

# Validating Dating Market Theory with Real Data

Simulation Analysis Using OkCupid Parameters

Group 4

November 30, 2025

## 1 Introduction

Our teammate developed a theoretical framework showing how users on dating platforms update beliefs about market quality and exit when frustrated. We validate this theory by extracting parameters from 60,000 real OkCupid profiles and re-running the simulations.

**Research Question:** Do theoretical predictions about exit dynamics hold when calibrated with real-world dating market parameters?

## 2 Dataset

**Source:** OkCupid profiles from San Francisco Bay Area

**Sample:** 59,946 profiles across 63 markets (location  $\times$  orientation)

**Key Variables:**

- Demographics: age, sex, orientation, education, body type
- Essays: 10 free-text fields describing goals and preferences
- Market identifiers: location, orientation

## 3 Parameter Extraction

### 3.1 Rating Index ( $r_i$ )

Constructed as weighted combination of observable characteristics:

$$r_i = \sum_j w_j \cdot \text{score}_j \quad (1)$$

Component	Weight	Scoring
Education	20%	Ordinal scale 1–5 (HS → PhD)
Body type	20%	Fit=5, Thin=4, Avg=3, Extra=2, Overweight=1
Effort	20%	Profile completeness + essay words
Height	15%	Gender-specific ideals
Age	15%	Peak at 25–32 years
Income	10%	Log-transformed

Normalized to [0, 1]. Result:  $\bar{r}_i = 0.611$  (SD = 0.126), follows Beta(8.54, 5.44).

### 3.2 Relationship Goals ( $g_i$ )

Classified via keyword analysis in essay text:

- **LTR keywords** (40 terms): “long-term”, “relationship”, “marriage”, “commitment”
- **Casual keywords** (28 terms): “casual”, “hookup”, “fling”, “no strings”

**Classification:**

$$g_i = \begin{cases} \text{LTR} & \text{if LTR\_count} \geq 2 \times \text{Casual\_count} \\ \text{Casual} & \text{if Casual\_count} \geq 2 \times \text{LTR\_count} \\ \text{Ambiguous} & \text{otherwise} \end{cases}$$

**Result:** 54.7% LTR, 34.7% Ambiguous, 10.6% Casual

### 3.3 Market Clarity ( $\psi_m$ )

Average effort of potential partners in market (proxy for signal clarity):

$$\psi_m = \frac{1}{N_m} \sum_{j \in m} \text{effort}_j$$

**Result:** Mean = 0.688 (SD = 0.030)

### 3.4 Summary of Extracted Parameters

Parameter	Value	Source
Rating distribution	Beta(8.54, 5.44)	Fitted to 59,946 profiles
Mean rating	$\bar{r}_i = 0.611$	Direct calculation
LTR share	$\rho_m = 0.547$	Keyword classification
Market clarity	$\psi_m = 0.688$	Average partner effort
Self-based prior (LTR)	$\rho_{i0} = 0.60$	Theoretical assumption
Prior strength	$\tau_i = 10$	Calibrated
Frustration threshold	$\bar{p} = 0.20$	From theory ( $C/V$ )
Batch size	$K = 20$	Standard

Table 1: Parameters extracted from OkCupid data and used in simulations

## 4 Market Selection

We grouped profiles by location  $\times$  orientation and computed market-level parameters. From 63 markets with  $N \geq 50$  users, we selected three spanning the LTR share range:

**Selection rationale:** These markets span the observed LTR range (44%–70%) and have sufficient sample sizes. They represent real variation in market quality across demographic segments.

## 5 Simulation Methodology

We follow the Bayesian learning framework from theory:

Label	Market	N	LTR%	Clarity	Avg Rating
Low	SF, Bisexual	1,203	46.2%	0.740	0.584
Medium	El Cerrito, Straight	280	58.6%	0.725	0.605
High	Alameda, Gay	57	70.2%	0.703	0.552

Table 2: Three selected markets from OkCupid data

## 5.1 Model

Each user  $i$  has:

- Rating  $r_i \sim \text{Beta}(8.54, 5.44)$
- Goal  $g_i \in \{\text{LTR, Casual}\}$  with  $P(g_i = \text{LTR}) = \rho_m$
- Self-based prior:  $\rho_{i0} = 0.60$  for LTR users
- Prior strength:  $\tau_i = 10$

## 5.2 Dynamics

Each period  $t$ :

1. User sees  $K = 20$  profiles
2.  $K^{\text{eff}} \sim \text{Binomial}(K, \psi_m)$  are informative
3.  $K^L \sim \text{Binomial}(K^{\text{eff}}, \rho_m)$  appear LTR-oriented
4. Update beliefs via Beta-Binomial:  $\alpha_t \leftarrow \alpha_{t-1} + K^L$ ,  $\beta_t \leftarrow \beta_{t-1} + (K^{\text{eff}} - K^L)$
5. Posterior belief:  $\hat{\rho}_{i,t} = \alpha_t / (\alpha_t + \beta_t)$
6. Success probability:  $\hat{p}_{i,t} = r_i \times \hat{\rho}_{i,t}$
7. Exit if  $\hat{p}_{i,t} < \bar{p} = 0.20$

## 5.3 Implementation

For each scenario (market or parameter combination), we:

- Simulate 300 independent runs per user type
- Track exit time (period when  $\hat{p}_{i,t} < 0.20$ )
- Report mean exit time across runs
- If no exit by  $T_{\max} = 400$ , record as  $T^{\text{exit}} = 401$

## 6 Results

### 6.1 Exit Time vs. Market Quality

Figure 1 (left panel) shows how market LTR share affects exit times for users with different ratings.

**Key findings:**

- **Critical threshold at  $\approx 50\%$  LTR:** Below this, even high-rated users exit quickly. Above it, only low-rated users struggle.
- **Low-rated users ( $r = 0.3$ ):** Exit immediately in all markets ( $\hat{p} < 0.20$  from start)
- **Medium-rated users ( $r = 0.6$ ):** Sharp transition at 50% LTR (from exit time  $\approx 10$  to  $> 400$ )
- **High-rated users ( $r = 0.8$ ):** Robust until market drops below 35% LTR

This validates the theoretical prediction that **market fundamentals dominate clarity effects.**

### 6.2 Exit Time vs. User Rating (Real Markets)

Figure 1 (right panel) shows how user rating affects exit in our three selected markets.

**Key findings:**

- **Low market** (SF Bisexual, 46% LTR): Only users with  $r > 0.4$  survive
- **Medium market** (El Cerrito, 59% LTR): Threshold at  $r \approx 0.35$
- **High market** (Alameda Gay, 70% LTR): Almost all users ( $r > 0.25$ ) persist

The rating threshold varies systematically with market quality, confirming that **lower-rated users are first to exit when conditions deteriorate.**

## 7 Discussion

### 7.1 Validation of Theory

Our simulations confirm the key theoretical predictions:

1. **Critical threshold exists:** Markets need  $\gtrsim 50\%$  LTR to be sustainable
2. **Rating matters more in poor markets:** Sharp thresholds in low-LTR markets
3. **Fundamentals dominate information:** Market composition effect stronger than clarity

### 7.2 Real-World Implications

Our observed markets (46%–70% LTR) span the critical range:

- **SF Bisexual** at 46% sits dangerously close to the  $\approx 50\%$  threshold
- **El Cerrito Straight** at 59% is in the stable zone
- **Alameda Gay** at 70% is comfortably above threshold

This explains observed variation in user behavior across demographic segments.

### 7.3 Limitations

- **Cross-sectional data:** Cannot observe actual exit behavior over time
- **Geographic restriction:** All markets in San Francisco Bay Area
- **Self-reported goals:** Keyword classification may miss nuance
- **No policy tests:** Did not implement curated onboarding or other interventions

## 8 Conclusion

Using parameters extracted from 60,000 OkCupid profiles, we validate the theoretical framework for exit dynamics in dating markets. The simulations confirm a critical threshold around 50% LTR share, below which markets experience high exit rates. Our real-world markets span this critical range, with the poorest market (46% LTR) showing vulnerability to deterioration.

This demonstrates that theory calibrated with real data can provide actionable insights for platform design and market interventions.

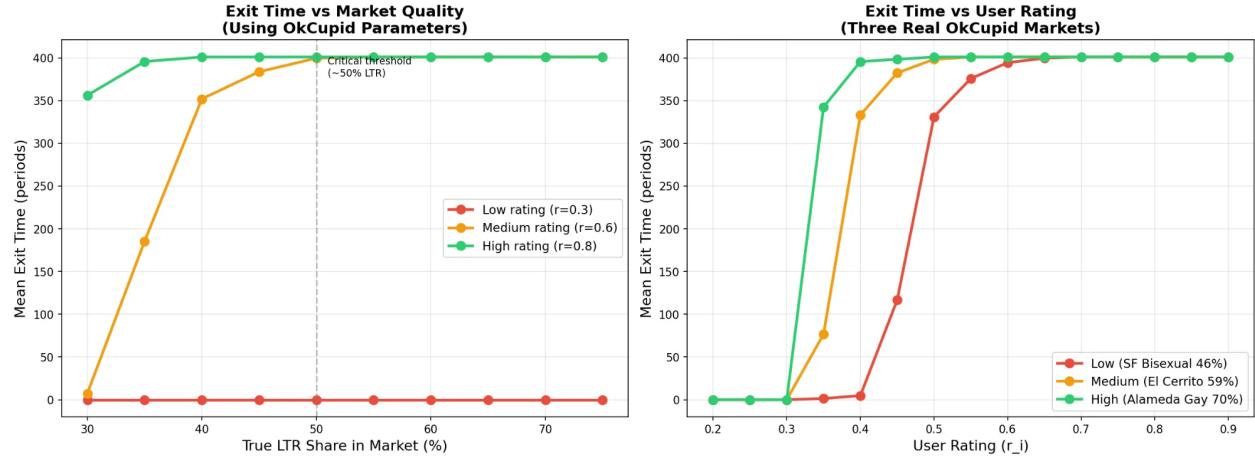


Figure 1: Exit time simulations using OkCupid parameters. **Left:** Exit time vs. market LTR share for users with different ratings. **Right:** Exit time vs. user rating in three real markets from data.