Categorical Mixed-Effects Analysis (Monotonicity × Display)

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

1.1 Design and data

We analyzed trial-level truth-value judgments from $N = 42$ participants (1260 observations). Each participant responded to 30 items crossing two within-participant factors: DISPLAY $(0, 2, 4)$ and MONOTONICITY (NEG, POS). Responses were recorded in three categories: *Completely false*, *Neither completely true nor completely false*, and *Completely true*. Participants were assigned to one of six questionnaire versions (indexed by Group).

1.2 Model

To avoid imposing an ordinal relationship among the three response options, we fit a Bayesian categorical (multinomial-logit) mixed-effects model. *Completely false* served as the reference category, so the model estimates two logit-linear predictors: (i) *Neither* vs. *False* and (ii) *True* vs. *False*. Both predictors included fixed effects of MONOTONICITY, DISPLAY, and their interaction, with random intercepts for participants and questionnaire version:

$$\text{Response}_{cat} \sim \text{MONOTONICITY} \times \text{DISPLAY} + (1 \mid \text{participant}) + (1 \mid \text{Group}). \quad (1)$$

Convergence diagnostics indicated good mixing (all $\hat{R} \approx 1.00$; effective sample sizes were generally adequate). Because several cells exhibited near-deterministic behavior (probabilities near 0 or 1), some log-odds coefficients were necessarily large; we therefore summarize the fitted effects primarily using posterior predictive probabilities and probability-scale contrasts.

1.3 Condition-wise predicted probabilities

Table 1 reports population-level predicted probabilities (random effects marginalized out) for each MONOTONICITY×DISPLAY cell. Values are posterior means with 95% credible intervals (CrI).

MONOTONICITY	DISPLAY	P(False)	P(Neither)	P(True)
NEG	0	0.00832 [0.00091, 0.0261]	0.00142 [0.00006, 0.00612]	0.990 [0.971, 0.999]
POS	0	0.996 [0.987, 0.999]	0.00412 [0.00053, 0.0130]	0.000053 $[< 10^{-6}, 0.000$
NEG	2	0.0485 [0.0112, 0.125]	0.945 [0.857, 0.987]	0.00633 [0.00069, 0.021
POS	2	0.134 [0.0388, 0.303]	0.860 [0.684, 0.959]	0.00627 [0.00057, 0.022
NEG	4	0.994 [0.982, 0.999]	0.00305 [0.00032, 0.0105]	0.00277 [0.00006, 0.012
POS	4	0.000031 $[< 10^{-6}, 0.00022]$	0.000676 [0.000008, 0.00353]	0.999 [0.996, 1.000]

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

1.4 Key posterior contrasts

We report three probability-scale contrasts that capture the main interaction pattern.

Monotonicity effect at the endpoints (Display 0 and 4). At DISPLAY=0, NEG yields far more *Completely true* responses than POS:

$$\Delta P(\text{True} \mid \text{DISPLAY} = 0; \text{ NEG} - \text{POS}) = 0.990 \quad [0.971, \ 0.999]. \tag{2}$$

At DISPLAY=4, the direction reverses: POS yields far more *Completely true* responses than NEG:

$$\Delta P(\text{True} \mid \text{DISPLAY} = 4; \text{ POS} - \text{NEG}) = 0.997 \quad [0.987, \ 1.000]. \tag{3}$$

Monotonicity effect at the intermediate level (Display 2). At DISPLAY=2, both conditions concentrate on the intermediate category, but NEG shows a higher probability of *Neither* than POS:

$$\Delta P(\text{Neither} \mid \text{DISPLAY} = 2; \text{ NEG} - \text{POS}) = 0.085 \quad [0.016, \ 0.211]. \tag{4}$$

1.5 Summary

The categorical analysis yields a strong crossover interaction between MONOTONICITY and DISPLAY. At DISPLAY=0, NEG responses are almost always *Completely true*, whereas POS responses are almost always *Completely false*. At DISPLAY=4, the pattern reverses, with POS responses almost always *Completely true* and NEG responses almost always *Completely false*. At the intermediate level DISPLAY=2, both monotonicity conditions produce predominantly intermediate ("Neither") judgments, with NEG showing a modestly higher probability of *Neither* responses than POS. Overall, the multinomial model corroborates the primary ordinal analysis while avoiding the assumption that the three response options form an ordered scale.