Supplementary Materials for

Designing Shared Information Displays for Agents of Varying Strategic Sophistication

Dongping Zhang*, Jason Hartline, and Jessica Hullman

*Corresponding author. Email: dzhang@u.northwestern.edu

1 Experimental Setup

This section of the supplementary materials focuses on the data processing pipeline and exploratory data analysis (EDA) conducted on the Chicago taxi trips dataset. Additionally, it includes examples illustrating the application of our algorithm to infer the search strategies of drivers, as outlined in the Appendix of our paper. For access to the data and code used in our study, please visit our GitHub repository at: https://github.com/dpzhang/shared-info-display.

1.1 Background

We are interested in the search-pickup dynamics of three Chicago Community Areas (CAs): North Loop, West Loop, and the Loop. Our goal is to create a counterfactual model trained on real taxi drivers' search and pickup dynamics that can support our strategic setting by its capability to predict the number of successful pickups in each action CA at 9 AM, given the total number of drivers who choose to search in each CA. As described in the paper, while we can aggregate the model's dependent variable (i.e., pickups) by directly counting the total number of 9 AM pickups in each target CA from the taxi origin-destination (OD) data, this data does not provide information on the drivers' search decisions (i.e., flows). To address this data limitation, we leverage the power of big data and design an algorithm capable of deducing individual drivers' search decisions by using their prior pickup history during AM Peaks from the past 10 days, which we refer to as the driver's search dyad.

1.2 Trip Stratification

We control the temporal variability of taxi flows by implementing a stratification scheme that groups taxi trips by start timestamps along two time dimensions: (1) days of the week (i.e., weekday and weekend) and (2) time intervals of a day. Assuming the flow dynamics are similar within each stratum, we define the pickup session of our strategic setting to be 9 AM, and hence select the strata containing 5.8 million trips from 522 weekday AM Peaks (7 AM - 10 AM) as the relevant trips to deduce each driver's 9 AM search decisions.

1.3 Identify Homogeneous Weekdays

We define two external factors that can influence taxi flows: weather and a city's special events.

Special Events: We identify a total of 37 weekdays in 2014-2015 when Chicago hosts special events, including US federal holidays. The total number of pickups, which are commonly used as proxies for demand, from these weekdays drastically differs from the rest of the weekdays. Hence, we exclude these 37 weekdays from the trip strata used to fit the models.

Weather: As described in the paper, we consider four meteorological variables: (1) apparent temperature (i.e., calculated by combining the heat index and wind chill to measure the temperature the body feels), (2) precipitation, (3) snowfall, and (4) wind speed. We define four criteria and include weekdays that have (1) apparent temperature falling between -10 °C (14 °F) and 23 °C (73 °F), (2) precipitation below "light rain" (0.25 cm or 0.1 inch) according to the scale defined by the American Meteorological Society, (3) no snowfall, and (4) wind speeds below "gentle breeze" (i.e., 10 knots) according to the Beaufort Wind Scale.

After conditioning on weather and special events, we identify 221 (42%) homogeneous weekdays from the trip strata. The main focus of our work is not to create a model that can rigorously predict the supply and demand dynamics but simply one that is trained on the real-world data and capable of providing exogenous and time-specific predictions of payoffs and evaluating players' decisions. Conditioning on weekdays allows us to (1) make our strategic settings as consistent as possible so that agents can use and improve their mental models of competitors' actions through repeated decision-making using predictions controlled by the influential external factors, (2) make the context of the strategic setting as simple as possible for explanation purposes, (3) evaluate decisions without requiring additional features to account for the external factors, and (4) facilitate the optimization of social welfare for which we need to use the model coefficients in the objective function to compute the maximal social welfare.

1.4 EDA on Pickups

We aggregate the total number of pickups originating from the three action CAs during the AM Peak by weekdays. We provide some descriptive statistics of the AM Peak pickups in Table S1 and visualize their distribution in Figure S1. The distribution of total AM Peak pickups follows a bell-shaped curve with a slight skew to the left.

Table S1: Descriptive of Total AM Peak Pickups in Action CAs by 221 Weekdays

Min.	1st. Qu.	Median	Mean	3rd Qu.	Max	Std. Dev.
4059	6264	6859	6766	7368	9068	823

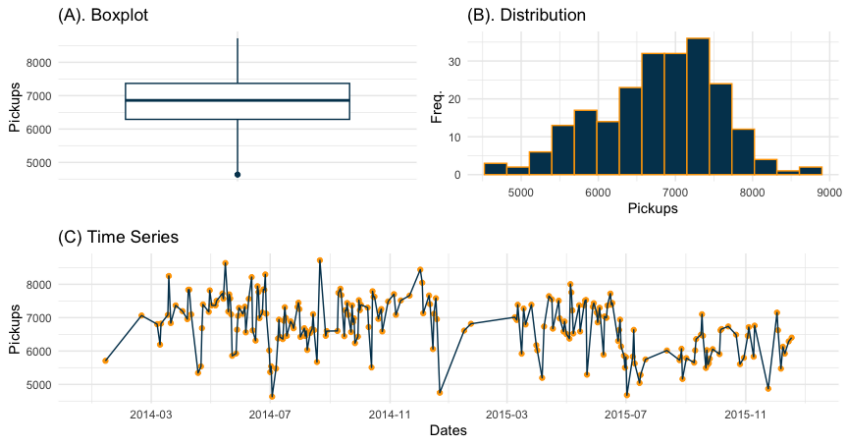

Fig. S1: AM Peak Pickups in Action CAs from 221 Weekdays

We present the time series of 9 AM pickups by CA in Figure S2. The series appear visually stationary, indicating that the conditioning was effective.

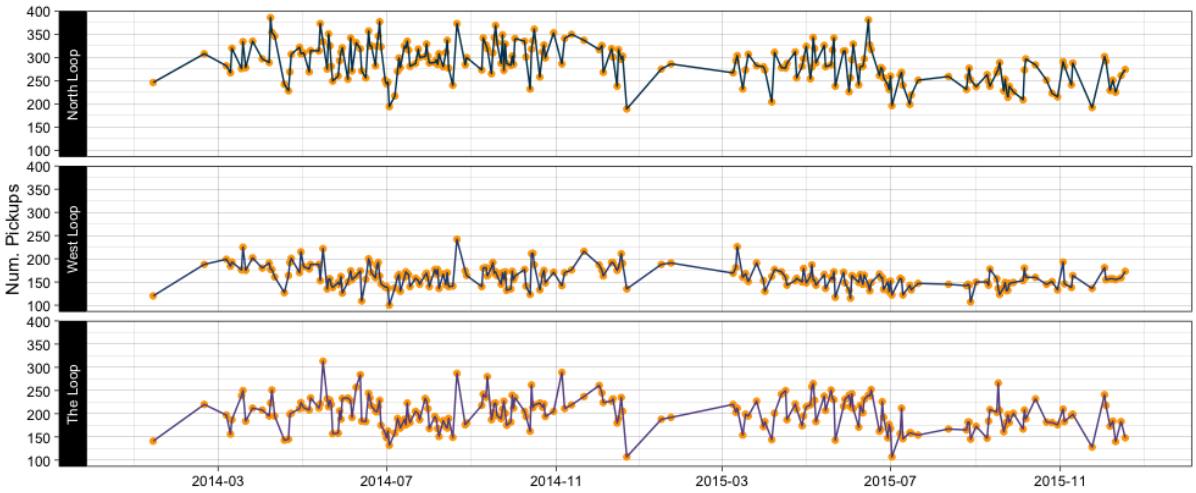

Fig. S2: Time series of total AM Peak pickups from 221 weekdays by Action CA

We present the descriptives of the total pickups by action CA in Table S2. North Loop appears to be relatively more popular and has a greater standard deviation (SD) than the other CAs. However, as shown in Figure S3, the distributions of 9 AM pickups for action CAs are all bell-shaped, which is expected since the data is aggregated from a large number of observed trips.

Table S2: Descriptive statistics of total AM Peak pickups from 221 weekdays by Action CAs

CA	Min.	1st. Qu.	Median	Mean	3rd Qu.	Max	Std. Dev.
North Loop	189	260	287	288	319	386	40
West Loop	100	145	160	161	175	242	23
The Loop	107	180	201	201	223	313	34

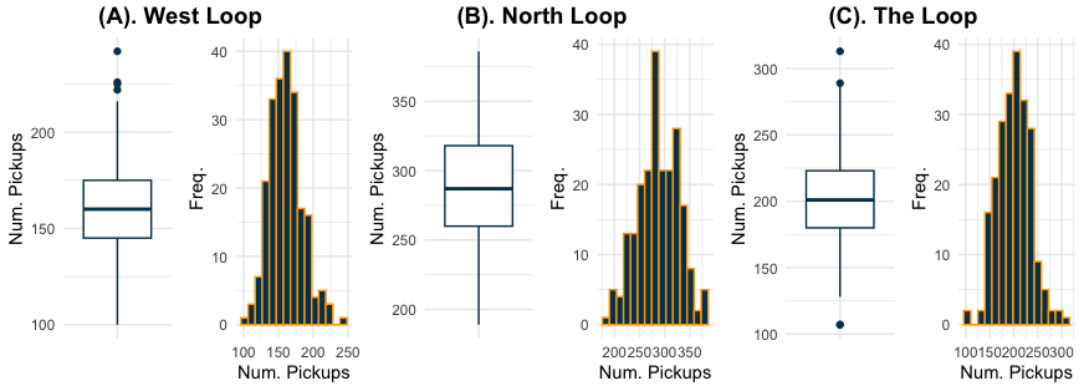

Fig. S3: Distributions of total AM Peak pickups from 221 weekdays by Action CAs

1.5 Deduce Search Flows

In Appendix A.3 of the paper, we describe the data limitation that the taxi OD data does not provide information about drivers who choose to search in an action CA but fail to secure pickups.

To overcome this challenge and obtain the independent variable of the model, we have designed an algorithm capable of deducing each driver's search decisions on each weekday, utilizing the driver's pickup history. We demonstrate the algorithm's pipeline using real examples from the data.

1.5.1 Create Trace Dyads

Suppose we aim to determine the total 9 AM search flow to the Loop on a random weekday (e.g., 2014-04-29). The algorithm queries all drivers who successfully secure 9 AM pickups in the Loop on that weekday from the data. It also gathers information on these drivers' most recent drop-off locations and the drop-off timestamps immediately before their 9:00 AM Loop pickups. Table S3 presents the trace dyads in an edge list, tracking these drivers' movements.

Table S3: Unique drop-off locations before 9 AM pickup in the Loop on 2014-04-29

Previous Drop-off	9 AM Pickup	Count
The Loop (32)	The Loop (32)	150
West Loop (28)	The Loop (32)	19
North Loop (8)	The Loop (32)	18
South Loop (33)	The Loop (32)	6
Midway Airport (56)	The Loop (32)	3
Lincoln Park (7)	The Loop (32)	1

We can visualize the drop-off Community Areas (CAs) of the trace dyads from Table S3 in Figure S4 through a heatmap. This visualization helps us inspect the unique previous drop-off CAs for drivers who picked up in the Loop at 9 AM on 2014-04-29. We observe that drivers who could pick up in the Loop at 9 AM mostly originated from one of the red CAs. Furthermore, we notice that most of the highlighted CAs are in close spatial proximity to the Loop, which is intuitive. Particularly, there is a strong pattern of intra-CA flows, indicating that the majority of drivers who found pickups in the Loop at 9 AM had just completed their last trip in the Loop.

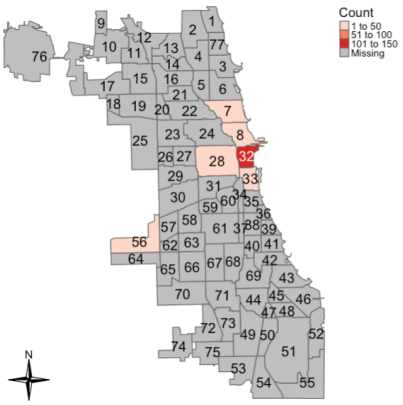

Fig. S4: Example of all previous drop-off CAs before 9 AM Loop pickup

1.5.2 Maximum Idle Time

Knowing the trace dyads is not sufficient. It is also necessary to understand how long it took each driver to find their 9 AM Loop pickups from their previous drop-off locations (i.e., idle time).

Figure S4 provides a good example of why analyzing idle time is important. Among the previous drop-off CAs, one of them is the Midway Airport (CA 56). Naturally, the time taken for a driver to go from a previous drop-off at the airport to picking up a passenger in the Loop at 9 AM is expected to be longer than the time for trips that originate from locations closer to the Loop. The chance of getting a pickup from such a distant location in the Loop at 9 AM could occur merely by chance. Since our strategic setting assumes that drivers are in nearby locations of the action CAs and can choose to search in one of the action CAs within 15 minutes, the algorithm processes idle time by removing extreme outliers using IQR, grouping them by unique origin CAs (i.e., drivers' previous drop-off location), and finding the maximum idle time within each group.

The results are presented in Table S4, which provides the information needed to identify candidate drivers who can find pickups in the Loop at 9 AM on the target weekday. We include drivers who (1) last dropped off in the Loop 20 minutes before 9 AM, (2) last dropped off in the North Loop 15 minutes before 9 AM, (3) last dropped off in the West Loop 15 minutes before 9 AM, and (4) last dropped off in the South Loop 14 minutes before 9 AM. The outlier removal is necessary because it ensures that the drivers we include in the flows into an action CA are drivers who were actively searching for pickups from a nearby community.

Table S4: Unique dropoff CAs with maximum idle time threshold in minutes.

Pre-Loop Drop-off Location (CA ID)	Maximum Waiting Time (Mins)
The Loop (32)	20
West Loop (28)	15
North Loop (8)	15
South Loop (33)	14

1.5.3 Candidate Drivers

By utilizing the trace dyads, we identified 843 unique candidate drivers who are eligible to find pickups in the Loop at 9 AM. Of the 199 total drivers who actually secured pickups according to the data, 171 (86%) are included in our candidate list. The remaining 28 (14%) drivers could be attributed to drivers who came organically online, starting their day by finding a pickup in the Loop at around 9 AM. This indicates that our candidate list is comprehensive enough to encompass most eligible drivers.

For our 843 candidates, we ascertain 171 of them did indeed pick up in the Loop, affirming that these drivers must have searched the Loop for pickups. However, the status of the remaining 672 drivers raises questions. By querying from the data, we discover 376 of these drivers found pickups elsewhere, not in the Loop, leading us to remove these drivers from the candidate list. This leaves us with a critical question: "What happened to the remaining 296 drivers, and how many of these drivers would choose to search the Loop based on their current location?" To answer this question, we delve into these drivers' search dyad.

1.5.4 Deduce Search Decisions

To deduce the potential search locations of these 296 unknown candidates, the algorithm examines each driver's search dyads, summarized by trips a driver completed during the AM peaks of the preceding 10 days. The algorithm organizes each driver's search history by drop-off timestamps, linking the drop-off Community Area (CA) of one trip to the pickup CA of the next. This process generates a set of search dyads that aids in inferring possible search decisions.

For instance, consider a candidate driver's set of search dyads as illustrated in Table S5. If the candidate's current location is the North Loop (as shown in rows 1 and 2), the algorithm assesses search dyads starting with North Loop and infers that the driver would likely search in the North Loop again, given a higher likelihood of securing pickups based on past search history. Conversely, if the candidate's current location is the West Loop and the counts are evenly split, both The Loop and West Loop are considered equally likely choices. The algorithm then resolves the search decision through a random draw. In situations where a candidate lacks any prior search history in their current CA, the algorithm deduces the search decision based on the CA from which the driver has historically picked up the most, resulting in the North Loop for this example. There might be corner cases where a driver may have no search history within the last 10 days; in such scenarios, we randomly select a CA from a list of neighboring CAs to the driver's current location, following the rook contiguity spatial weights.

Table S5: Example of a driver's unique search prior represented by search dyads

Previous Drop-off	Next Pickup	Counts
North Loop	West Loop	10
North Loop	North Loop	15
West Loop	The Loop	6
West Loop	West Loop	6

Using the conditional prior, we determine that 147 (out of 296) drivers would opt to search the Loop. By adding these 147 drivers to the 171 drivers who actually secured pickups in the Loop, we are able to calculate the total flow to the Loop (i.e., $318 = 147 + 171$) at 9 AM and the pickup probability for the Loop (i.e., $54\% = 171/318$) on that example weekday of 2014-04-29.

It is important to highlight that our algorithm is both realistic and comprehensive, as it can accurately deduce the pickup probability based on the total number of real taxi drivers (1) whose current location enables them to search for pickups in the Loop at 9 AM on that specific day, and (2) who will actually decide to search the Loop from their current location.

2 Counterfactual Model

The data we use to fit the model is presented in Figure S5. It is evident that the intercepts exhibit large variation across action CAs. Although all slopes are positive, they are less than 1, indicating that within each group, there is expected to be greater variation among intercepts and relatively smaller variation among slopes. Given our decision-making context, we hypothesize a certain degree of correlation between slopes and intercepts. This hypothesis has guided us in fitting the model based on the following priors.

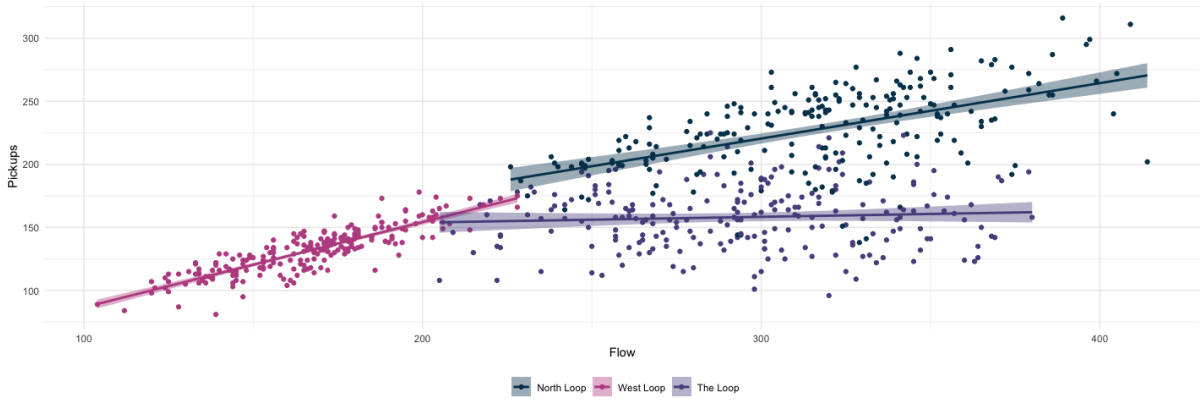

Fig. S5: Model Data

$$\text{pickup}_i \sim \text{Normal}(\mu_i, \sigma_i) \quad \text{[likelihood]} \tag{S1}$$

$$\mu_i = \alpha_{\text{CA}[i]} + \beta_{\text{CA}[i]} \times \text{flow}_i \quad \text{[linear model]} \tag{S2}$$

$$\begin{bmatrix} \alpha_{\text{CA}} \\ \beta_{\text{CA}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{S} \right) \quad \text{[population of varying effects]} \tag{S3}$$

$$\mathbf{S} = \begin{bmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{bmatrix} \mathbf{R} \begin{bmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{bmatrix} \quad \text{[construct covariance matrix]} \tag{S4}$$

$$\text{where } \mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{[correlation matrix]} \tag{S5}$$

$$\omega \sim \text{Student's T}(3, 60, 10) \quad \text{[prior for average intercept]} \tag{S6}$$

$$\beta \sim \text{Student's T}(3, 0, 15) \quad \text{[prior for average slope]} \tag{S7}$$

$$\sigma \sim \text{HalfCauchy}(15) \quad \text{[prior std. dev. within CAs]} \tag{S8}$$

$$\sigma_\alpha \sim \text{Student's T}(3, 75, 20) \quad \text{[prior std. dev. among intercepts]} \tag{S9}$$

$$\sigma_\beta \sim \text{Exponential}(2) \quad \text{[prior std. dev. among slopes]} \tag{S10}$$

$$\mathbf{R} \sim \text{LKJcorr}(5) \quad \text{[prior for correlation matrix]} \tag{S11}$$

We fit our counterfactual model, as stated above, using BRMS (version 2.16.7) in R (version 4.1.1) using four chains with 10,000 iterations and 5,000 warm-ups. Figure S6 displays the trace plots, illustrating that our Markov Chain Monte Carlo (MCMC) chains have mixed well.

We briefly discuss our MCMC chain convergence in Figure S7 using the $\hat{R}$ statistic (i.e., Gelman-Rubin convergence diagnostic) and the number of effective sample size ratios. If the chains have converged to equilibrium, the $\hat{R}$ values should be equal to 1. While our $\hat{R}$ are not exactly 1, they closely approach the target. The effective sample size ratio (i.e., n_{eff} ratio) serves as an indicator of sampling efficiency. Ideally, this ratio should be as high as possible. Each of our parameters shows an n_{eff}/N greater than 0.2, indicating that our chains have successfully identified effective samples. The posterior distributions of model coefficients are displayed in Figure S8, and the posterior predictive checks are shown in Figure S9.

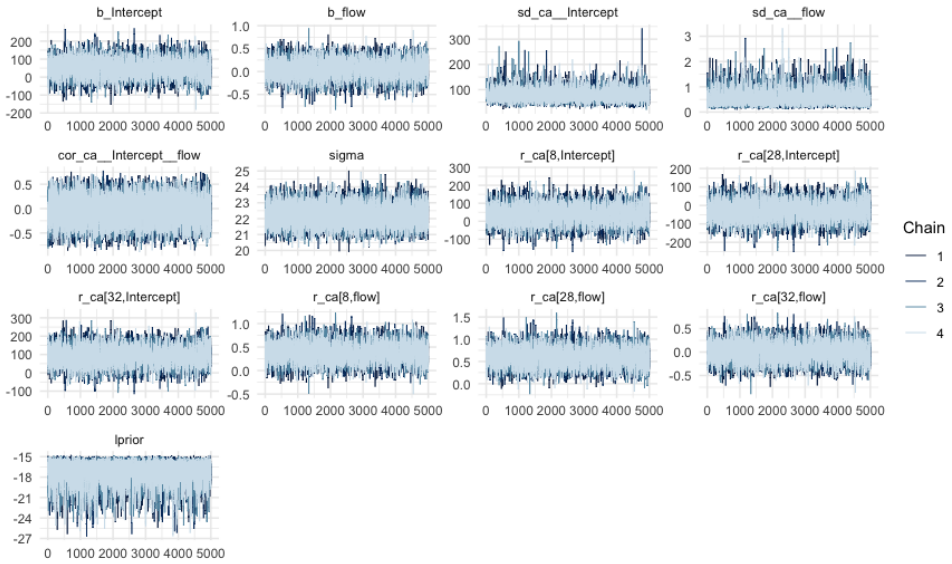

Fig. S6: Trace plots of the counterfactual model by coefficients

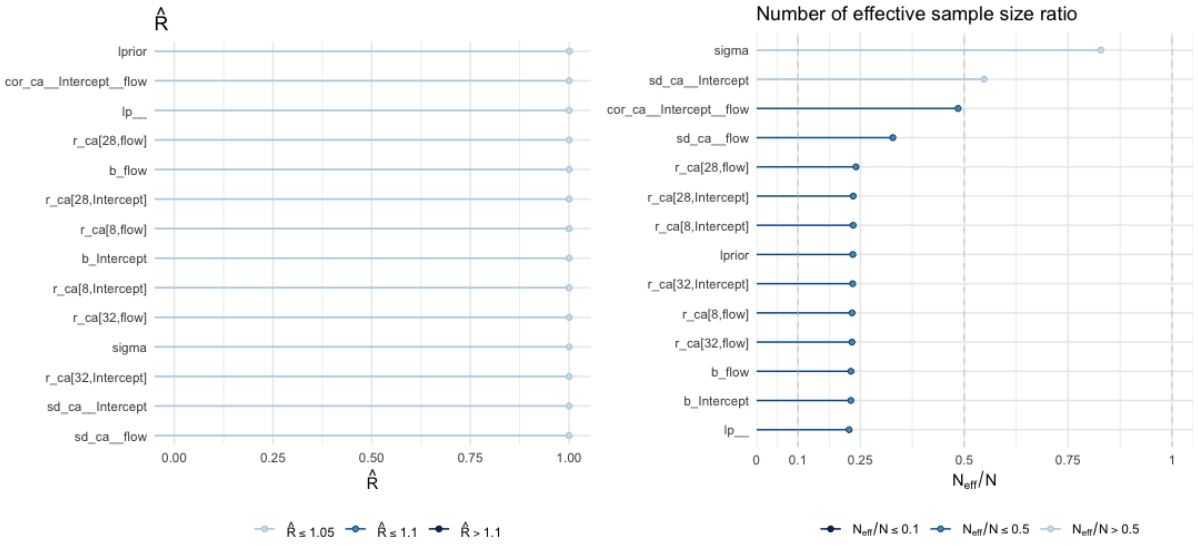

Fig. S7: Counterfactual Model Diagnostics

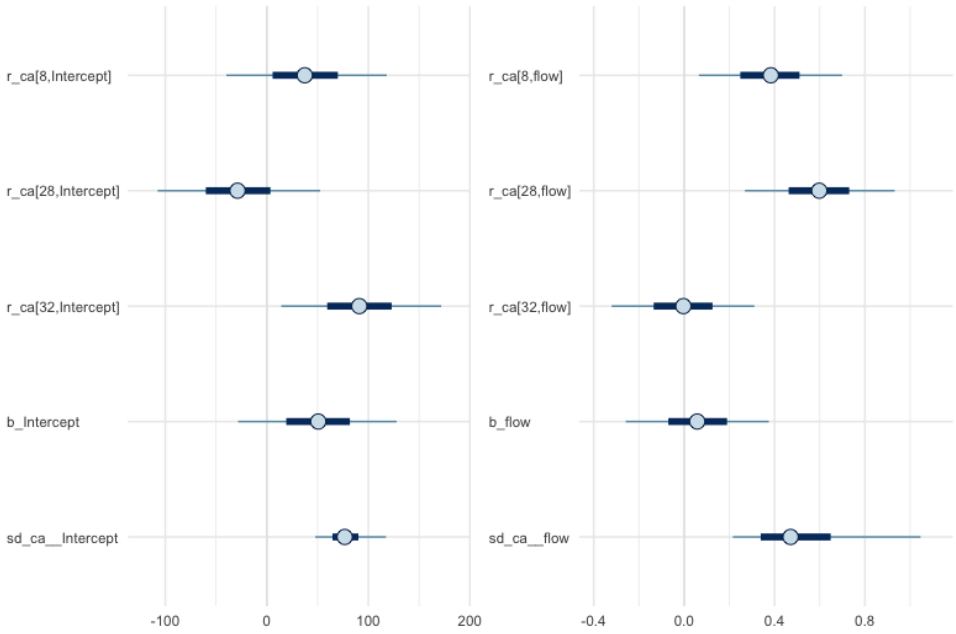

Fig. S8: Posterior of Counterfactual Model Coefficients

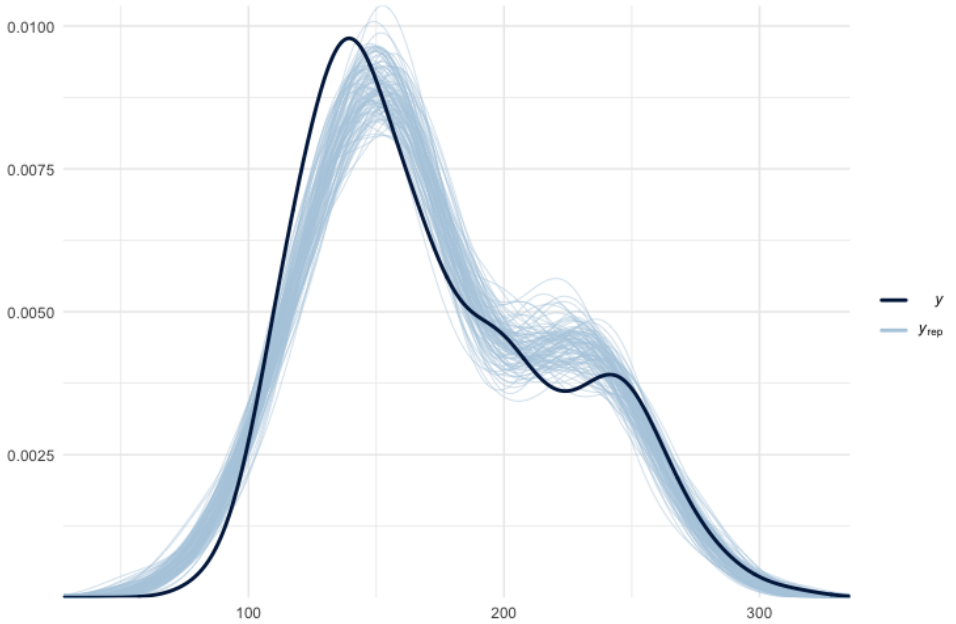

Fig. S9: Counterfactual Model Posterior Predictive Check

3 Earth Mover's Distance

In Section 3.2.8 of the paper, we detail the application of the Earth Mover's Distance (EMD) to quantify both the distribution shift (i.e., aggregate-level response variable) and anticipation error (i.e., individual-level response variable). EMD evaluates the minimum amount of "work" required to match the two flow distributions. To calculate EMD, we initially map the two distributions on a 2D grid as illustrated in Figure S10. Suppose a participant's anticipated flow distribution is

$$A = \{(a_1, w_1), \cdots, (a_i, w_i)\} \tag{S12}$$

and the flow of level-specific outcome based on the endowed belief is

$$E = \{(e_1, w_1), \cdots, (e_i, w_i)\} \tag{S13}$$

Because both flows are formed by decisions of N drivers, that is $|A| = |E| = N$, and the weight of each point is

$$w_i = w_j = 1 \quad \forall i, j \in N \tag{S14}$$

We hence have an integer solution property and can one-to-one point match the distributions. Suppose $F(A, E)$ is a set of movements needed to shift a level-k participant's anticipated flow and align with the flow of level-specific outcome, such as the solid lines shown in Figure S10. Each shift of F has a cost, which is the Euclidean distance defined as $L(a_i, e_i)$. Hence the total cost of a feasible shift is the sum of distances over all movements expressed as

$$Cost(F, A, E) = \sum_i \sum_j f_{ij} L(a_i, e_j) \tag{S15}$$

We use Equation (S16) below to optimize the minimum amount of Euclidean distance to compute the number of shifts and then standardize it by the total number of drivers. The EMD measure can be interpreted as the average amount of distance each circle in the anticipated flow distribution must move in order to match the flows of the level-specific outcome.

$$EMD(A, E) = \frac{min_{F=(f_{ij}) \in F(a,e)} Cost(F, A, E)}{N} \tag{S16}$$

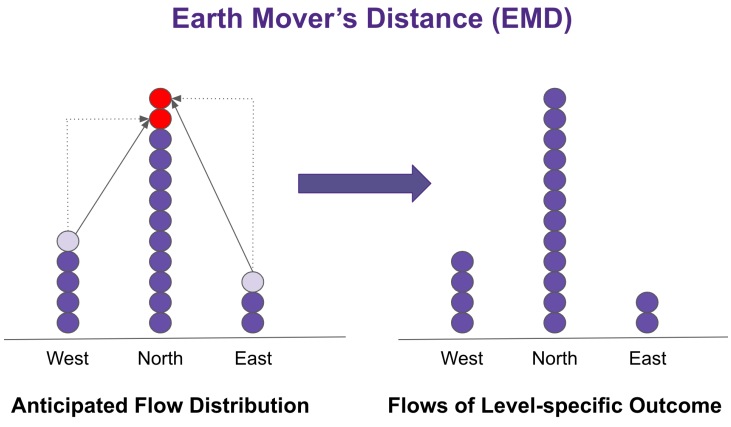

Fig. S10: An example showing EMD computation between a level-k participant's anticipated flow and the flow of level-specific outcome. This example has an EMD of 0.84.

4 Welfare Optimization

As described in Section 3.2.8 of the paper, we use a non-linear optimization via the augmented Lagrange method, which uses random initialization and multiple restarts to find a flow distribution that can render maximum pickups for each decision scenario. This optimization is carried out in R (version = 4.1.1) using 'Rsolnp' (version = 1.16). The goal of the optimization is to determine a flow vector of length three, $(d1, d2, d3)$, corresponding to the search flows towards each district in a scenario involving N players so that this flow will result in the highest number of pickups as predicted by our counterfactual model.

We first identify the bounds for each district, d_i, from simulated system outcomes and the actual data used to fit the model. We then construct an objective function that takes an input flow vector $(d1, d2, d3)$ and outputs a numeric value indicating the count of drivers who did not secure pickups. This function assesses the proposed flows with the goal of minimizing the number of drivers without pickup, subject to constraints ensuring that the flow vector's sum equals to the total number of drivers N and that the minimum and maximum flow to a district does not exceed the bounds. As detailed in Appendix A.4, because we use two heuristics to ensure that the counterfactual model can make plausible predictions, it makes the objective function non-smooth and it becomes harder to evaluate the optimalities. Therefore, we opt for initiating multiple restarts of the solver from a set of randomly selected parameters.

5 Models of Response Variables

We report our model diagnostics, posteriors, and posterior predictive checks in subsections below.

5.1 Aggregate-level: Welfare Ratio (Ratio)

The welfare ratio model employs beta regressions. From Figure S11, we see $\hat{R}$ values are centered at 1 (in row 1) with large n_{eff} ratios across collection arms (in row 2), signaling model convergence. We illustrate the posterior distributions of the coefficients for each collection arm in Figure S12, and the posterior predictive checks in Figure S13. From the posterior predictive, we detect bimodality that is not fully captured, which is likely due to trial-level differences.

5.2 Aggregate-level: Distribution Shift (DS)

Our DS models are a multiple linear regressions, and we see evidence of model convergence from the $\hat{R}$ and n_{eff} ratio depicted in Figure S14. The posterior distribution and the posterior predictive check are presented in Figure S15 and Figure S16, respectively.

5.3 Individual-level: Best Response (BR)

For our BR models, which incorporate both fixed and random effects by participants, we present the MCMC diagnostics for $\hat{R}$ and n_{eff} ratio as distributions in Figure S17. The $\hat{R}$ values for model coefficients are consistently close to 1 across all collection arms, indicating chain converges. However, for the n_{eff} ratio, several parameters associated with random effects (i.e. participants), especially for level-1s, have n_{eff} ratio less than 0.1, This suggests potential inefficiency in our chain's sampling process for these parameters.

We present the posterior distributions of all model coefficients in Figure S18 and illustrate four draws of the binned residual plot to assess the posterior predictive in Figure S19. We see that most points reside within the confidence limits across the collection arms, suggesting that the models are reliable.

5.4 Individual-level: Anticipation Error (AE)

In Figure S20, the $\hat{R}$ values for all coefficients in our AE models are predominantly at 1, signaling model convergence. However, several parameters associated with random effects by participants have n_{eff} ratio less than 0.1, indicating inefficiencies in sampling. Despite these sampling challenges, our AE model demonstrates good fits, as evidenced in Figure S22. The posterior distribution of the model's coefficients is detailed in Figure S21.

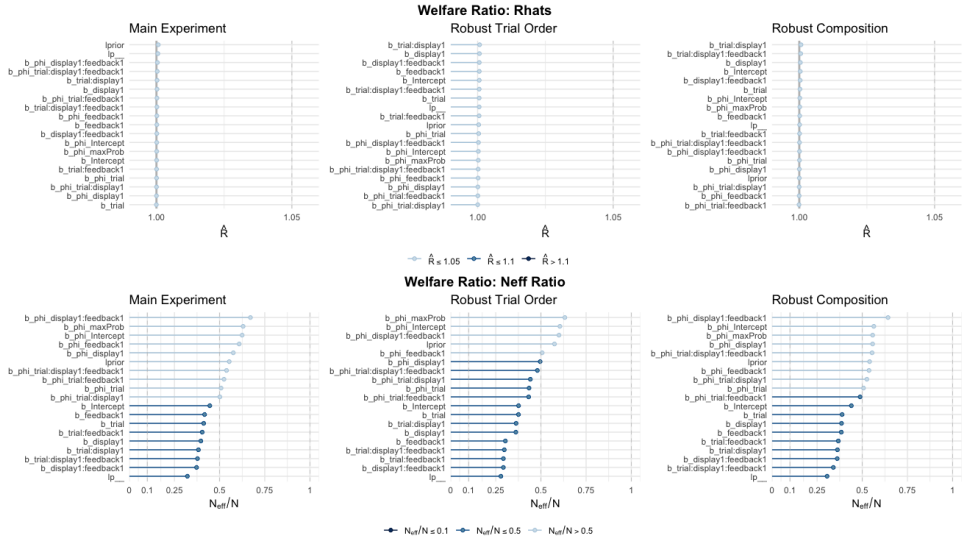

Fig. S11: Ratio Model Diagnostics

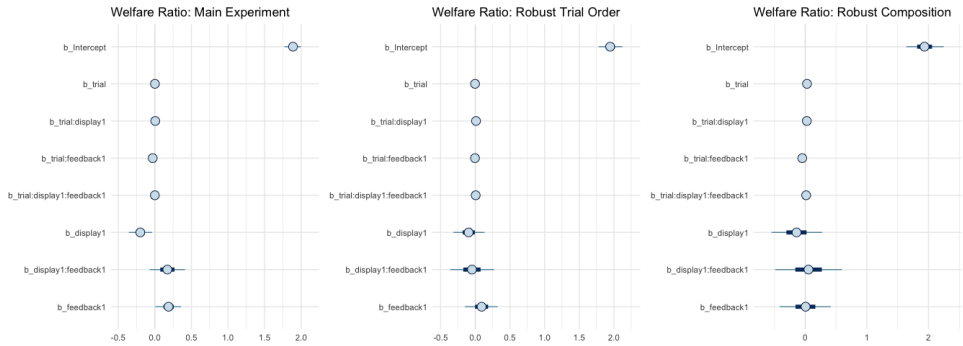

Fig. S12: Ratio Model Posteriors

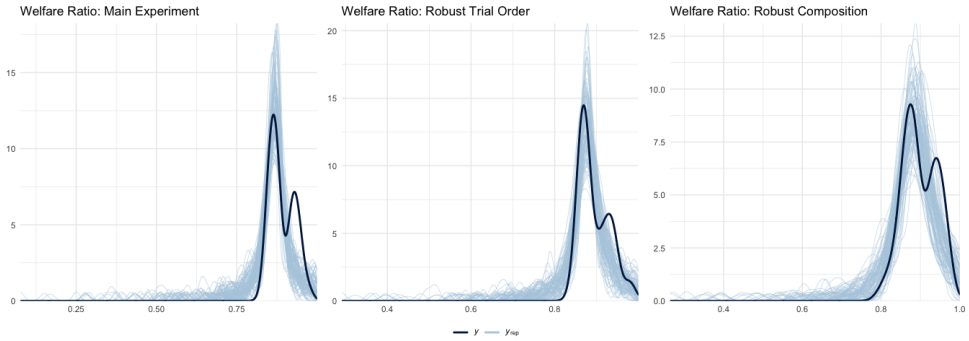

Fig. S13: Ratio Model Posterior Predictive Checks

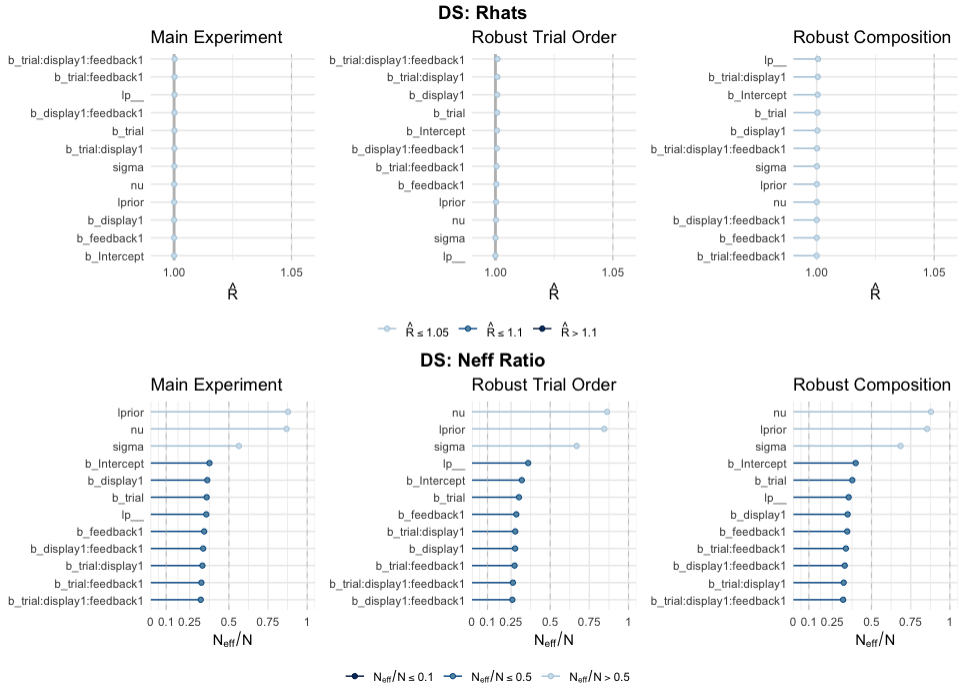

Fig. S14: DS Model Diagnostics

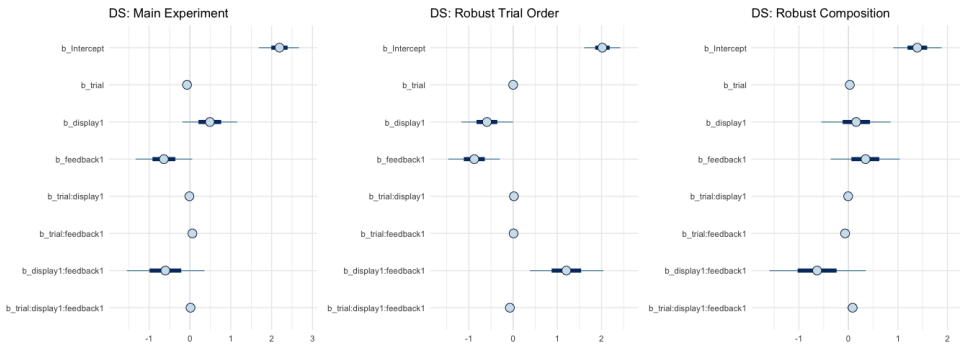

Fig. S15: DS Model Posteriors

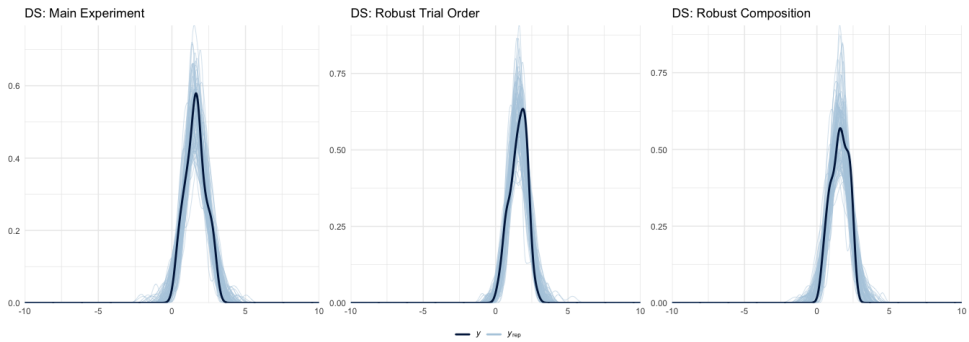

Fig. S16: DS Model Posterior Predictive Checks

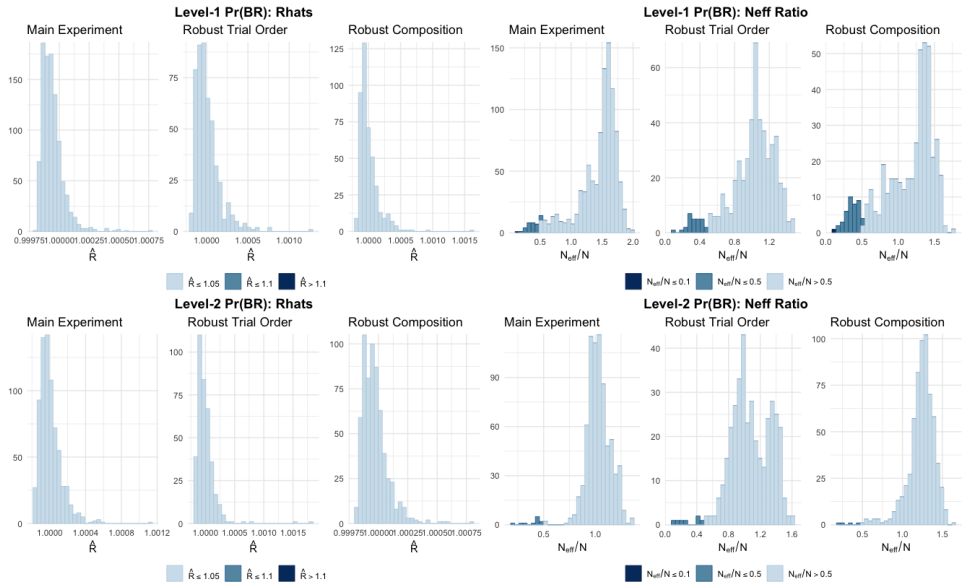

Fig. S17: BR Model Diagnostics

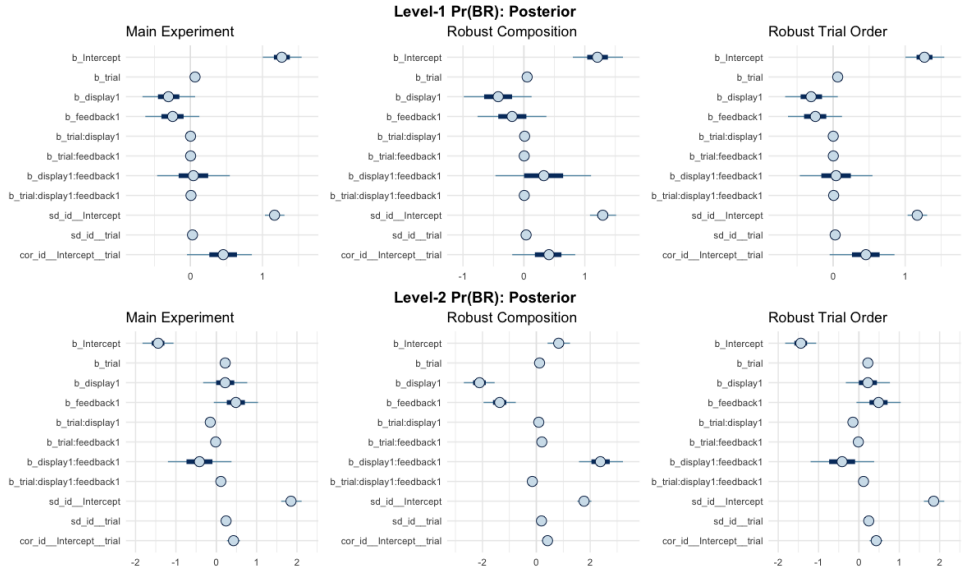

Fig. S18: BR Model Posteriors

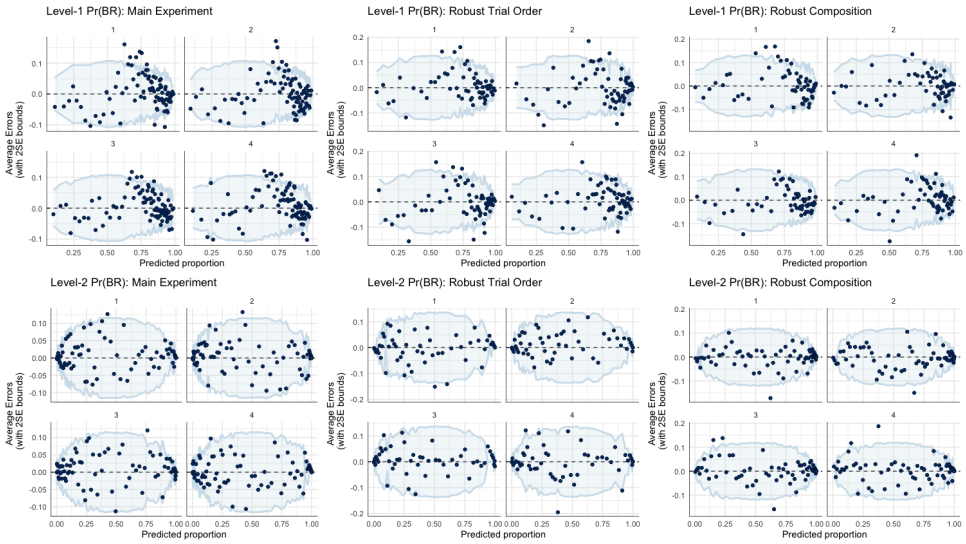

Fig. S19: BR Model Posterior Predictive Checks

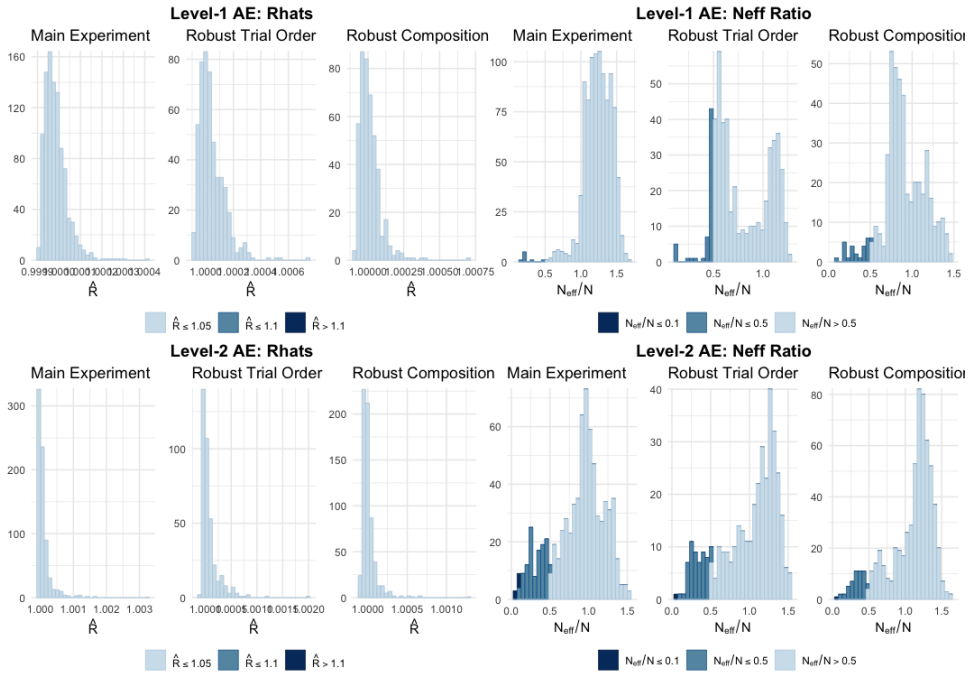

Fig. S20: AE Model Diagnostics

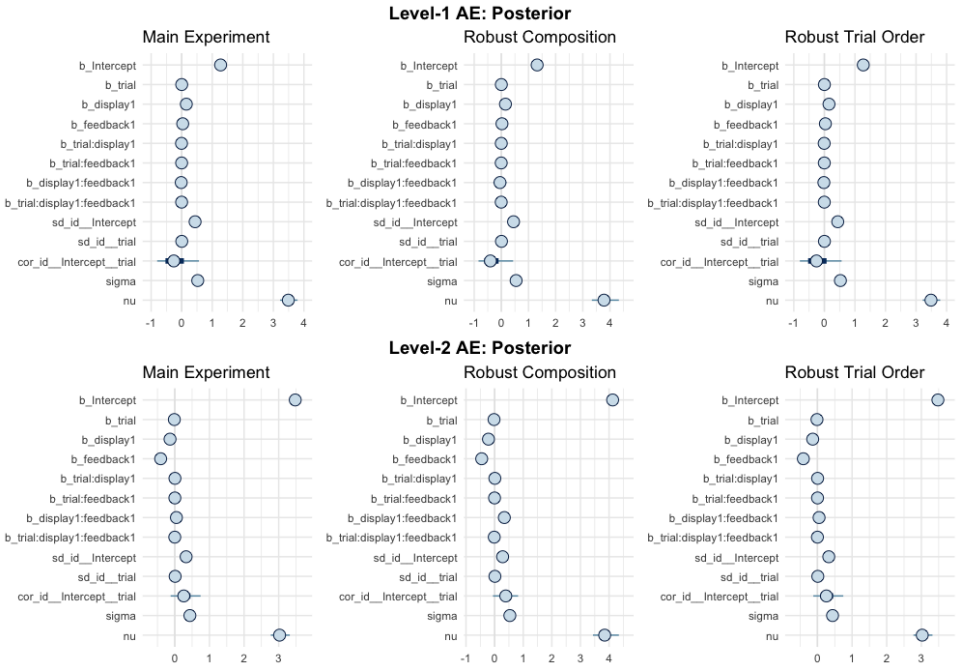

Fig. S21: AE Model Posteriors

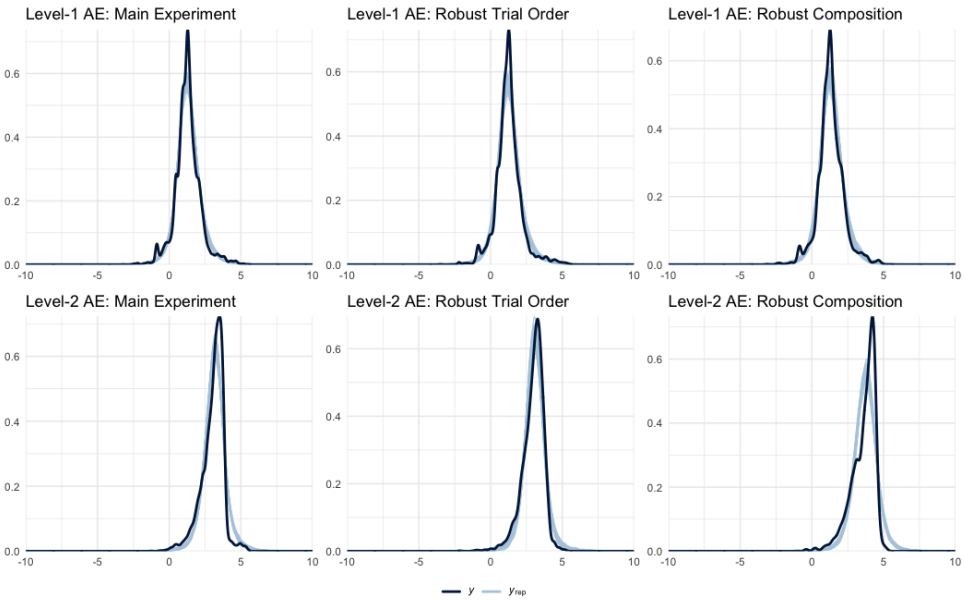

Fig. S22: AE Model Posterior Predictive Checks