

Behavioral Experiment Simulation Tool – Methods

Summary

Version 2.1 | Created by Dr. [Eugen Dimant](#)

Overview

The Behavioral Experiment Simulation Tool generates realistic synthetic pilot data for behavioral science experiments. It automates the creation of datasets that mirror real survey response patterns, enabling researchers and students to:

- Test analysis pipelines before data collection
- Practice data cleaning and exclusion procedures
- Verify survey logic and variable coding
- Develop analysis scripts on realistic data structures
- Learn to interpret statistical outputs on data with known properties

This tool is designed for **educational and pilot testing purposes**; it does not replace actual human subjects research but provides a rigorous training ground for developing analytical skills.

Key Features (v2.1)

Automatic Survey Parsing

Upload your Qualtrics QSF file and the tool automatically:

- Detects experimental conditions from survey flow and randomizers
- Identifies Likert scales with their specific point ranges (5-point, 7-point, 10-point, etc.)
- Extracts open-ended question specifications
- Infers factorial design structure (2×2, 2×3, etc.)
- Detects attention checks and manipulation checks

Mandatory Scale Verification

To prevent data-reality mismatches, the tool now requires users to:

- Review all detected scales before simulation
- Verify that scale point ranges match the actual Qualtrics survey

- Explicitly confirm scale configurations (blocking requirement)
- Edit scales if automatic detection was incorrect

Attention & Manipulation Check Management

Users can now:

- Review automatically detected attention checks
- Review detected manipulation checks
- Manually add or edit check questions
- Track which checks will be simulated

Behaviorally Realistic Response Generation

The simulator generates responses that reflect actual human survey behavior:

- **Response styles:** Engaged, satisficing, extreme responding, acquiescence
- **Attention patterns:** Realistic attention check pass/fail rates
- **Individual differences:** Trait-based variation across participants
- **Sentiment alignment:** Open-ended responses consistent with numeric ratings
- **Careless responding:** Realistic patterns of inattention for exclusion practice

Complete Output Package

Each simulation produces:

File	Description
<code>Simulated_Data.csv</code>	Generated dataset with all variables
<code>Study_Summary.md</code>	Summary report in Markdown format
<code>Study_Summary.html</code>	Same summary in browser-viewable HTML
<code>Metadata.json</code>	Simulation parameters and settings
<code>Schema_Validation.json</code>	Data quality checks and validation
<code>R_Prepare_Data.R</code>	Analysis-ready R script
<code>Data_Codebook_Handbook.txt</code>	Complete variable descriptions

Instructor-Only Reports (sent via email, not in student download):

- Comprehensive statistical analysis with visualizations
- T-tests, ANOVA, regression results
- Effect size calculations
- Hypothesis evaluation against pre-registration

Persona-Based Simulation Methodology

Theoretical Foundation

The persona-based approach is grounded in decades of survey methodology research demonstrating that participants systematically differ in how they respond to surveys. Rather than treating all respondents as homogeneous, this tool simulates the heterogeneous response patterns observed in real data.

Core Response Style Personas

The following personas represent well-documented response styles from the survey methodology literature:

Persona	Weight	Theoretical Basis	Key Characteristics
Engaged Responder	30%	Ideal respondent model	High attention ($\mu=0.95$), full scale use, consistent responses
Satisficer	20%	Krosnick (1991) satisficing theory	Midpoint tendency, faster completion, minimal text
Extreme Responder	8%	Greenleaf (1992) ERS literature	Endpoint use (1s and 7s), emphatic text
Acquiescent Responder	7%	Paulhus (1991) yes-saying bias	Agreement tendency regardless of content
Careless Responder	5%	Meade & Craig (2012)	Low attention, fails checks, inconsistent

Trait-Based Individual Differences

Each persona defines probability distributions for behavioral traits. When a participant is assigned a persona, their individual trait values are sampled from these distributions, creating realistic within-persona variation:

Core Traits:

- Attention Level:** Probability of reading carefully (affects attention check performance)
- Response Consistency:** Stability across similar items (affects reliability)
- Scale Use Breadth:** Tendency toward endpoints vs. midpoints
- Acquiescence:** Agreement bias independent of item content
- Social Desirability:** Tendency to present favorably
- Reading Speed:** Affects completion time simulation

Domain-Specific Personas

The tool includes specialized personas activated by keywords in your study description:

Consumer Behavior & Marketing:

- Brand Loyalist (high attachment, low price sensitivity)
- Deal Seeker (high price sensitivity, low brand loyalty)
- Impulse Buyer (low deliberation, high novelty seeking)
- Conscious Consumer (values-driven purchasing)
- Early Adopter (high novelty seeking, tech enthusiasm)

AI & Technology Research:

- Tech Enthusiast (high AI trust, optimistic about automation)
- Tech Skeptic (privacy concerned, cautious about AI)
- AI Pragmatist (balanced, use-case dependent attitudes)
- Privacy-Concerned User (data sensitivity, security focus)

Behavioral Economics:

- Loss-Averse Agent (prospect theory patterns)
- Present-Biased Agent (hyperbolic discounting)
- Rational Deliberator (utility-maximizing patterns)
- Status Quo Defender (endowment effect, inertia)

Organizational Behavior:

- High Performer (engaged, achievement-oriented)
- Disengaged Employee (cynical, low commitment)
- Team Player (prosocial, collaborative)

Social Psychology:

- Prosocial Individual (high empathy, cooperation)
 - Individualist (self-focused, competitive)
 - Conformist (social influence susceptibility)
-

Open-Ended Response Generation

Methodology

Open-ended text responses are generated using a multi-layer approach:

1. **Sentiment Alignment:** Text sentiment matches numeric response patterns

- 2. **Persona Voice:** Language style reflects the assigned persona
- 3. **Topic Relevance:** Responses address the study topic appropriately
- 4. **Natural Variation:** Hedging, fillers, and varied sentence structures

Response Characteristics by Persona

Persona	Verbosity	Detail	Coherence	Example Pattern
Engaged	Moderate	Specific	High	Thoughtful, balanced responses
Satisficer	Minimal	Vague	Low	"It was fine." "OK I guess."
Extreme	Moderate	Emphatic	Moderate	Strong opinions, superlatives
Acquiescent	Moderate	Agreeable	Moderate	Positive framing, agreement
Careless	Minimal	Irrelevant	Very Low	Off-topic, nonsensical

Research Foundations

Core LLM Simulation Research

The use of language models and algorithmic approaches for simulating human survey responses is supported by a growing body of rigorous research:

Foundational Validation Studies:

- 1. **Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023).** Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.
 - DOI: [10.1017/pan.2023.2](https://doi.org/10.1017/pan.2023.2)
 - **Key Finding:** LLMs achieve "algorithmic fidelity"—replicating human response distributions across demographic subgroups on political attitudes and social issues.
- 2. **Horton, J. J. (2023).** Large language models as simulated economic agents: What can we learn from Homo Silicus? *NBER Working Paper* No. 31122.
 - DOI: [10.3386/w31122](https://doi.org/10.3386/w31122)
 - **Key Finding:** LLM-simulated agents exhibit canonical economic behaviors including downward-sloping demand, status quo bias, and anchoring effects.
- 3. **Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023).** Using large language models to simulate multiple humans and replicate human subject studies. *ICML 2023*, PMLR 202:337-371.
 - **Paper:** proceedings.mlr.press/v202/aher23a

- **Key Finding:** "Turing Experiments" successfully replicate classic psychology studies including the Milgram paradigm and ultimatum games.

High-Impact Validation:

4. **Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023).** Generative Agents: Interactive Simulacra of Human Behavior. *ACM UIST*.
 - **DOI:** [10.1145/3586183.3606763](https://doi.org/10.1145/3586183.3606763)
 - **Key Finding:** Stanford/Google research demonstrating that simulated agents can maintain consistent personalities and exhibit believable autonomous behavior over extended interactions.
5. **Binz, M. & Schulz, E. (2023).** Using cognitive psychology to understand GPT-3. *PNAS*, 120(6), e2218523120.
 - **DOI:** [10.1073/pnas.2218523120](https://doi.org/10.1073/pnas.2218523120)
 - **Key Finding:** Systematic comparison showing LLMs exhibit human-like patterns in decision-making, analogical reasoning, and cognitive biases.
6. **Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023).** Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27, 597-600.
 - **DOI:** [10.1016/j.tics.2023.04.008](https://doi.org/10.1016/j.tics.2023.04.008)
 - **Key Finding:** 0.95 correlation between GPT-3.5 and human moral judgments; provides framework for when LLM simulation is appropriate.

Survey Methodology Foundations

The persona system draws on classic survey methodology research:

7. **Krosnick, J. A. (1991).** Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213-236.
 - **Key Concept:** Satisficing theory—respondents may provide "good enough" rather than optimal responses when cognitive demands are high.
8. **Greenleaf, E. A. (1992).** Measuring extreme response style. *Public Opinion Quarterly*, 56(3), 328-351.
 - **Key Concept:** Extreme response style (ERS) is a stable individual difference affecting scale use.
9. **Paulhus, D. L. (1991).** Measurement and control of response bias. In *Measures of Personality and Social Psychological Attitudes* (pp. 17-59).
 - **Key Concept:** Response biases including acquiescence, social desirability, and extreme responding.

10. **Meade, A. W., & Craig, S. B. (2012).** Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437-455.

- **DOI:** [10.1037/a0028085](https://doi.org/10.1037/a0028085)
- **Key Concept:** Methods for detecting and handling careless/inattentive responding in surveys.

On LLM Detection & Survey Validity

11. **Veselovsky, V., Ribeiro, M. H., & West, R. (2023).** Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv:2306.07899*.

- **Key Finding:** Raises important questions about data quality that simulation tools must address.

12. **Westwood, S. J. (2025).** The potential existential threat of large language models to online survey research. *PNAS*, 122(47).

- **DOI:** [10.1073/pnas.2518075122](https://doi.org/10.1073/pnas.2518075122)
- **Key Finding:** LLM-generated responses can evade 99.8% of attention checks; underscores the importance of rigorous simulation standards and realistic careless responding patterns.

Market Research Applications

13. **Brand, J., Israeli, A., & Ngwe, D. (2023).** Using GPT for Market Research. *Harvard Business School Working Paper* 23-062.

- **Paper:** hbs.edu/ris/download.aspx?name=23-062.pdf
- **Key Finding:** LLMs can generate realistic willingness-to-pay estimates and consumer preference data at minimal cost.

Standardization Features

Why Standardization Matters

When multiple teams simulate pilot data for coursework or research training, standardized defaults ensure:

- **Comparable datasets** across teams and projects
- **Consistent exclusion criteria** for fair assessment
- **Reproducible simulation parameters** for verification
- **Fair evaluation** of different analysis approaches

Default Settings (Simple Mode)

Parameter	Default Value	Rationale
Gender distribution	50% male	Balanced representation
Mean age	35 years	Adult population midpoint
Age SD	12 years	Realistic adult variance
Attention check pass rate	95%	Typical online sample quality
Random responder rate	5%	Expected careless rate
Min completion time	60 seconds	Impossibly fast threshold
Max completion time	1800 seconds	30 minute maximum
Straight-line threshold	10 items	Consecutive identical responses

Advanced mode allows full customization of all parameters, including effect size specification for power analysis practice.

Exclusion Criteria Simulation

The tool simulates realistic data quality issues to provide practice with exclusion decisions:

Flag	Description	Detection Method
Flag_Speed	Completion time outside acceptable range	Time < min or > max
Flag_Attention	Failed attention check(s)	Incorrect attention responses
Flag_StraightLine	Excessive straight-line responding	≥ threshold consecutive identical
Flag_Careless	Overall careless responding pattern	Multiple indicators combined
Exclude_Recommended	Recommended for exclusion	Any flag triggered

These flags enable students to practice:

- Applying pre-registered exclusion criteria
- Making defensible exclusion decisions
- Understanding the impact of exclusions on results

Effect Size Simulation (Advanced Mode)

For power analysis training, Advanced Mode allows specification of expected effects:

- **Select dependent variable:** Choose which scale shows the effect
- **Select factor:** Which experimental factor creates the effect
- **Set magnitude:** Cohen's d from 0.0 (null) to 1.5 (very large)
- **Set direction:** Which condition scores higher/lower

This allows students to:

- Understand how effect sizes manifest in real data
 - Practice power analysis interpretation
 - See how sample size affects statistical significance
-

Instructor Analysis Reports

The tool generates comprehensive instructor-only reports (not included in student downloads) containing:

Statistical Analyses

- **Descriptive statistics:** Means, SDs, confidence intervals by condition
- **Primary tests:** Independent samples t-tests, one-way ANOVA
- **Effect sizes:** Cohen's d with confidence intervals
- **Factorial ANOVA:** For multi-factor designs
- **Non-parametric alternatives:** Mann-Whitney U, Kruskal-Wallis

Visualizations

- Distribution plots by condition
- Means comparison charts with error bars
- Effect size forest plots

Pre-registration Alignment

- Extracted hypotheses from uploaded pre-registration
 - Results mapped to registered predictions
 - Badges indicating PRE-REGISTERED vs. ADDITIONAL analyses
-

Reproducibility

- **Unique Run ID:** Each simulation receives a timestamped identifier

- **Simulation Seed:** Deterministic seeding from study metadata
 - **Reproducible Output:** Same seed = identical results
 - **Complete Metadata:** All parameters captured in JSON for replication
-

Technical Requirements

- **Python 3.9+**
 - **Core Dependencies:** streamlit, pandas, numpy
 - **PDF Generation:** reportlab
 - **Email Delivery:** sendgrid (optional)
 - **PDF Parsing:** pypdf (optional, for pre-registration upload)
-

Ethical Considerations

This tool is designed for **educational and pilot testing purposes only**. It should not be used to:

- Fabricate research data for publication
- Misrepresent simulated data as human subjects data
- Bypass IRB requirements for actual research

The tool includes clear labeling in all outputs indicating data is simulated.

Citation

If you use this tool in your research or teaching, please cite:

Dimant, E. (2026). Behavioral Experiment Simulation Tool (Version 2.1). University of Pennsylvania.

Contact & Support

For questions, issues, or feature requests:

- **Website:** <https://eugendimant.github.io/>
- **Email:** edimant@sas.upenn.edu
- **Issues:** [GitHub Repository](#).