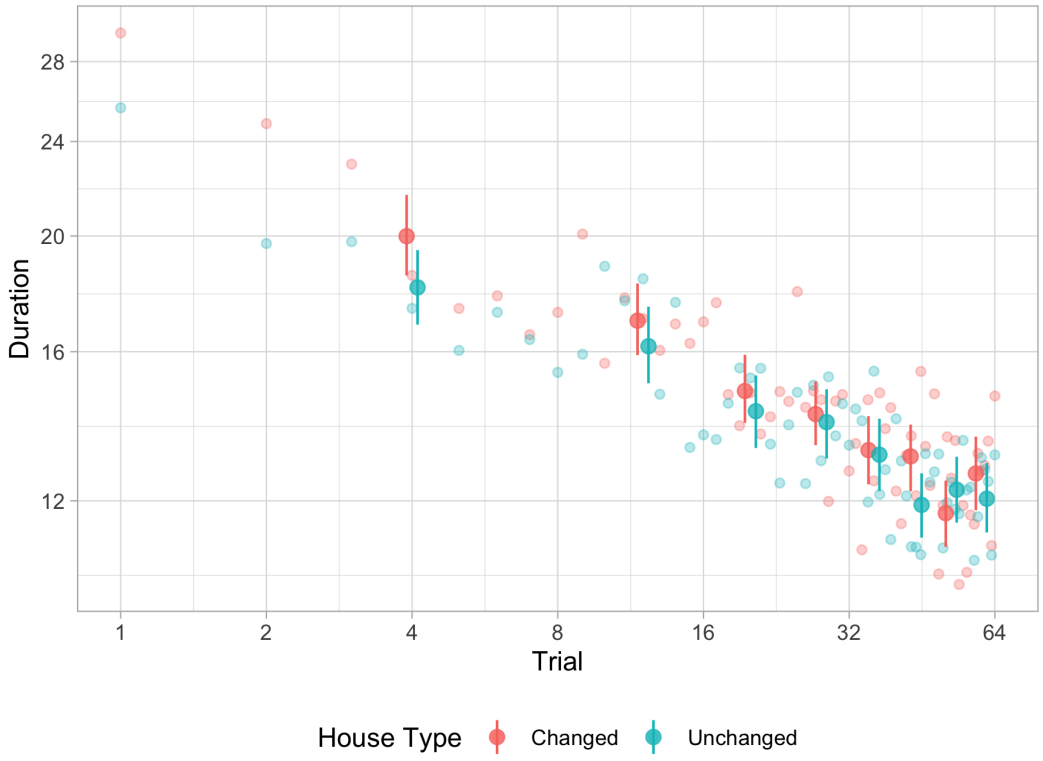

Figure 8: Effect of change in house type on response time (correct responses only). Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

8.2 Tutorial House Type

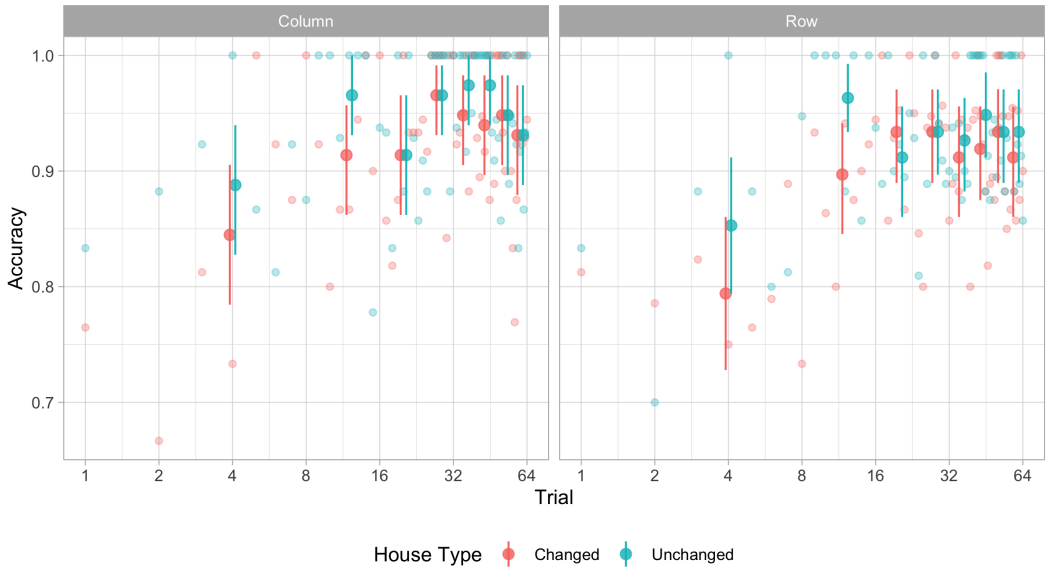

Figure 9: Effect of change in house type on accuracy. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale. Left: participants with column puzzles for tutorials. Right: participants with row puzzles for tutorials.

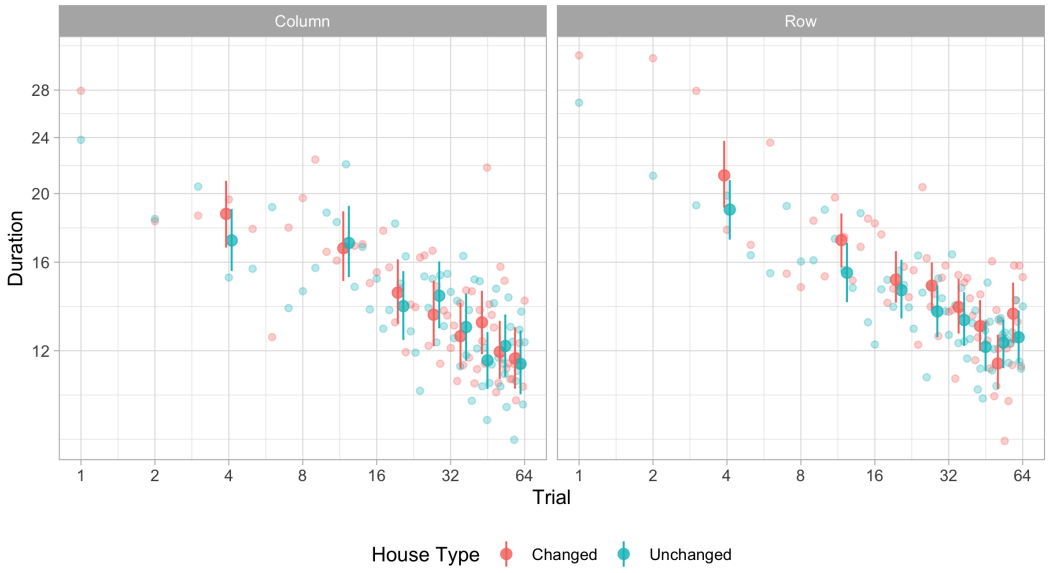

Figure 10: Effect of change in house type on response time (correct responses only). Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

8.3 Change in House and Cell Indices

8.3.1 House Type Unchanged

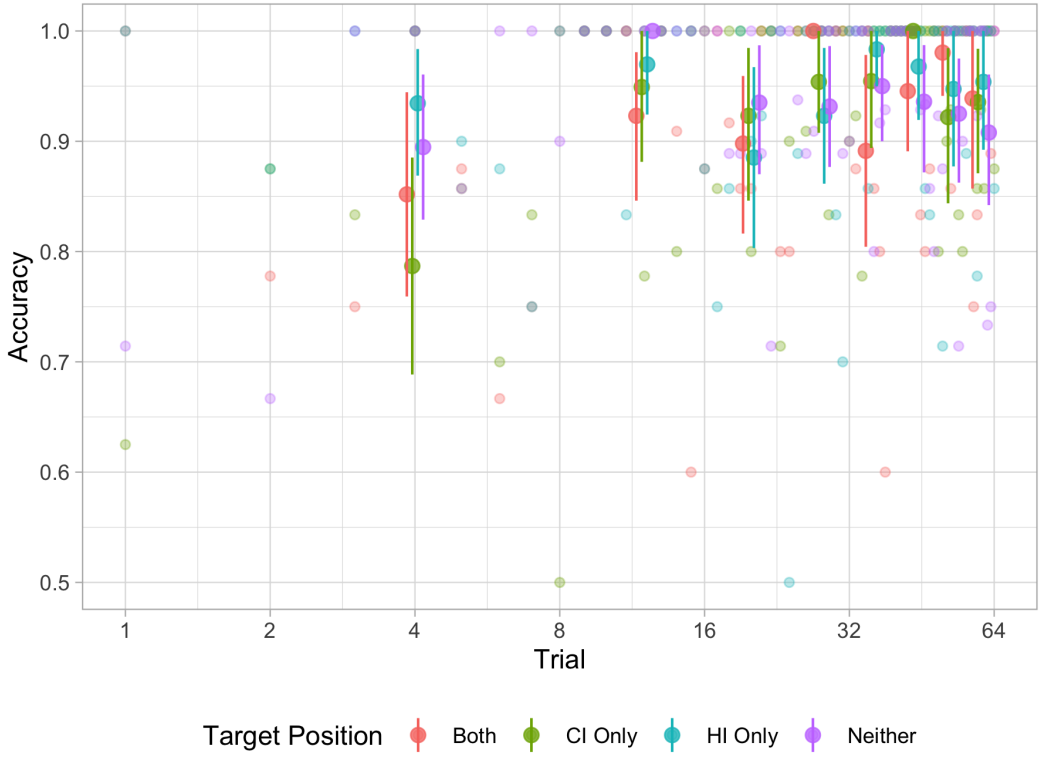

Figure 11: Effect of change in house and cell indices on accuracy when house type has been unchanged. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

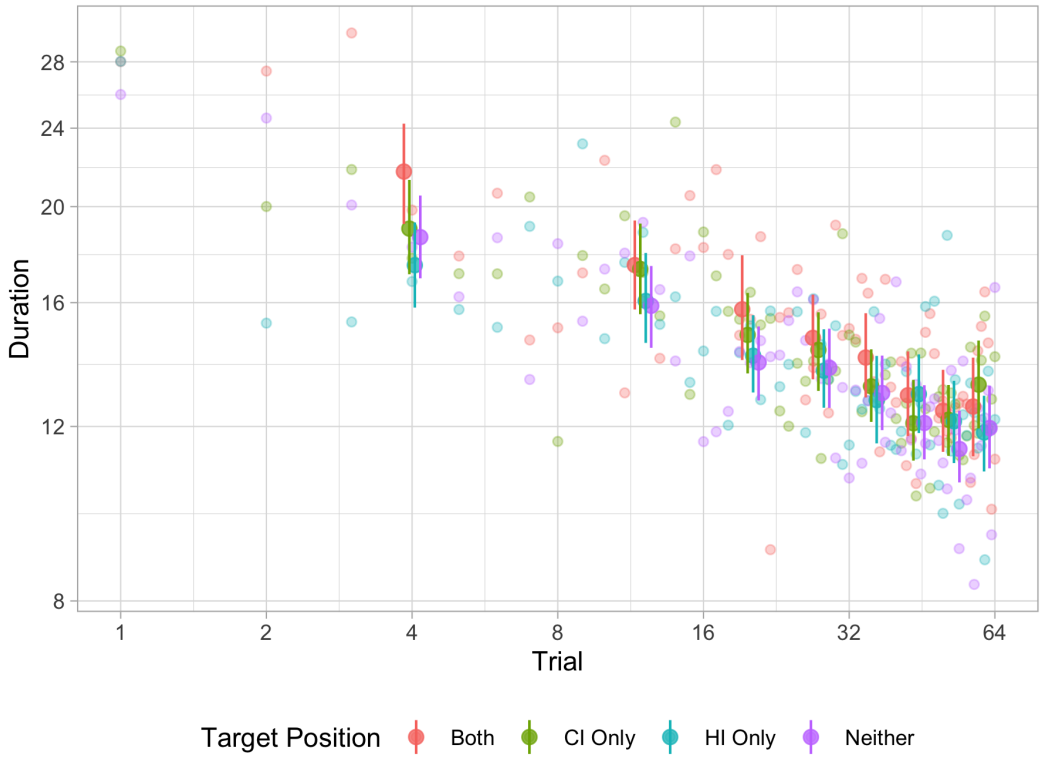

Figure 12: Effect of change in house and cell indices on response time (correct responses only) when house type has been unchanged. Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

8.3.2 House Type Changed

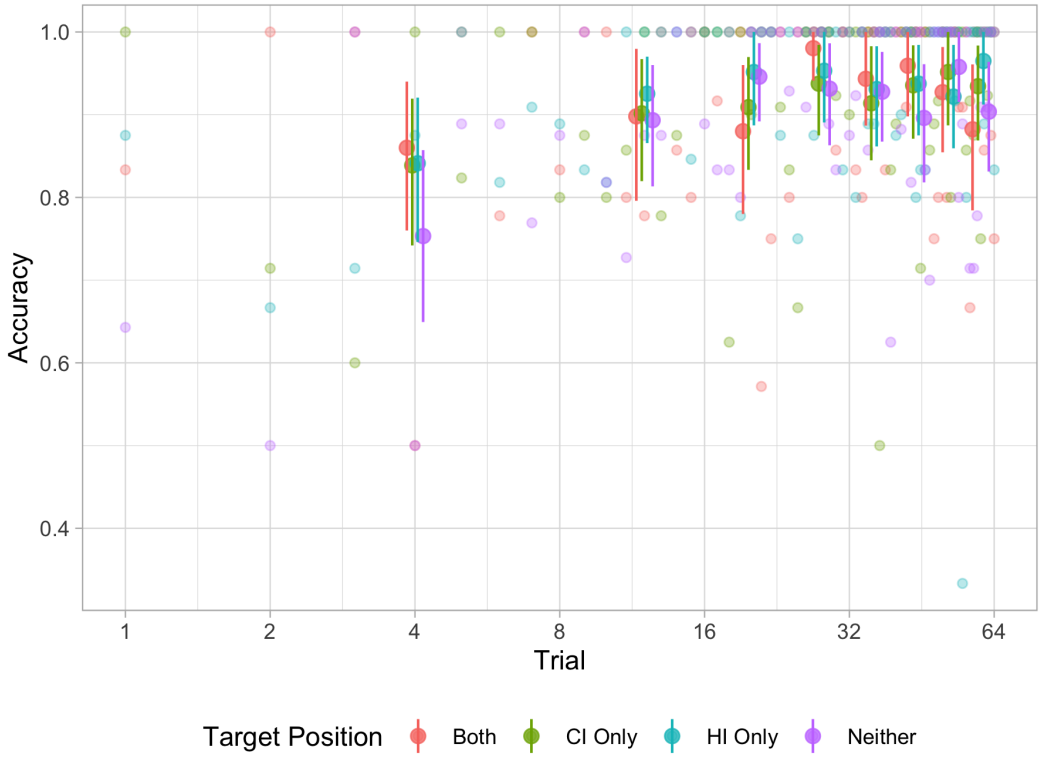

Figure 13: Effect of change in house and cell indices on accuracy when house type has been changed. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

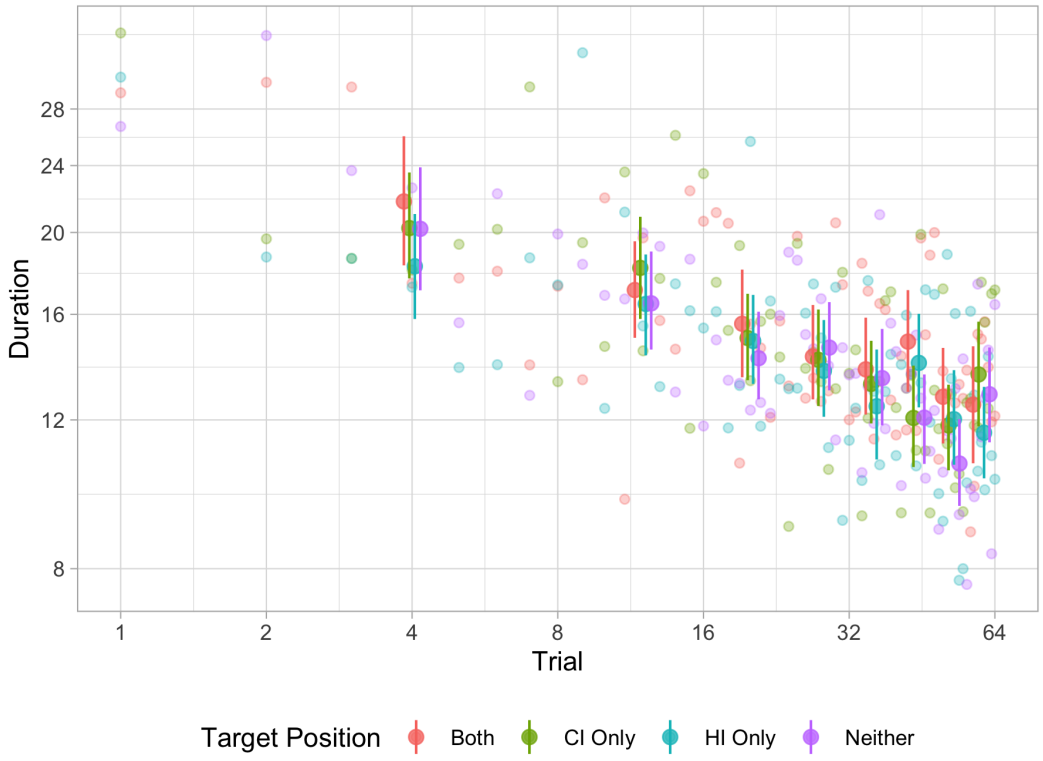

Figure 14: Effect of change in house and cell indices on response time (correct responses only) when house type has been changed. Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

8.4 Change in Digit Set

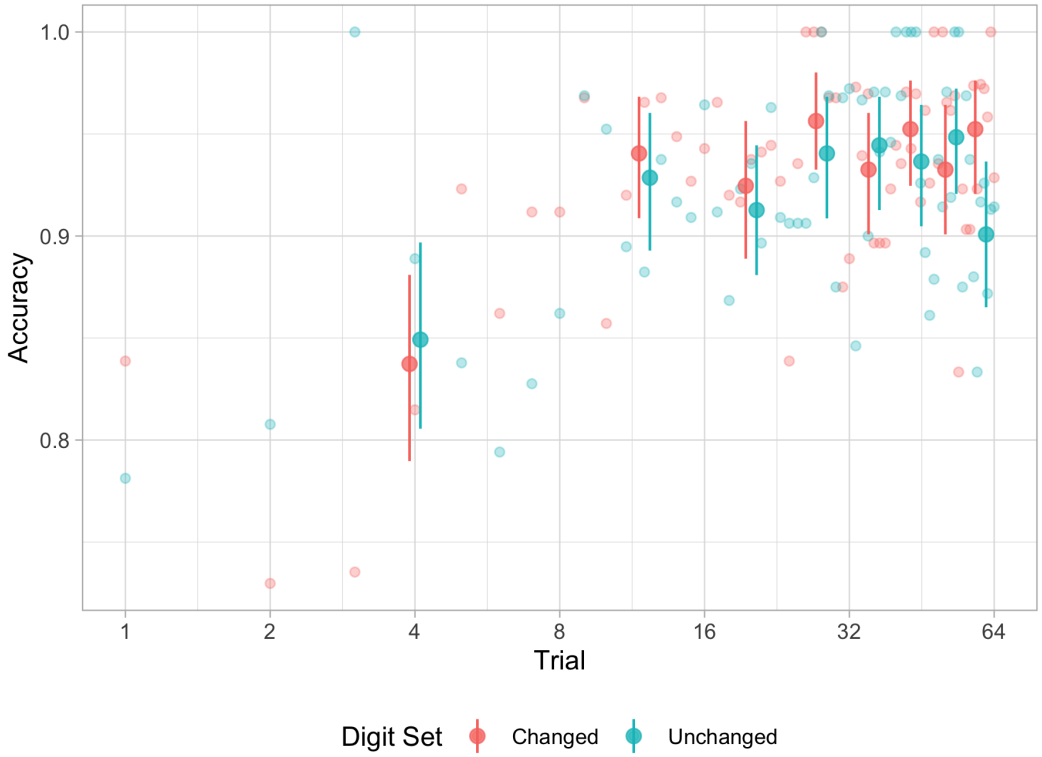

Figure 15: Effect of change in digit set on accuracy. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

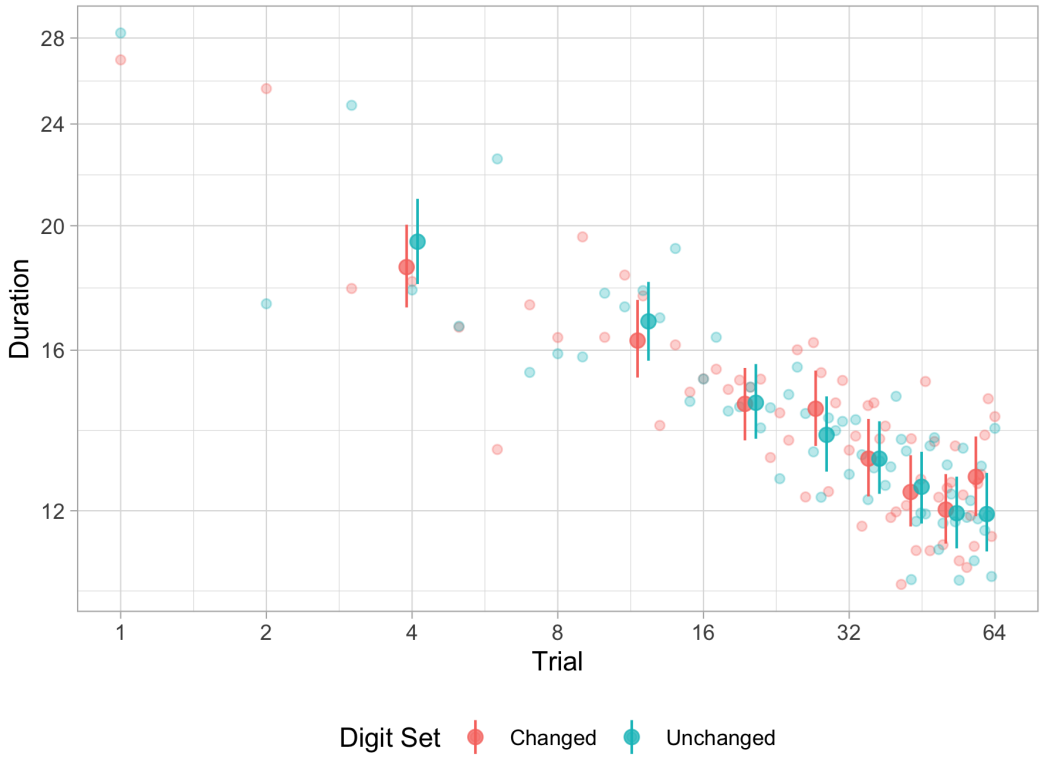

Figure 16: Effect of change in digit set on response time (correct responses only). Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

9 Regression Figures

Solid lines are predicted values of regressions trained on all 8 sets. Shaded regions indicate intervals between first and ninth deciles of actual data. Dashed lines indicate the boundaries of a 80% prediction interval.

9.1 Change in House Type

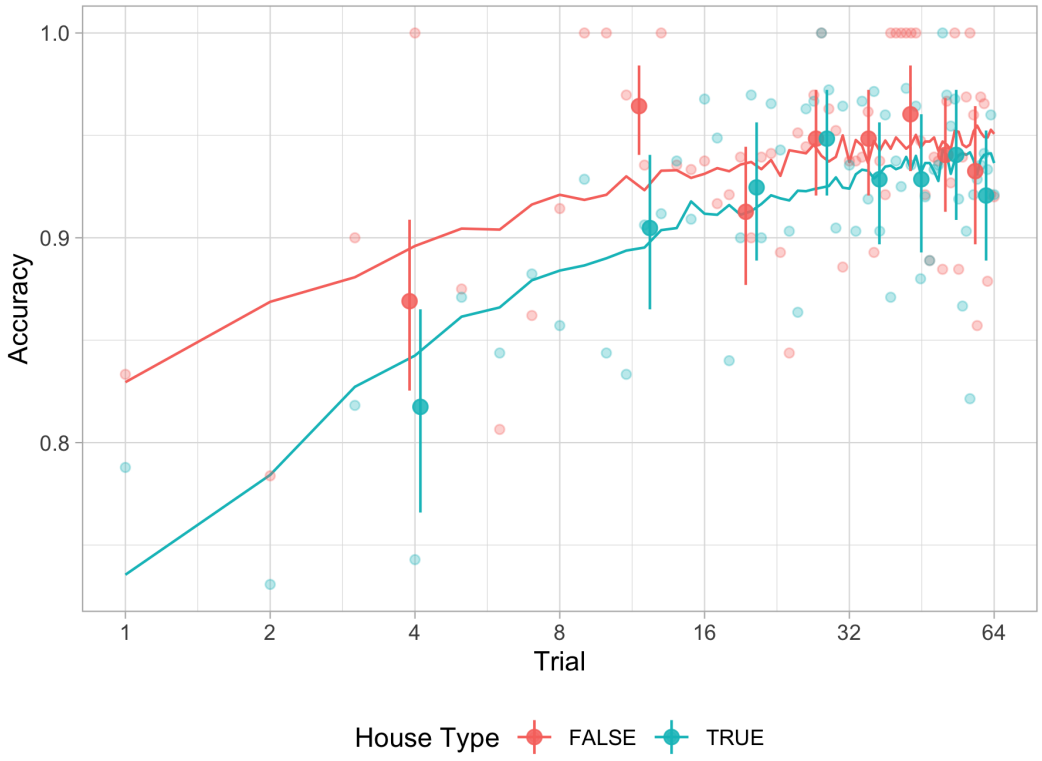

Figure 17: Effect of change in house type on accuracy. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

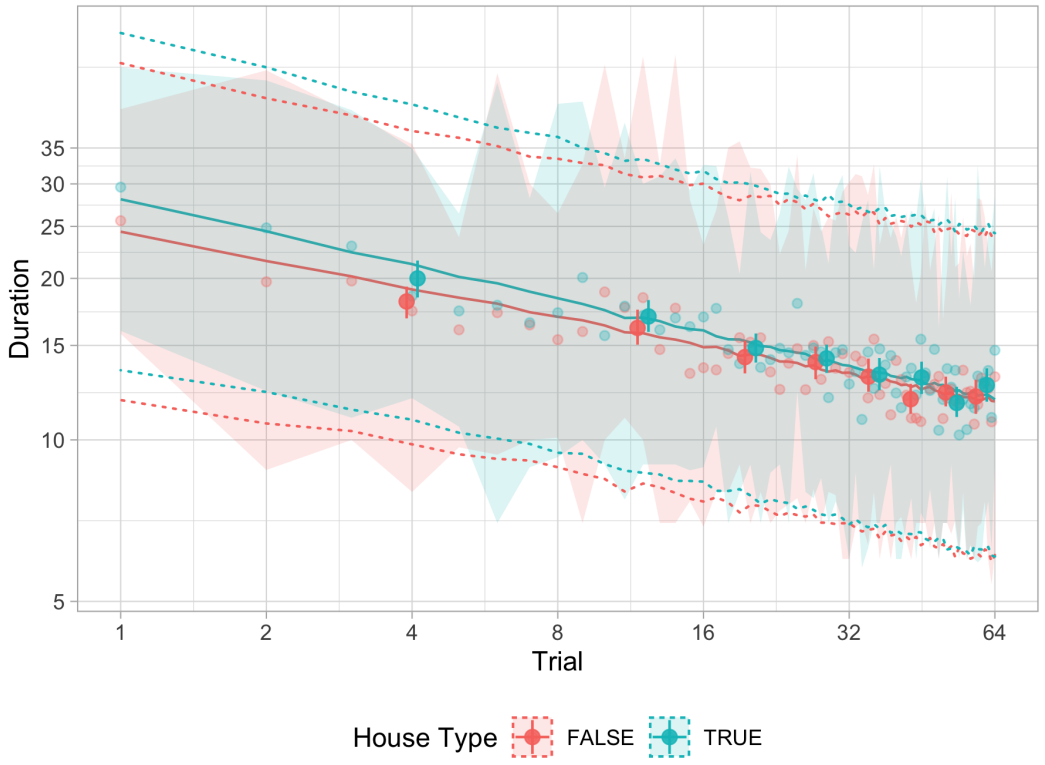

Figure 18: Effect of change in house type on response time (correct responses only). Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

9.2 Tutorial House Type

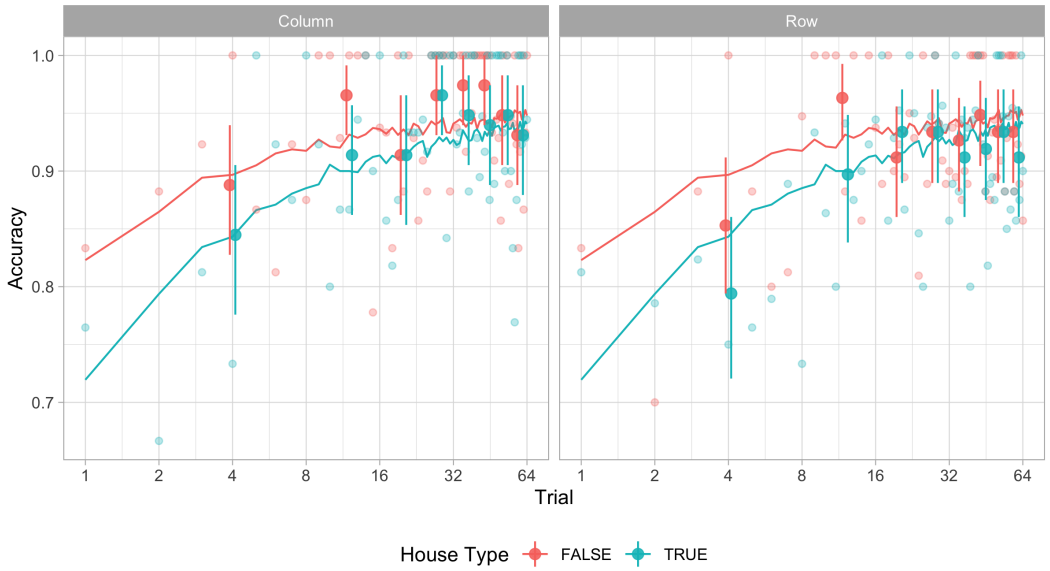

Figure 19: Effect of change in house type on accuracy. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale. Left: participants with column puzzles for tutorials. Right: participants with row puzzles for tutorials.

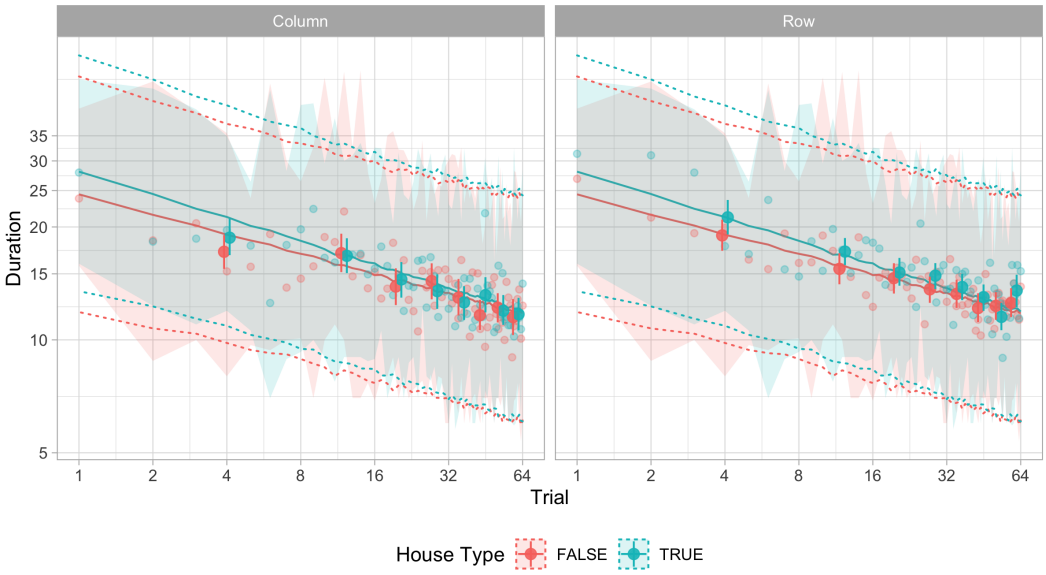

Figure 20: Effect of change in house type on response time (correct responses only). Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

9.3 Change in House and Cell Indices

9.3.1 House Type Unchanged

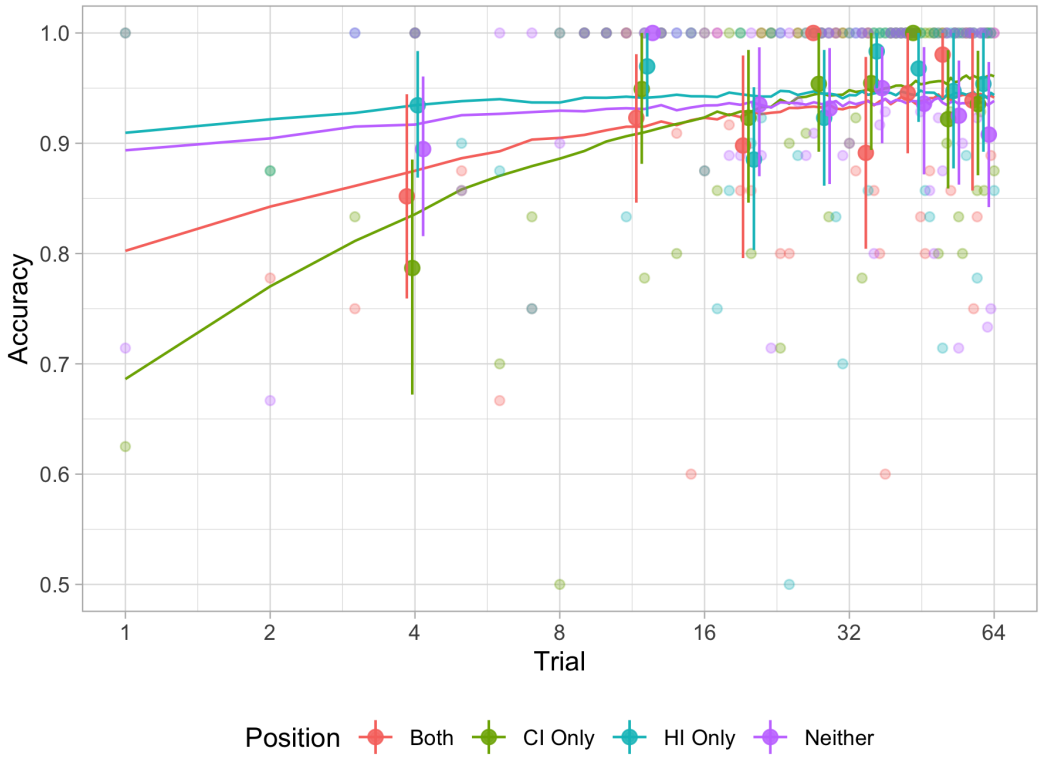

Figure 21: Effect of change in house and cell indices on accuracy when house type has been unchanged. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

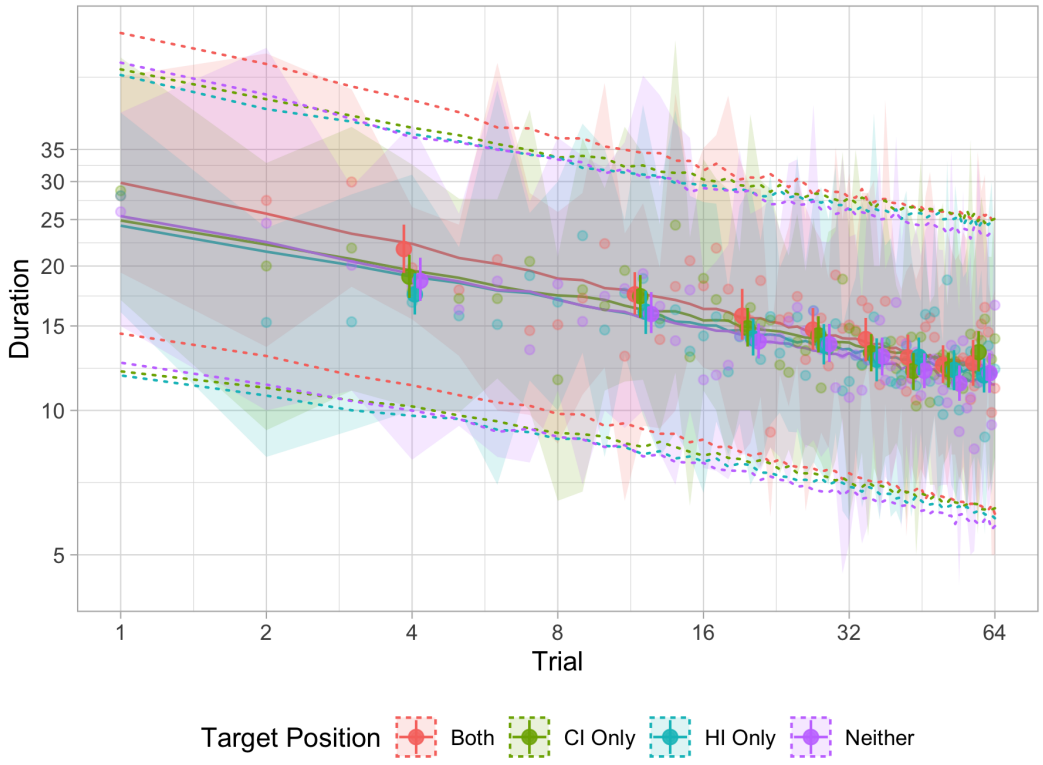

Figure 22: Effect of change in house and cell indices on response time (correct responses only) when house type has been unchanged. Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

9.3.2 House Type Changed

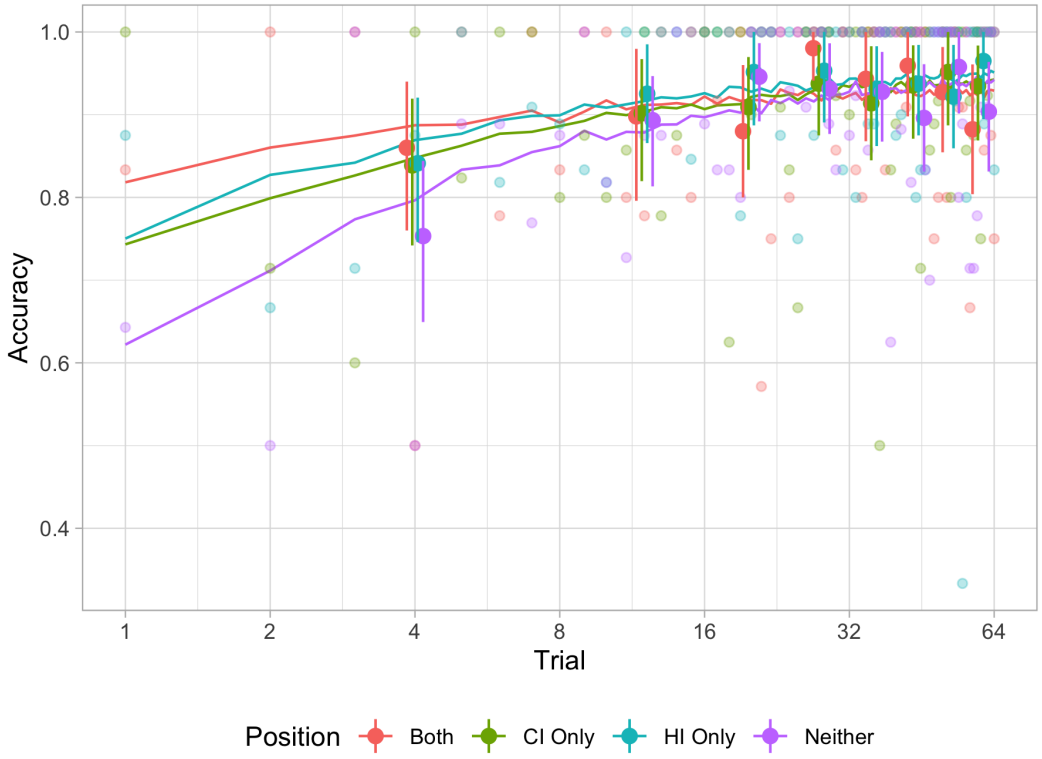

Figure 23: Effect of change in house and cell indices on accuracy when house type has been changed. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

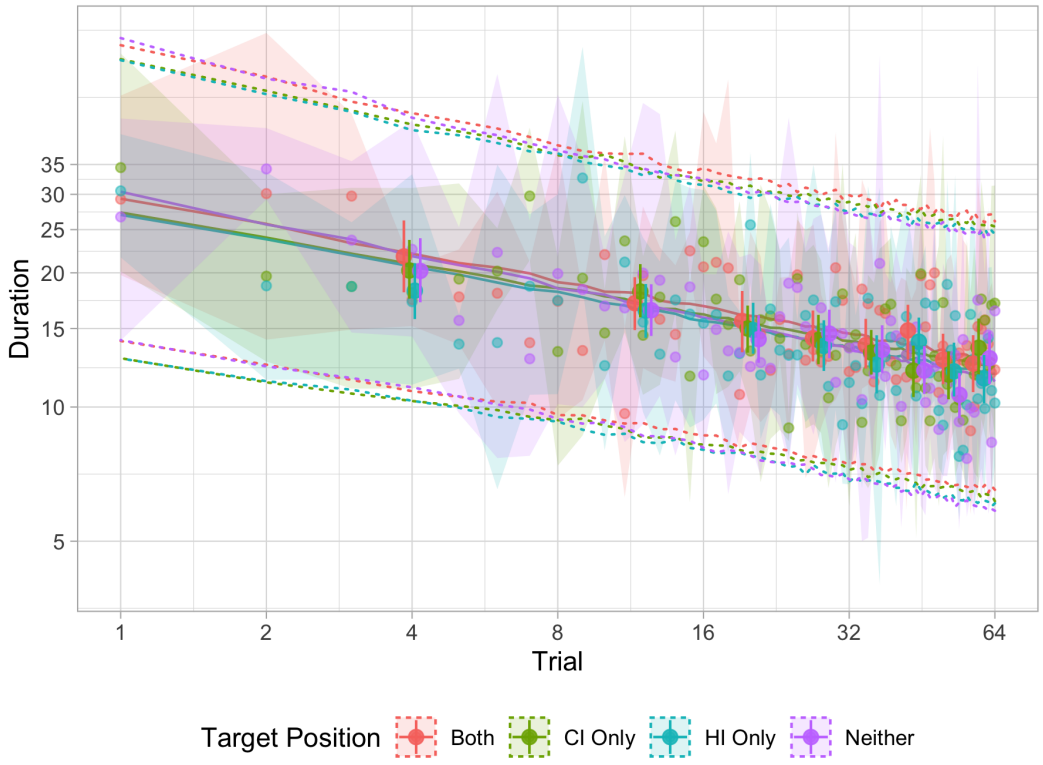

Figure 24: Effect of change in house and cell indices on response time (correct responses only) when house type has been changed. Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.

9.4 Change in Digit Set

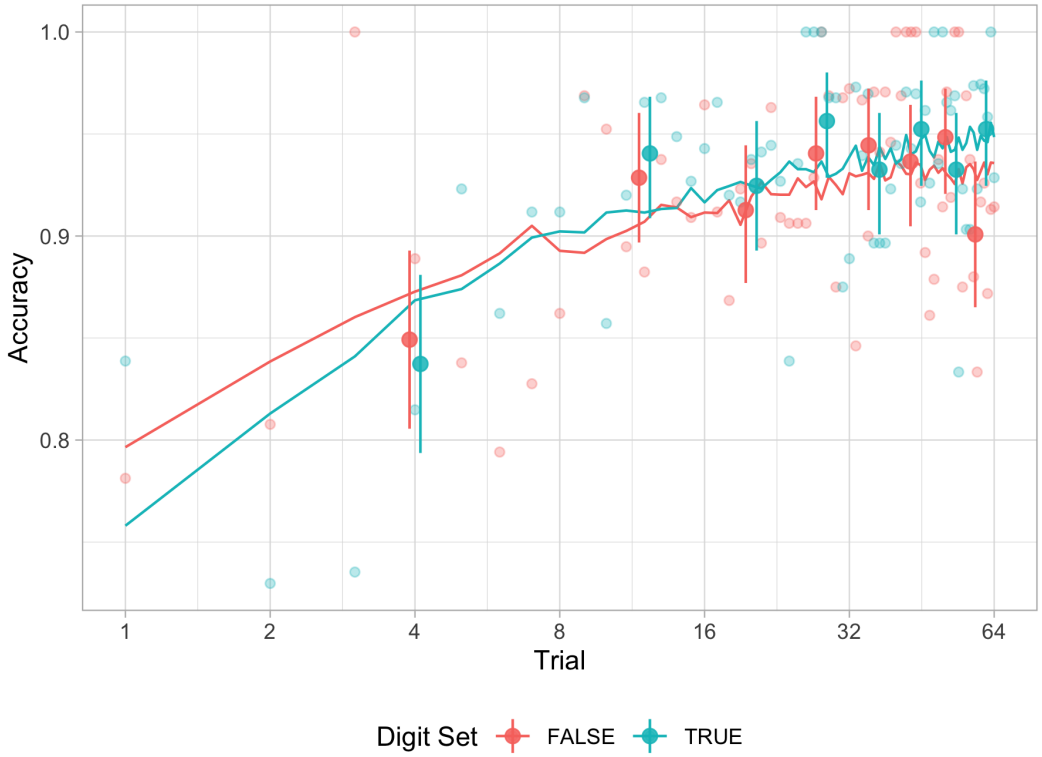

Figure 25: Effect of change in digit set on accuracy. Larger, dark points with error bars indicate mean and SE accuracy of each set (8 trials). Smaller, lighter points indicate mean accuracy of each trial. Note that trials (x-axis) are in log-2 scale.

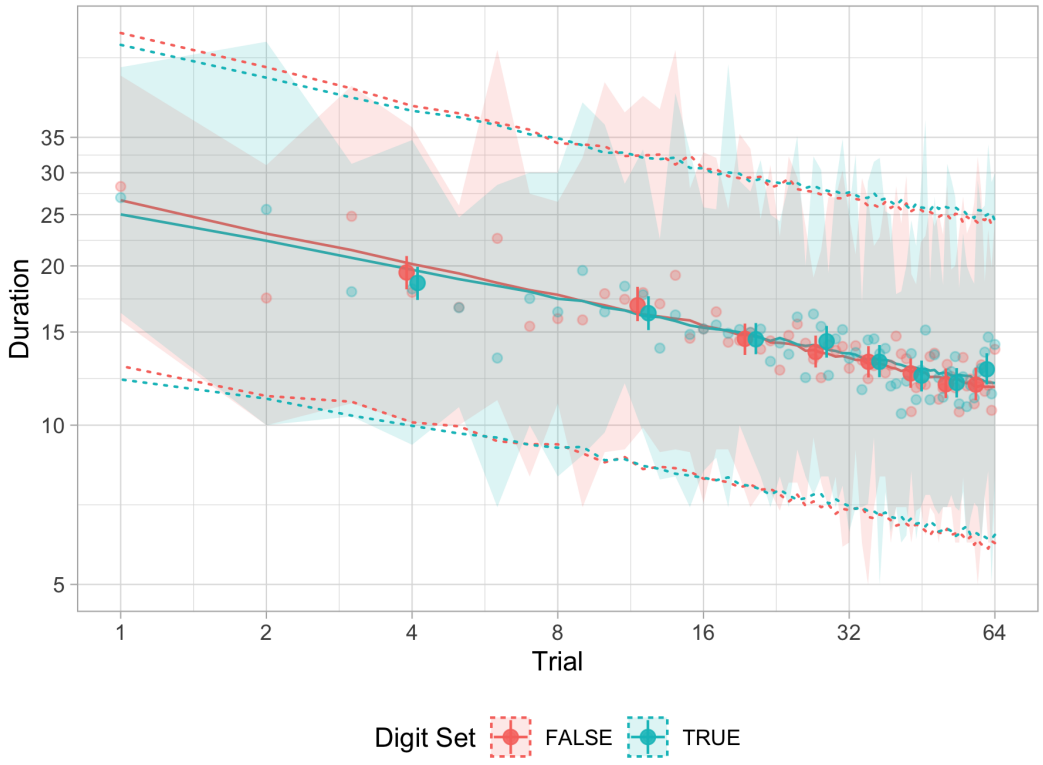

Figure 26: Effect of change in digit set on response time (correct responses only). Larger, dark points with error bars indicate mean and SE duration of each set (8 trials). Smaller, lighter points indicate mean duration of each trial. Note that seconds (y-axis) are in log-2 scale.