Data Quality Assessment Protocol

Overview

This document outlines our systematic approach to identifying potentially problematic participants in the UI Ethics Evaluation study, with particular focus on detecting AI-generated responses and poor data quality.

Quality Assessment Criteria

1. AI Usage Indicators

- **Low variance in text response length**: Participants with very consistent character counts across explanations (variance < 100) may be using AI to generate similar-length responses
- **Extremely long text responses**: Participants with unusually verbose explanations (avg > 300 characters) may be using AI tools that generate detailed responses
- **Tendency-Release Decision Mismatch**: Discrepancies between numerical tendency scores and actual release decisions may indicate inconsistent or automated responding

2. Poor Data Quality Indicators

- **Very low tendency variance**: Participants who give very similar scores across all interfaces (straightlining behavior)
- **Very high tendency variance**: Participants with extreme variation in scores may be clicking randomly
- **Very short text responses**: Minimal explanations (avg < 20 characters) suggest low engagement
- **High character count variance**: Extreme inconsistency in explanation length (variance > 10,000) may indicate variable engagement or mixed response strategies

Flagging Criteria

Participants are flagged for manual review if they meet any of the following criteria:

1. **AI Suspicious**: Low text variance (< 100) OR very long responses (> 300 chars avg)
2. **Poor Quality**: Very low tendency variance (straightlining) OR very high tendency variance (random) OR very short responses (< 20 chars avg)
3. **Inconsistent**: High character variance (> 10,000) OR tendency-release mismatches

Manual Review Process

1. **Automated Flagging**: Script identifies participants meeting the above criteria
2. **Text Extraction**: All text responses (explanations, open feedback, general feedback) extracted for flagged participants
3. **Manual Annotation**: Researcher reviews full text and marks participants as AI_SUSPICIOUS (TRUE/FALSE)
4. **Final Classification**: Based on manual review, participants are classified for potential exclusion

Data Sources

- **Interface Explanations**: Text explanations for tendency scores across all evaluated interfaces

- **Open Feedback**: Any open-ended feedback provided during the study
- **General Feedback**: Overall study feedback comments
- **Demographic/Technical Info**: Browser, device, and completion time data where available

This systematic approach ensures consistent and thorough quality assessment while maintaining transparency in our exclusion criteria.