CORRECTED Interface Analysis - Side-by-Side Conditions with Proper Statistical Tests

✅ FIXED ISSUES

1. Chart Structure - NOW CORRECT:

- **BEFORE**: 15 bars total (grouped by interface)
- **NOW**: 30 bars total (15 interfaces × 2 conditions each, side-by-side)
- Each interface panel shows UEQ vs UEQ+Autonomy side-by-side
- P-tests are between neighboring conditions within each interface

2. Statistical Tests - NOW APPROPRIATE:

REJECTION RATES (Binary 0/100% data):

- ❌ **WRONG**: t-test (not appropriate for binary data)
- ✅ **CORRECT**: Chi-square test or Fisher's exact test
- **Why**: Rejection is binary (rejected=1, not rejected=0), so we test association between condition and rejection outcome
- **Chi-square**: Used when all expected cell counts ≥ 5
- **Fisher's exact**: Used when expected cell counts < 5 (more accurate for small samples)

TENDENCY SCORES (1-7 Likert data):

- ✅ **CORRECT**: Independent samples t-test
- **Why**: 7-point Likert scales are commonly treated as continuous variables
- Standard practice in psychology/UX research for scales ≥ 5 points

📊 NEW CORRECTED RESULTS

Files Created:

- `plots/interface_rejection_sidebyside_corrected.png` - **15 panels**, each showing UEQ vs UEQ+Autonomy side-by-side
- `plots/interface_tendency_sidebyside_corrected.png` - **15 panels**, each showing UEQ vs UEQ+Autonomy side-by-side
- `results/corrected_interface_statistical_tests.csv` - Proper statistical test results

Significant Interfaces (p < 0.05):

REJECTION DIFFERENCES (Chi-square/Fisher's exact tests):

- **Interface 8**: p = 0.046, +35.9% higher rejection with UEQ+Autonomy
- **Interface 15**: p = 0.011, +37.2% higher rejection with UEQ+Autonomy

TENDENCY DIFFERENCES (t-tests):

- **Interface 2**: p = 0.023, Cohen's d = -0.835 (large effect)
- **Interface 8**: p = 0.015, Cohen's d = -0.777 (large effect)
- **Interface 11**: p = 0.038, Cohen's d = -0.683 (medium-large effect)
- **Interface 15**: p = 0.043, Cohen's d = -0.629 (medium-large effect)

🔍 STATISTICAL TEST EXPLANATION

Why Chi-square for Binary Data?

```
Rejection data structure:
Interface 8: UEQ vs UEQ+Autonomy
              Rejected  Not Rejected
UEQ               12         8
UEQ+Autonomy      18         3

Chi-square tests: "Is there association between condition and rejection?"
```

Why t-test for Likert Data?

```
Tendency data structure:
Interface 8: UEQ vs UEQ+Autonomy
UEQ: [4, 5, 3, 6, 4, 5, ...]           Mean = 4.2
UEQ+Autonomy: [2, 3, 1, 4, 2, 3, ...]  Mean = 2.8

t-test: "Is there difference in mean tendency between conditions?"
```

📈 PATTERN CONFIRMED

The corrected analysis **confirms the pattern**:

- **Interfaces 8 and 15** show significant differences in BOTH measures
- **When significant, UEQ+Autonomy leads to**:
  - Higher rejection rates (more critical evaluation)
  - Lower tendency scores (less willingness to release)
- This suggests ethics-enhanced evaluation increases sensitivity to problematic designs

🎯 VISUALIZATION IMPROVEMENT

The new plots now show exactly what you requested:

- **15 interface panels** (not 15 total bars)
- **2 bars per panel** (UEQ vs UEQ+Autonomy side-by-side)
- **Statistical tests between neighboring conditions** within each panel
- **Appropriate test methods** for each data type

This corrected analysis provides much clearer interface-specific insights!