# Categorical Mixed-Effects Analysis of Truth-Value Judgments

## 1 Statistical analysis: Categorical (multinomial) mixed-effects model

### 1.1 Rationale

As a robustness check that does not assume an ordinal relationship among the three response options, we additionally analyzed the judgments using a categorical (multinomial) mixed-effects model. This treats *Neither completely true nor completely false* as a qualitatively distinct response category rather than as an intermediate point on an ordered scale.

### 1.2 Design and data

The dataset comprised 1260 trial-level observations (42 participants × 30 trials). Two within-participant factors were manipulated: DISPLAY $(0, 2, 4)$ and PROMPT (`ALL`, `AND`). Participants were assigned to one of six questionnaire versions (`Group`). Responses had three categories: *Completely false*, *Neither completely true nor completely false*, and *Completely true.*

### 1.3 Model

We fit a Bayesian multinomial-logit mixed-effects model using *Completely false* as the reference category. The model estimates two logit-linear predictors: (i) *Neither* vs. *False* and (ii) *True* vs. *False.* Both predictors included fixed effects of PROMPT, DISPLAY, and their interaction, along with random intercepts for participants and questionnaire version:

$$\log \frac{P(\text{Neither})}{P(\text{False})} = \eta_{\text{Neither}} = \beta_0^{(N)} + \beta_{\text{PROMPT}}^{(N)} + \beta_{\text{DISPLAY}}^{(N)} + \beta_{\text{PROMPT}\times\text{DISPLAY}}^{(N)} + u_{\text{participant}}^{(N)} + v_{\text{Group}}^{(N)}, \tag{1}$$

$$\log \frac{P(\text{True})}{P(\text{False})} = \eta_{\text{True}} = \beta_0^{(T)} + \beta_{\text{PROMPT}}^{(T)} + \beta_{\text{DISPLAY}}^{(T)} + \beta_{\text{PROMPT}\times\text{DISPLAY}}^{(T)} + u_{\text{participant}}^{(T)} + v_{\text{Group}}^{(T)}. \tag{2}$$

All parameters were estimated via NUTS. Convergence diagnostics indicated good mixing (all $\hat{R} \approx 1.00$; effective sample sizes were large).

### 1.4 Condition-wise predicted probabilities

We fit a Bayesian multinomial-logit mixed-effects model with *Completely false* as the reference category. The model estimates two logit-linear predictors, one for *Neither* vs. *False* and one for *True* vs. *False.* Both predictors included fixed effects of PROMPT, DISPLAY, and their interaction, with random intercepts for participants and questionnaire version:

$$\texttt{Response}_{cat} \sim \text{PROMPT} \times \text{DISPLAY} + (1 \mid \texttt{participant}) + (1 \mid \texttt{Group}). \tag{3}$$

Convergence diagnostics indicated good mixing (all $\hat{R} \approx 1.00$; effective sample sizes large).

Table 1 reports population-level predicted probabilities (random effects marginalized out) for each PROMPT×DISPLAY cell, summarized as posterior means and 95% credible intervals (CrI).

| PROMPT | DISPLAY | $P$(False) | $P$(Neither) | $P$(True) |
|---|---|---|---|---|
| ALL | 0 | 0.994 [0.982, 0.999] | 0.00060 [0.00003, 0.00256] | 0.00580 [0.00080, 0.0177] |
| AND | 0 | 0.995 [0.984, 0.999] | 0.00167 [0.00012, 0.00583] | 0.00374 [0.00025, 0.0135] |
| ALL | 2 | 0.993 [0.980, 0.999] | 0.00524 [0.00071, 0.0169] | 0.00174 [0.00004, 0.00804] |
| AND | 2 | 0.484 [0.220, 0.743] | 0.516 [0.256, 0.780] | 0.00069 [0.00000, 0.00469] |
| ALL | 4 | 0.0140 [0.00317, 0.0344] | 0.00023 [0.00000, 0.00141] | 0.986 [0.965, 0.997] |
| AND | 4 | 0.00216 [0.00006, 0.00938] | 0.00011 [0.00000, 0.00081] | 0.998 [0.990, 1.000] |

Table 1: Population-level predicted response probabilities from the categorical mixed-effects model (posterior mean with 95% CrI).

## 1.5 Fixed effects (log-odds ratios)

We summarize the main fixed-effect patterns in terms of posterior means and 95% credible intervals (CrI). Coefficients are log-odds effects on the corresponding logit scale.

**(i) DISPLAY strongly increases True responding.** In the *True vs. False* equation, DISPLAY=4 (relative to 0) shows a large positive effect:

$$\beta^{(T)}_{\text{DISPLAY=4}} = 9.84 \quad [8.23,\ 11.72],$$

indicating that DISPLAY=4 strongly shifts responses toward *Completely true* relative to *Completely false.*

**(ii) A pronounced interaction at DISPLAY=2 for Neither responses.** In the *Neither vs. False* equation, the interaction term at DISPLAY=2 is robustly positive:

$$\beta^{(N)}_{\text{PROMPTAND}\times\text{DISPLAY=2}} = 4.51 \quad [2.45,\ 6.62],$$

indicating that, at DISPLAY=2, the `AND` prompt substantially increases the odds of giving a *Neither* response (relative to *False*) compared to `ALL`.

**(iii) Prompt differences at DISPLAY=4 are limited by ceiling effects.** While DISPLAY=4 yields near-ceiling *True* responding, the model provides evidence for an additional `AND` effect in the *True vs. False* equation at DISPLAY=4:

$$\beta^{(T)}_{\text{PROMPTAND}\times\text{DISPLAY=4}} = 3.00 \quad [0.34,\ 5.88],$$

consistent with `AND` producing slightly stronger movement toward *True* in a regime where both prompts are already near ceiling.

## 1.6 Random effects

Participant-level variability was substantial, especially for the *Neither vs. False* component (random-intercept SD $\approx$ 2.70 [1.85, 3.79]), reflecting considerable individual differences in the propensity to select the intermediate response. Questionnaire-version effects (`Group`) were smaller and more uncertain (SDs $\approx$ 0.61 and 0.43).

## 1.7 Summary of the categorical analysis

The categorical model reproduces the same qualitative pattern as the ordinal analysis. Responses at DISPLAY=0 are overwhelmingly *Completely false* in both prompts, while responses at DISPLAY=4 are overwhelmingly *Completely true*. The main divergence between prompts emerges at DISPLAY=2: under `ALL`, responses remain near-ceiling *Completely false* ($P(\text{False}) \approx 0.993$), whereas under `AND` the distribution shifts sharply toward the intermediate category ($P(\text{Neither}) \approx 0.516$ and $P(\text{False}) \approx 0.484$). Thus, even without an ordinality assumption, `AND` selectively increases *Neither* responding at the intermediate display level.