Modern Experimental Designs for Business Research

Course Notes for

BTE 693

Computational Methods and Field Experiments

University of Miami

Niam Yaraghi

Version 1.0 · September 2026

Modern Experimental Designs for Business Research
Course Notes for BTE 693: Computational Methods and Field Experiments
University of Miami

Version 1.0, September 2026

© 2026 Niam Yaraghi

Cite as:

Yaraghi, N. (2026). *Modern Experimental Designs for Business Research.* Version 1.0. University of Miami. `https://niamyaraghi.com/experimental-designs`

Preface

This book collects the notes for BTE 693, *Computational Methods and Field Experiments*, a doctoral course on experimental design for business research at the University of Miami. Each chapter corresponds to one session of the course, in the order the material was taught, so the chapters that open with a "Summary of the Previous Session" are picking up exactly where the preceding chapter left off.

The book is organized in four parts.

Part I – Foundations of Experimental Research. Why experiments, and what secondary data cannot do; internal, external and ecological validity and the spectrum of experiment types; and the statistical machinery of hypothesis testing, power, and sample size.

Part II – Optimizing Interventions. The Multiphase Optimization Strategy (MOST), factorial and fractional factorial designs, and Sequential Multiple Assignment Randomized Trials (SMART) for interventions that adapt between stages.

Part III – Adaptive and Real-Time Designs. Just-in-Time Adaptive Interventions (JITAIs) and the Micro-Randomized Trials (MRTs) used to estimate their proximal effects.

Part IV – Preferences, Scale, and New Tools. Conjoint analysis for preference elicitation, the scalability problem and machine-learning estimation of heterogeneous treatment effects, and the use of large language models in designing and running experiments.

A single running example threads through most of the book: a smartphone stress-reduction app for university students, built from three components — mindfulness exercises, physical activity goals, and social support. The same app is used to illustrate a factorial MOST design, a two-stage SMART, a JITAI, a micro-randomized trial, and a choice-based conjoint study, which makes the differences between the designs easier to see than a fresh example for each method would.

Worked analyses are given in SAS. Code listings are numbered per chapter and indexed in the List of Listings at the front of the book. The data behind them are simulated, so the numbers are there to make the mechanics of each procedure concrete rather than to report findings.

Each chapter closes with the required readings assigned for that session of the

course; the full references are collected in the bibliography. Appendix A holds the readings for the course's session on research ethics, which has no lecture notes of its own, together with the CITI training required before any human-subjects study at the University of Miami.

Contents

Preface iii

List of Figures ix

List of Tables xi

List of Listings xiii

I Foundations of Experimental Research 1

1 Foundations: From Secondary Data to Experiments 3
- 1.1 Overview of Secondary Data Analysis and Key Challenges 3
- 1.2 Advantages and Strengths of Experiments 4
- 1.3 Poor Econometrician's Time Machine 5
- 1.4 Reliability and Validity in Empirical Research 6
- 1.5 Key Principles of Experimental Design 7
- 1.6 Scalability in Experiments 7

2 Validity and the Spectrum of Experiment Types 11
- 2.1 Summary of the Previous Session 11
- 2.2 Ecological Validity 11
- 2.3 The Spectrum of Experiment Types based on List (2011) 12
- 2.4 Online Platforms in Experiments (Based on Fréchette et al., 2022) 14
- 2.5 A Simple Experiment 15
- 2.6 Advantages and Disadvantages of Including Multiple Images per Respondent vs. One Image per Respondent 17

3 Hypothesis Testing, Statistical Power, and Sample Size 19
- 3.1 Summary of the Previous Session 19
- 3.2 Overview of Hypothesis Testing, Type I and Type II Errors, and p-value 19
- 3.3 Statistical Power 22

3.4 Sample Size and Power Calculations 24
3.5 Calculate the Requeired Sample Size in Gu et al. 2022 28
3.6 Ethical Considerations in Determination of Sample Size 29

II Optimizing Interventions 33

4 The Multiphase Optimization Strategy and Factorial Designs 35
4.1 Summary of the Previous Session 35
4.2 Introduction to MOST 35
4.3 The Four Desiderata of MOST: EASE 38
4.4 Steps in Optimization 39
4.5 Factorial Designs 41

5 MOST in Practice: A Worked Example and Fractional Factorial Designs 47
5.1 Summary of the Previous Session 47
5.2 Implementing MOST for the Stress-Reduction App 47
5.3 Fractional Factorial Designs 57

6 Sequential Multiple Assignment Randomized Trials (SMART) 59
6.1 Summary of the Previous Session 59
6.2 What is SMART? 59
6.3 SMART Implementation of the Stress Reduction App at University of Miami 60
6.4 Empirical Methods for Analyzing SMART Data 65

III Adaptive and Real-Time Designs 77

7 Just-in-Time Adaptive Interventions (JITAIs) 79
7.1 Summary of the Previous Session 79
7.2 Introduction to JITAIs 79
7.3 Components of JITAIs 80
7.4 Extending the Stress-Reduction App Example with JITAIs 81

8 Micro-Randomized Trials (MRTs) 89
8.1 Summary of the Previous Session 89
8.2 Micro Randomized Trial (MRT) 89
8.3 How Does MRT Work? 90
8.4 Modeling Time-Varying Effects and Endogeneity in MRTs 90

IV Preferences, Scale, and New Tools **95**

9 Conjoint Analysis and Preference Elicitation **97**
9.1 Summary of the Previous Session 97
9.2 Types of Questions Conjoint Analysis Addresses 98
9.3 Foundations of Conjoint Analysis 99
9.4 Implementing Conjoint Analysis in the Stress Reduction App . . . 100

10 Scaling Experiments and Heterogeneous Treatment Effects **109**
10.1 Summary of the Previous Session 109
10.2 Scalability in Experiments 109
10.3 Application of Machine Learning to Estimate Heterogeneous Treatment effects 112

11 Large Language Models in Experimental Research **117**
11.1 Summary of the Previous Session 117
11.2 Why LLMs Matter for Experiments 117
11.3 How LLMs Improve the Experiments 119

A Research Ethics and IRB Training **123**
A.1 Ethical and Practical Challenges in Experimental Research 123
A.2 Registering for Required IRB Training at the University of Miami 124

Bibliography **129**

List of Figures

4.1 Intervention components and causal pathways in a weight loss experiment. 41

5.1 The direct and indirect effects of intervention components on the outcome, along with their interactions 49
5.2 G*Power Sample Size Calculation 53

7.1 Effectivness of Interventions 86

8.1 MRT Flowchart From Yaraghi et al. 2025 90

List of Tables

2.1 Comparison of Internal Validity, External Validity, and Ecological Validity . . . 12
2.2 Validity Across Different Types of Experiments . . . 13

3.1 Comparison of Cohen's d and Risk Difference (RD). . . . 25

4.1 2×2 Factorial Design: Experimental Conditions . . . 42
4.2 An easier way to show Experimental Conditions in a 2×2 Factorial Design . . . 46
4.3 Calculation of Main and Interaction Effects in Factorial Design . . . 46

5.1 2×2×2 Factorial Conditions for the Stress-Reduction App . . . 49
5.2 Overall ANOVA for Response . . . 55
5.3 Type I ANOVA for Main Effects and Interactions . . . 55
5.4 Parameter Estimates from the Model . . . 55

6.1 Baseline Stress by Stage 1 Treatment . . . 61
6.2 T-Test Results for Baseline Stress by Stage 1 Treatment . . . 62
6.3 SMART Paths: Responder Status and Treatment Assignment . . . 63
6.4 Stress at Week 8 by SMART Path . . . 64
6.5 GLM Analysis of Stress at Week 8 by Treatment Path (Reference: A_A) . . . 66
6.6 Contrast Results for GLM Analysis of Stress at Week 8 . . . 66
6.7 LSMeans Estimates for Treatment Paths from PROC GLM . . . 67
6.8 LSMeans Estimates for Treatment Paths from PROC GLM . . . 67
6.9 Analysis of Maximum Likelihood Parameter Estimates . . . 71
6.10 Path Least Squares Means . . . 71
6.11 Stage 2 LSMeans for Non-Responders (C vs. D) . . . 74
6.12 Stage 1 LSMeans (A vs. B) . . . 74

7.1 Penalized Maximum Likelihood Estimates . . . 88
7.2 Penalized Maximum Likelihood Estimates . . . 88

9.1 All Options in Each Choice Set Before Simulation . . . 101

9.2 Unique Profiles Chosen by Respondents 103
9.3 Frequency of Chosen Options Within Each Choice Set 103
9.4 Part-Worth Utility Estimates from PROC PHREG (Reference Level = 1) 107

List of Listings

5.1 SAS Code to Compute Critical F-Value 51
5.2 SAS Code to Generate Synthetic Data 52
5.3 SAS Code to Perform Factorial Analyis 54
6.1 SAS Code to Tabulate the Data and Test Randomization 61
6.2 SAS Code to Show the SMART Treatment Allocations 62
6.3 SAS Code to Caluclate Stress at Week 8 64
6.4 SAS Code to run the GLM model 65
6.5 SAS Code to summerize the optimal strategy 66
6.6 SAS Code to calculate and assign probabilities to each observationl 69
6.7 SAS Code to use logistic regression to calculate probabilities . . . 69
6.8 SAS Code to calculate stabilized IPWs 70
6.9 SAS code to estimate the smart coefficients using stabilized weights 71
6.10 SAS code to implement Q-Learning 73
7.1 SAS code to estimate the effects of interventions for those who responded 86
7.2 SAS code to estimate the intercation effects 87
9.1 SAS code to summerize the actual choices 102
9.2 SAS code examine the chosen profile within each set 103
9.3 SAS code to prepare the data for analysis 104
9.4 SAS code run the regression model 106

Part I

Foundations of Experimental Research

Chapter 1

Foundations: From Secondary Data to Experiments

1.1 Overview of Secondary Data Analysis and Key Challenges

Secondary data analysis leverages existing data sources to infer causal relationships. While it offers scalability and cost efficiency, it is particularly susceptible to biases and relies on strong assumptions for causal inference.

Key Susceptibilities to Bias

1. Omitted Variable Bias

- Definition: Occurs when a confounding variable is not accounted for, leading to misleading estimates of the treatment effect.
- Example: Estimating the effect of an advertising campaign on sales without considering seasonal trends.

2. Selection Bias

- Definition: Arises when the treatment group is not representative of the population due to non-random assignment.
- Example: Comparing outcomes for self-selected participants in a program versus non-participants.

3. Measurement Bias

- Definition: Errors in how variables are recorded or defined, affecting validity.
- Example: Using inaccurate proxies for customer satisfaction in survey data.

Required Assumptions for Causal Inference

Secondary data analysis often requires strong assumptions to mitigate these biases and enable credible causal interpretations.

1. Parallel Trends Assumption (Difference-in-Differences)

- What it Ensures: Ensures that, in the absence of treatment, treated and control groups would follow the same trend over time.
- Relationship with Bias: Helps control for omitted variables that affect both groups equally over time.

2. No Confounding Variables (Interrupted Time Series)

- What it Ensures: Assumes no unmeasured factors simultaneously influence the observed trends and the intervention's timing.
- Relationship with Bias: Minimizes omitted variable bias by isolating the intervention's effect.

3. Exogeneity (Instrumental Variables)

- What it Ensures: Requires that the instrument (variable inducing treatment assignment) affects the outcome only through the treatment.
- Relationship with Bias: Addresses selection bias by mimicking randomization through an exogenous instrument.

1.2 Advantages and Strengths of Experiments

Experiments offer unique advantages that make them a powerful tool for causal inference in research. These strengths include:

1. Random Assignment

- Elimination of Selection Bias: Random assignment ensures that treatment and control groups are statistically equivalent, eliminating systematic differences between the groups. This allows researchers to attribute observed differences in outcomes directly to the treatment.
- Foundation for Causal Inference: By creating comparable groups, random assignment minimizes confounding variables and ensures the validity of causal claims.

2. Controlled Environments

- Manipulation of Variables: Experiments enable researchers to directly manipulate independent variables and observe their impact on outcomes, offering unparalleled flexibility in testing hypotheses.

- Flexibility in Testing Theories: Researchers are not constrained by existing datasets, allowing them to explore a wider range of questions and theories that observational data cannot accommodate.

3. Direct Causal Inference

- Bias-Free Measurement: Experiments measure the treatment effect directly without the biases commonly found in secondary data analysis.
- Simplified Analysis: The absence of biases and the straightforward structure of treatment and control groups often simplify statistical analysis, making results more robust and interpretable.

1.3 Poor Econometrician's Time Machine

The most definitive method to examine the causal effects of a treatment is to use a time machine! Ideally, researchers would treat participants, observe the outcomes, then travel back in time to retrieve the same individuals, this time withholding treatment, and observe the outcomes again. This approach allows the researcher to compare the results from both scenarios. The difference between these outcomes would then represent the pure causal impact of the treatment:

$$E\left[y_i \mid D_i = 1\right] - E\left[y_i \mid D_i = 0\right].$$

We can write the observed outcome as follows:

$$y_i = \begin{cases} y_{1i} & \text{if } D_i = 1, \\ y_{0i} & \text{if } D_i = 0. \end{cases}$$

This allows us to rewrite the outcome as:

$$y_i = y_{0i} + \left(y_{1i} - y_{0i}\right) D_i.$$

Therefore:

$$E\left[y_i \mid D_i = 1\right] = E\left[y_{1i} \mid D_i = 1\right],$$
$$E\left[y_i \mid D_i = 0\right] = E\left[y_{0i} \mid D_i = 0\right].$$

Using this notation, we can rewrite the initial formula as:

$$E\left[y_i \mid D_i = 1\right] - E\left[y_i \mid D_i = 0\right] = E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 0\right].$$

By subtracting and then adding the term $E\left[y_{0i} \mid D_i = 1\right]$, we have:

$$\left\{E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 1\right]\right\} + \left\{E\left[y_{0i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 0\right]\right\}.$$

The first term, $E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 1\right]$, is the average causal treatment effect on the treated (ATT). This shows the difference between the treated outcome

and the counterfactual outcome for the untreated for the same individuals. This assumes that the researcher, equipped with a hypothetical time machine, could travel back in time, retrieve the same individuals, withhold treatment, and observe what would have happened to these experiment participants had they not been treated.

The second term introduces significant complications. It involves the researcher using the time machine to go back and find the individuals who were originally treated, withhold treatment this time, and observe the outcomes. These results are then compared to the actual outcomes of those who were never treated in the first place. If the researcher had any biases in selecting whom to enroll in his experiments, that would show here. If the researcher had any biases in selecting whom to enroll in the experiments, that would show here. That is, the counterfactual outcomes for the treated would be different from the actual outcomes of the untreated individuals. There should ideally be no differences between treated and untreated individuals unrelated to the treatment.

If there were no selection bias, the following equation would hold:

$$E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 0\right] = E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 1\right].$$

That is, the average difference in the outcome would be the average treatment effect on the treated (ATT). This simplifies further as:

$$E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 0\right] = E\left[y_{1i} \mid D_i = 1\right] - E\left[y_{0i} \mid D_i = 1\right]$$

$$= E\left[y_{1i} - y_{0i} \mid D_i = 1\right] = E\left[y_{1i} - y_{0i}\right].$$

The point is, **random assignment of treatment acts as the poor econometrician's time machine!**

1.4 Reliability and Validity in Empirical Research

Reliability and validity are fundamental concepts in econometrics and experimental design, ensuring the robustness of research findings.

- *Reliability* refers to the consistency and stability of measurements. Reliable methods produce the same results under the same conditions. For example:
  - A scale is reliable if it shows the same weight every morning when conditions are unchanged. In business research, reliability ensures that repeating a survey with the same population yields consistent outcomes.
- *Validity* assesses whether a method measures what it is intended to measure. Validity ensures the accuracy of the causal relationships inferred. For example:
  - A scale is valid if it measures weight rather than an unrelated attribute like height. In a study on advertising's impact on sales, validity requires

isolating the causal effect of advertising without interference from factors like seasonality or competitor actions.

Together, reliability and validity ensure that experiments yield meaningful, reproducible results, forming the backbone of empirical research.

1.5 Key Principles of Experimental Design

Experimental design refers to a structured and systematic approach to identifying causal relationships. The goal is to isolate the effect of a treatment or intervention on an outcome of interest, controlling for confounding factors and ensuring that observed effects are attributable solely to the treatment.

- Randomization: Ensuring Validity
  Randomization addresses selection bias and omitted variable bias by evenly distributing both observed and unobserved factors across treatment and control groups. This ensures internal validity by isolating the causal effect of the treatment. Without randomization, differences in outcomes might reflect pre-existing disparities rather than the treatment effect.
- Replication: Ensuring Reliability
  Replication checks for consistency and reliability by repeating experiments with new samples or under different conditions. Reliable findings demonstrate stability and reproducibility, forming the foundation of empirical research.
- Control: Supporting Validity
  Control minimizes measurement bias and allows researchers to directly manipulate variables of interest while holding others constant. This precise manipulation ensures that the observed changes in outcomes can be attributed solely to the treatment, supporting the validity of the findings.

1.6 Scalability in Experiments

Scalability refers to the ability to expand an experimental design to larger populations, different contexts, or real-world settings without compromising the validity, reliability, or interpretability of results. While experiments, particularly randomized controlled trials (RCTs), are a gold standard for causal inference, scaling them up introduces unique challenges.

Key Challenges in Scalability

Representativeness:

- *Definition:* Experimental samples often do not reflect the broader population due to limited resources or constraints in participant recruitment.

- *Problem:* Findings from unrepresentative samples may lack external validity, making it difficult to generalize results to other populations or settings.
- *Example:* A study on an educational intervention conducted only in urban schools may not apply to rural areas.

Voltage Effects:

- *Definition:* Voltage effects occur when the effectiveness of an intervention diminishes as it scales from a controlled experimental setting to real-world applications.
- *Problem:* Factors such as reduced adherence, variations in implementation fidelity, or unforeseen external influences in real-world settings can undermine effectiveness.
- *Example:* A personalized marketing campaign showing strong results in a small pilot may falter when scaled due to less precise targeting or broader audience variability.

Heterogeneous Treatment Effects:

- *Definition:* The impact of an intervention may vary across different subgroups or settings.
- *Problem:* Experiments often identify average treatment effects, which may obscure important differences in how various populations respond.
- *Example:* A job training program may benefit younger participants more than older ones, but this nuance might not be evident without subgroup analyses.

Required Readings

- Omar Al-Ubaydli, Min Sok Lee, John A. List, Claire L. Mackevicius, and Dana Suskind. How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, 5(1):2–49, 2021. doi: 10.1017/bpp.2020.17.
- Yue Katherine Feng, Jennifer L. Claggett, Elena Karahanna, and Kar Yan Tam. A randomized field experiment to explore the impact of herding cues as catalysts for adoption. *MIS Quarterly*, 46(2):1135–1164, 2022. doi: 10.25300/MISQ/2022/16141.
- Francesco Cecchi, Robert Lensink, and Edwin Slingerland. Ambiguity attitudes and demand for weather index insurance with and without a credit bundle: Experimental evidence from Kenya. *Journal of Behavioral and Experimental Finance*, 41:100885, 2024. doi: 10.1016/j.jbef.2023.100885.
- Moran Lazar, Ella Miron-Spektor, Gilad Chen, Brent Goldfarb, Miriam Erez, and Rajshree Agarwal. Forming entrepreneurial teams: Mixing business and friendship to create transactive memory systems for enhanced success. *Academy of Management Journal*, 65(4):1110–1138, 2022. doi: 10.5465/amj.2020.0393.

- Omid Rafieian, Anuj Kapoor, and Amitt Sharma. Multiobjective personalization of marketing interventions. *Marketing Science*, 44(2):457–477, 2025. doi: 10.1287/mksc.2023.0122.
- Niam Yaraghi. An empirical analysis of the financial benefits of health information exchange in emergency departments. *Journal of the American Medical Informatics Association*, 22(6):1169–1172, 2015. doi: 10.1093/jamia/ocv068.

Chapter 2

Validity and the Spectrum of Experiment Types

2.1 Summary of the Previous Session

In the previous session, we discussed the foundational principles of experimental design, focusing on *randomization*, *replication*, and *control* as tools to address biases such as selection bias and omitted variable bias. We highlighted how these principles ensure *internal* validity by isolating causal effects. Additionally, we examined the challenges of scalability and *external* validity, including issues like representativeness, voltage effects, and heterogeneous treatment effects, which complicate the generalization of findings to broader contexts.

2.2 Ecological Validity

Definition: Ecological validity refers to the extent to which the results of an experiment can be generalized to real-world settings. It assesses how closely the experimental conditions, tasks, and participant behaviors mirror those found in the natural environment. High ecological validity ensures that the findings of an experiment are applicable and relevant outside the controlled experimental setup.

Example from the Real World

Consider a psychological study on memory conducted in a laboratory where participants are asked to memorize and recall lists of random words. While the study might yield accurate data on memory processes in controlled settings, its ecological validity is questionable because memorizing lists of random words is far removed from how people use memory in daily life (e.g., remembering grocery lists or recalling important conversations).

A more ecologically valid approach would involve observing participants' memory in a realistic context, such as recalling items they had seen on a simulated online shopping website or during a guided tour.

Example in a Business Context

In a business experiment, suppose a company tests the effectiveness of a new email marketing strategy in a simulated environment where participants are asked to "imagine" they are potential customers and respond to promotional emails. While the results might suggest the strategy is effective, the study lacks ecological validity because participants are aware they are part of an experiment and may behave differently.

A more ecologically valid approach would involve testing the email marketing strategy in a real-world setting by sending actual promotional emails to a segment of the company's customer base and tracking their click-through and purchase behaviors. This real-world test would provide insights into how the strategy performs under natural conditions, accounting for factors like competing emails, time constraints, and real customer motivations.

The Interplay between Internal, External, and Ecological Validity

Table 2.1: Comparison of Internal Validity, External Validity, and Ecological Validity

Validity Type	Focus	Goal	Example
Internal	Ensures the observed effects are due to the treatment or intervention, not confounding variables.	Ensure the study is methodologically sound and free of biases (e.g., selection bias, omitted variable bias).	In an RCT for a new drug, ensuring differences in health outcomes are due to the drug and not other factors.
External	Determines whether the study's findings can generalize to other populations, settings, or times.	Apply findings beyond the specific conditions of the experiment.	A drug trial conducted with young adults might lack generalizability to elderly populations.
Ecological	Assesses how well the experimental conditions represent real-world scenarios.	Ensure the study mirrors the context in which the phenomenon naturally occurs.	Testing an advertising strategy in a controlled lab environment might not reflect real-world market behavior.

2.3 The Spectrum of Experiment Types based on List (2011)

1. **Laboratory Experiments**

- **Explanation:** Conducted in a highly controlled environment where participants are aware they are part of a study. Researchers manipulate variables to establish causal relationships while minimizing external influences.
- **Example:** Testing the impact of time pressure on decision-making by asking business students to complete investment allocation tasks in a lab setting with varying time constraints.

2. **Artefactual Field Experiments**
   - **Explanation:** Similar to laboratory experiments but conducted outside the lab with participants from a relevant population. The setting is still controlled, but the participants are more representative.
   - **Example:** Investigating how professional managers allocate budgets under uncertainty by inviting them to participate in a structured simulation.
3. **Framed Field Experiments**
   - **Explanation:** Conducted in a natural environment with participants who are unaware of being studied, but they are informed about the purpose or context of the experiment. The environment reflects real-world conditions but with some framing by the researchers.
   - **Example:** Studying the effect of incentive structures on employee productivity by informing workers in a firm about a temporary bonus scheme and observing how their output changes.
4. **Natural Field Experiments**
   - **Explanation:** Participants are unaware they are part of an experiment, and the study occurs in their natural environment without framing or artificial interventions. This approach closely mirrors real-world conditions.
   - **Example:** Testing the effect of algorithmic changes on consumer behavior by adjusting the recommendation engine of an e-commerce platform for certain users without their knowledge and observing purchasing patterns.

Table 2.2: Validity Across Different Types of Experiments

Experiment Type	**Internal Validity**	**External Validity**	**Ecological Validity**
Laboratory Experiments	High	Low	Low
Artefactual Field Experiments	Moderate	Moderate	Moderate
Framed Field Experiments	Moderate to High	High	Moderate to High
Natural Field Experiments	Low to Moderate	High	High

2.4 Online Platforms in Experiments (Based on Fréchette et al., 2022)

Online platforms like Amazon Mechanical Turk (MTurk) and Prolific have revolutionized behavioral and experimental economics by providing accessible and cost-effective ways to conduct experiments. These platforms bring unique advantages and challenges:

Advantages

1. *Increased Participant Diversity:* Online platforms allow researchers to access a global pool of participants with diverse demographics. This enhances external validity by ensuring the findings are more representative of broader populations.
    - *Example:* A researcher can test the impact of framing effects on purchasing decisions with participants from different age groups, educational backgrounds, and geographic regions.
2. *Lower Costs and Faster Recruitment:* Online platforms significantly reduce the time and financial resources required for recruitment and data collection compared to traditional methods.
    - *Example:* Running an experiment to assess the effects of ad targeting can be done at a fraction of the cost on MTurk compared to organizing in-person focus groups.

Role of Online Platforms in Enhancing Validity

Internal Validity: Online platforms allow for randomization, a key method for reducing confounding factors, thereby improving internal validity.

- *Example:* Randomly assigning participants to different marketing messages and tracking their engagement removes systematic differences between groups.

External Validity: Platforms like MTurk enable access to diverse and globally distributed samples, improving the representativeness of findings and their generalizability.

- *Example:* A study testing app usability using participants from various countries and demographics enhances the relevance of results.

Ecological Validity: Online experiments conducted on platforms where participants use real-world interfaces (e.g., e-commerce websites or digital ads) can closely mimic natural behaviors.

- *Example:* Observing ad click-through rates in a simulated online shopping environment mirrors real-world consumer decision-making.

Challenges of Online Platforms

Noise and Attention: Participants on online platforms may multitask or engage in minimal-effort behaviors, reducing the quality of data and potentially lowering internal validity.

- *Mitigation:* Employ attention checks or incentivize high-quality participation.

Learning Effects: Frequent MTurk participants might recognize experimental setups, altering their responses and lowering ecological validity.

- *Mitigation:* Use naive participants or design experiments to reduce predictability.

Sampling Bias: While online platforms are diverse, they may not fully represent specific target populations, limiting external validity in certain cases.

- *Mitigation:* Stratified sampling or targeting specific demographics on the platform.

2.5 A Simple Experiment

Objective: To investigate how the label (*Created by AI* vs. *Created by Human*) affects the perceived quality of an image, where all images are actually generated by AI.

Key Details:

- **Images:** All are AI-generated.
- **Labels:** Randomly assign each image one of two labels:
  1. *Created by AI*
  2. *Created by Human*
- **Response Scale:** Participants will rate the quality of the image on a Likert scale (e.g., 1 to 7, where 1 = Very Poor and 7 = Excellent).

Implementation Instructions in Qualtrics

This guide provides step-by-step instructions to set up the AI vs. Human-labeled image experiment using Qualtrics:

1. **Create a New Project:**
   - Log in to Qualtrics and create a new survey project.
   - Name the project (e.g., "AI vs. Human-Labeled Image Experiment").
2. **Set Up Random Assignment:**

- In the *Survey Flow* section, add a *Randomizer* block.
- Configure the Randomizer to assign participants evenly to one of two conditions:
  - *Condition 1:* Images labeled as "AI-created."
  - *Condition 2:* Images labeled as "Human-created."
- Under each condition, create blocks to display the images and questions specific to that label.

3. **Embed the Images:**
   - In each block (Condition 1 and Condition 2), upload or embed the AI-generated images. Use the image upload feature or paste links to hosted images.
   - Add captions for each image, such as "Created by AI" or "Created by a Human," depending on the condition.
4. **Control the Number of Images per Respondent:**
   - *Single Image per Respondent:* Limit the number of questions to one per block and use the *Randomizer* in the Survey Flow to show only one block to each participant.
   - *Multiple Images per Respondent:* Add multiple questions (one for each image) within the assigned block.
5. **Add Survey Questions:**
   - After each image, add evaluation questions such as:
     - "How creative is this image?" (e.g., Likert scale: 1 = Not at all creative, 5 = Very creative).
     - "How likely do you think this image was created by a human?"
   - Use *Skip Logic* or *Display Logic* to ensure questions are displayed appropriately based on the random assignment.
6. **Insert Attention Checks:**
   - Include an attention check question within the survey, such as:
     - "What label was displayed for the image you just evaluated?" (e.g., options: "AI-created," "Human-created").
   - Use Qualtrics' *Validation* features to ensure responses meet quality standards.
7. **Preview and Test the Survey:**
   - Use the *Preview* feature to test the survey flow.
   - Ensure randomization, image display, and question routing work as intended.
8. **Set Up Survey Distribution:**

- Generate a survey link from the *Distributions* tab in Qualtrics.
- Set survey options to limit one response per participant.

9. **Recruit Participants on MTurk:**
   - (Only if you have IRB approval), Post the Qualtrics survey link on Amazon Mechanical Turk (MTurk).
   - Specify participant criteria.
10. **Collect and Export Data:**
    - Monitor responses in the *Data & Analysis* tab.
    - Export the data for statistical analysis.

2.6 Advantages and Disadvantages of Including Multiple Images per Respondent vs. One Image per Respondent

2.6.1 Single Image per Respondent

Advantages:

- **Minimized Learning Effects:** By showing only one image, respondents are less likely to detect patterns or anticipate the experiment's purpose, preserving the ecological validity of their responses.
  - *Example:* If respondents see multiple images labeled as "AI-created," they might suspect that the study is testing perceptions of AI-generated content, potentially biasing their evaluation.
- **Focused Attention:** Respondents can dedicate their full attention to evaluating one image, reducing noise and enhancing the internal validity of their responses.
- **Simplified Analysis:** Each respondent provides a single independent data point, simplifying statistical analyses and avoiding within-subject correlations that might complicate interpretation.

Disadvantages:

- **Larger Sample Size Needed:** Since each respondent contributes only one data point, achieving statistical power requires a larger number of participants, increasing the cost of the study.
- **Limited Context for Comparison:** Respondents may evaluate the single image without a reference point, potentially increasing variability in responses.
  - *Example:* One respondent's "excellent" rating may differ significantly from another's due to a lack of comparative context.

2.6.2 Multiple Images per Respondent

Advantages:

- **Cost Efficiency:** Each respondent provides multiple data points, reducing the number of participants needed to achieve statistical power. This approach can lower costs, especially on platforms like MTurk.
- **Comparative Context:** Respondents can compare images, which may improve the reliability of their ratings by reducing subjective variability.
  - *Example:* A respondent might rate the second image more accurately after evaluating the first, as it provides a reference point for their judgment.

Disadvantages:

- **Increased Risk of Learning Effects:** Respondents might notice experimental patterns, such as consistent labeling of "AI-created," which could bias their subsequent responses and lower ecological validity.
- **Fatigue Effects:** Evaluating multiple images can lead to respondent fatigue, increasing noise in later responses and potentially reducing internal validity.
- **Within-Subject Correlation:** Responses from the same participant are not independent, complicating statistical analysis and requiring more advanced modeling techniques (e.g., mixed-effects models).

Required Readings

- Guillaume R. Fréchette, Kim Sarnoff, and Leeat Yariv. Experimental economics: Past and future. *Annual Review of Economics*, 14(1):777–794, 2022. doi: 10.1146/annurev-economics-081621-124424.
- John A. List. Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25(3):3–16, 2011. doi: 10.1257/jep.25.3.3.
- Niam Yaraghi, Darrell M. West, Ram D. Gopal, and Ram Ramesh. (How) did attack advertisements increase Affordable Care Act enrollments? *PLOS ONE*, 15(2): e0228185, 2020. doi: 10.1371/journal.pone.0228185.

Chapter 3

Hypothesis Testing, Statistical Power, and Sample Size

3.1 Summary of the Previous Session

In the previous session, we explored the use of online platforms like MTurk in experimental research, emphasizing their potential to enhance participant diversity, lower costs, and increase scalability. We discussed key challenges such as response noise, attentiveness, and learning effects, alongside mitigation strategies to address these issues. Additionally, we examined the role of randomization, external validity, and ecological validity in designing robust experiments, with a focus on balancing realism and control across laboratory, artefactual, framed, and natural field experiments. Lastly, we introduced a hands-on experiment design exercise using Qualtrics.

Building on these foundations, this session focuses on sample size and statistical power—key elements in designing robust experiments that are both credible and replicable. Understanding how to determine an appropriate sample size and ensure sufficient power is essential for overcoming challenges related to variability and ensuring reliable results.

3.2 Overview of Hypothesis Testing, Type I and Type II Errors, and p-value

3.2.1 Hypothesis Testing Framework

- **Null Hypothesis (H_0):** Assumes no effect or no difference (e.g., "The intervention has no impact").
- **Alternative Hypothesis (H_1):** Suggests there is an effect or difference (e.g., "The intervention has a significant impact").

3.2.2 Type I and II Errors

Type I and Type II errors are critical concepts in statistical analysis and hypothesis testing. Understanding these errors is essential for designing robust studies and interpreting results accurately.

- *Type I Error (False Positive):* Occurs when the null hypothesis (H_0) is incorrectly rejected, suggesting an effect or difference that does not exist. In other words, it occurs when H_0 is rejected even though it is true. The probability of this error is denoted by the significance level, α.
- *Type II Error (False Negative):* Occurs when the null hypothesis is incorrectly *not rejected*, failing to detect a real effect or difference. In other words, it occurs when H_0 is not rejected even though H_1 is true. The probability of this error is denoted by β.

Simple Life Example: Home Pregnancy Test

- **Hypotheses:**
  - H_0: The person is not pregnant.
  - H_1: The person is pregnant.
- **Type I Error:** The test indicates pregnancy (reject H_0) when the person is not pregnant. This can lead to unnecessary emotional and logistical preparations.
- **Type II Error:** The test fails to detect pregnancy (fail to reject H_0) when the person is actually pregnant, causing missed opportunities for preparation.

3.2.3 The Concept of the p-value

Definition: The p-value represents the probability of obtaining a test statistic as extreme as the observed one, assuming H_0 is true.

Common Misconception: The p-value does not indicate the probability that H_0 is true but rather the compatibility of the data with H_0.

A small p-value indicates that the observed data would be highly unlikely if the null hypothesis (H_0) were true. However, this does **not** mean that the probability of H_0 being true, given the data, is low. Here's why:

A. Direction of the Conditional

- $P(\text{data} \mid H_0)$: The p-value represents the probability of obtaining data as extreme or more extreme than the observed data, *assuming* that H_0 is true.
- $P(H_0 \mid \text{data})$: This is the probability that the null hypothesis is true, *given* the observed data—which is what is often mistakenly inferred from the

p-value.

The key distinction is that the p-value provides $P(\text{data} \mid H_0)$, **not** $P(H_0 \mid \text{data})$. These are fundamentally different, and confusing them is a logical error known as the *transposed conditional* or *prosecutor's fallacy.*

Suppose that in a murder trial, the prosecution presents DNA evidence found at the crime scene and argues:

"The probability of this DNA match occurring by chance if the defendant were innocent is only 1 in 1,000,000. Therefore, the probability that the defendant is innocent is only 1 in 1,000,000."

This reasoning is incorrect because even though the chance of a false match for any one person is very small (1 in 1,000,000), when you apply it to a large population (10 million people), you end up with multiple innocent people who match the DNA. This dramatically reduces the probability that the matched person is actually guilty.

B. Lack of Prior Information

In frequentist statistics, the p-value is calculated without incorporating any prior belief or probability about H_0. It does not account for how plausible H_0 was before the data were observed.

To estimate $P(H_0 \mid \text{data})$, one typically needs a **Bayesian approach**, which involves prior probabilities and uses Bayes' theorem to update beliefs based on the data:

$$P(H_0 \mid \text{data}) = \frac{P(\text{data} \mid H_0)P(H_0)}{P(\text{data})}$$

where:

- $P(H_0)$ is the **prior probability** of H_0, representing its plausibility before observing the data.
- $P(\text{data} \mid H_0)$ is the likelihood of the observed data assuming H_0 is true (this is what the p-value measures).
- $P(\text{data})$ is the overall probability of the data under all possible hypotheses.

Unlike the frequentist approach, Bayesian inference allows us to compute $P(H_0 \mid \text{data})$, which directly answers how probable H_0 is given the data. However, it requires specifying a prior, which can be subjective.

3.2.4 p-value is Not Enough

A high p-value in hypothesis testing can arise from two key scenarios:

1. **The Null Hypothesis is True:** In this case, the observed data aligns well with the assumption of no effect or no difference. The high p-value correctly reflects that there is insufficient evidence to reject the null hypothesis (H_0).

For instance, if a new drug is genuinely no better than the placebo, the p-value will likely be high, and the study correctly fails to reject H_0.

2. **The Study Lacked the Power to Detect a Real Effect:** Even if the alternative hypothesis (H_1) is true, the study may fail to detect the effect if it is underpowered. An underpowered study can result from:
   - **Small Sample Size:** Insufficient data reduces the study's ability to distinguish between the null and alternative hypotheses.
   - **Large Variability in Data:** High variability in measurements obscures the effect, making it harder to achieve statistical significance.
   - **Small Effect Size:** If the actual difference or effect is subtle, a study with inadequate power might not detect it.

   This situation leads to a Type II error, where we fail to reject H_0 despite H_1 being true.

3.3 Statistical Power

3.3.1 Definition and Meaning

Statistical Power is the probability that a test will correctly reject the null hypothesis (H_0) when the alternative hypothesis (H_1) is true. Mathematically, it is expressed as:

$$\text{Power} = 1 - \beta$$

where β is the probability of making a Type II error.

3.3.2 Importance of Statistical Power

- **Minimizing Errors:** High power reduces the likelihood of Type II errors, ensuring that real effects are not overlooked.
- **Efficient Resource Use:** Adequate power prevents wasted resources on inconclusive studies.
- **Reliable Conclusions:** Ensures that detected effects are meaningful and not due to chance.

3.3.3 Factors Affecting Statistical Power

Several factors influence the power of a study:

- **Sample Size:** Larger samples reduce variability and increase power.
- **Effect Size:** Detecting larger differences or effects is easier and requires less power.

- **Significance Level (α):** Lowering α (e.g., from 0.05 to 0.01) reduces Type I errors but requires more power to detect an effect.
- **Variability in Data:** Lower variability increases the study's ability to detect effects.

3.3.4 Simple Example: Flashlight Testing

Imagine you are testing whether a flashlight works, which corresponds to testing the null and alternative hypotheses:

- **Null Hypothesis (H_0):** The flashlight does not work.
- **Alternative Hypothesis (H_1):** The flashlight works.

The role of the battery in this test is analogous to the concept of **statistical power**:

- **High Power (Fully Charged Battery):** A fully charged battery ensures the flashlight will emit light if it works. In this case:
  - If H_1 (the flashlight works) is true, the high power ensures that the flashlight lights up, and we correctly reject H_0.
  - If H_0 (the flashlight does not work) is true, the high power still allows us to correctly conclude that the flashlight is faulty.
- **Low Power (Weak Battery):** A weak or nearly depleted battery reduces the chance of detecting whether the flashlight works, even if H_1 is true. In this case:
  - If H_1 is true (the flashlight works), the weak battery might fail to light the flashlight, leading to a Type II error (failing to reject H_0 when H_1 is true).
  - If H_0 is true (the flashlight does not work), the weak battery might not affect the outcome, as the flashlight would not light up regardless of the battery's state.

Takeaway: The "battery" (statistical power) in this analogy determines how well the test can detect the truth of the alternative hypothesis. A high-power study (like using a fully charged battery) ensures that if the flashlight works (H_1), we will detect it. A low-power study (like using a weak battery) risks failing to detect the truth of H_1, leading to incorrect conclusions.

3.3.5 Business Example: Product Launch Decision

- **Scenario:** A company tests whether a new product improves customer satisfaction.
- **High Power Study:** A sufficiently large sample size and robust design reveal that the product significantly increases satisfaction.

- **Low Power Study:** An underpowered study might fail to detect the improvement, leading the company to wrongly assume the product has no impact.

3.3.6 Ensuring Adequate Power

Researchers can enhance power by:

- Increasing the sample size.
- Reducing variability through better measurement techniques.
- Choosing a suitable significance level and effect size for the study context.

3.4 Sample Size and Power Calculations

3.4.1 Definition

Sample size refers to the minimum number of observations required to achieve *reliable* and *valid* conclusions in a study. Determining the appropriate sample size is critical for:

- Avoiding wasted resources (overestimating sample size).
- Preventing underpowered studies (leading to Type II errors).
- Ensuring statistical validity (by reducing the liklihood of type I and type II errors).

3.4.2 Key Parameters in Sample Size Calculations

1. **Research Question:** Use frameworks like PICO (Population, Intervention, Comparison, Outcome) to formulate research questions.
    - Example (Medical): Testing the efficacy of a new diabetes drug versus a placebo on reducing HbA1c levels on a group of Medicare enrolees.
    - Example (Business): Evaluating the effect of a new loyalty program on customer retention rates.
2. **Null and Alternative Hypotheses:**
    - H_0: No significant difference or effect.
    - H_1: Significant difference or effect.
3. **Significance Level (α):**
    - The probability of a Type I error.
    - Common values: 0.05 (5%), 0.01 (1%).
4. **Power ($1 - \beta$):**
    - The probability of correctly rejecting H_0 when H_1 is true.
    - Common target: 80% or 90%.

Aspect	Cohen's d	Risk Difference (RD)
Definition	Standardized mean difference between two groups, measuring the magnitude of the effect.	Absolute difference in proportions of an outcome between two groups.
Mathematical Formula	$d = \frac{\bar{X}_1 - \bar{X}_2}{s}$, where $\bar{X}_1$ and $\bar{X}_2$ are group means, and s is the pooled standard deviation.	$RD = p_1 - p_2$, where p_1 and p_2 are the proportions of an outcome in the intervention and control groups.
Type of Outcome	Continuous (e.g., test scores, blood pressure).	Binary (e.g., disease presence/absence, treatment success/failure).
Unit	Standardized (in units of standard deviation).	Absolute difference (e.g., percentage or proportion).
Interpretation	$d = 0.2$: Small, $d = 0.5$: Medium, $d = 0.8$: Large effect size.	$RD > 0$: Increased risk in intervention group. $RD < 0$: Decreased risk. $RD = 0$: No difference.
Example Context	Comparing average test scores or blood pressure between two groups.	Comparing infection rates or treatment success between intervention and control groups.

Table 3.1: Comparison of Cohen's d and Risk Difference (RD).

5. **Effect Size (ES):**
   - Magnitude of the difference between groups.
   - Small ES requires larger sample sizes.
   - Example: Cohen's d, Risk Difference (RD) (see Table 3.1 for details).
6. **Variability (Standard Deviation):** Variability refers to the spread or dispersion of data within a population, often measured by the standard deviation (σ). It reflects how much individual data points differ from the population mean.
7. **Margin of Error (d):** The margin of error represents the maximum allowable difference between the true population parameter (e.g., mean or proportion) and the estimate obtained from the sample. It defines the level of precision desired in the results. For example, $d = 5\%$ means the estimate will be within $\pm 5\%$ of the true value with the specified confidence level.

   Relationship Between Variability and Margin of Error (d):
   - Variability and margin of error are inversely related:
     - Higher Variability: Increases the required sample size to achieve a smaller margin of error, as it makes the data less precise.
     - Lower Variability: Reduces the sample size needed to maintain precision, allowing for a tighter margin of error.
   - For a fixed sample size, greater variability results in a wider margin of error, reducing the precision of the estimate.
8. **Dropout Rate:**
   - Adjust sample size for anticipated losses (e.g., 10–20%).

3.4.3 General Formula for Sample Size Calculation

- **Categorical Outcomes (Proportions/Prevalence):**

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot [p_1(1-p_1) + p_2(1-p_2)]}{d^2}$$

  Where:

  - n: Required sample size **per group**.
  - $Z_{1-\alpha/2}$: Z-value corresponding to the desired confidence level (e.g., 1.96 for 95% confidence).
  - $Z_{1-\beta}$: Z-value corresponding to the desired power (e.g., 0.84 for 80% power).
  - p_1: Proportion of success in the first group (e.g., control group).
  - p_2: Proportion of success in the second group (e.g., intervention group).
  - d: Minimum detectable difference between proportions ($|p_1 - p_2|$).

- **Continuous Outcomes (Means):**

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot \sigma^2}{d^2}$$

  Where:

  - n: Required sample size **per group**.
  - $Z_{1-\alpha/2}$: Z-value corresponding to the desired confidence level (e.g., 1.96 for 95% confidence).
  - $Z_{1-\beta}$: Z-value corresponding to the desired power (e.g., 0.84 for 80% power).
  - σ: Standard deviation of the population (a measure of variability in the data).
  - d: Minimum detectable difference or effect size (the margin of error or precision desired).

  Note that this formula is very similar to the one given in Columb and Atkinson (2016).

$$n > \frac{2K\sigma^2}{d^2}$$

  Where:

  - K is a constant determined by the desired confidence level (α) and power ($1 - \beta$).
  - Typically, for $\alpha = 0.05$ (two-tailed) and $\beta = 0.20$ (80% power), $K = 7.9$.
  - This form is useful for quick approximations.

Simplified Formula:

$$n > \frac{16}{(\text{standardized difference})^2}$$

Where:

- standardized difference $= \frac{d}{\sigma}$.
- If $d = \sigma$, then $n = 16$ per group.
- This formula assumes $\alpha = 0.05$ (two-tailed) and $1 - \beta = 0.80$.

3.4.4 Examples

Categorical Outcome – Free Shipping and Purchase Rate

Scenario: A business wants to test whether offering free shipping increases the proportion of customers who make a purchase compared to no free shipping. The goal is to calculate the required sample size to detect a meaningful difference in purchase rates between the two groups.

Inputs:

- Proportion in the control group (p_1): $p_1 = 0.1$ (10% of customers make a purchase without free shipping).
- Proportion in the intervention group (p_2): $p_2 = 0.15$ (15% of customers are expected to make a purchase with free shipping).
- Margin of error (d): $d = |p_2 - p_1| = 0.05$ (minimum detectable difference in purchase rates).
- Confidence level ($1 - \alpha$): 95% ($Z_{1-\alpha/2} = 1.96$).
- Power ($1 - \beta$): 80% ($Z_{1-\beta} = 0.84$).

Formula:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot [p_1(1-p_1) + p_2(1-p_2)]}{d^2}$$

Calculation:

$$n = \frac{(2.8)^2 \cdot 0.2175}{0.0025} = \frac{7.84 \cdot 0.2175}{0.0025} = \frac{1.7052}{0.0025} \approx 683$$

Adjustment for Dropout: Assume a 10% dropout rate:

$$n^* = \frac{n}{1 - \text{Dropout Rate}} = \frac{683}{0.9} \approx 759$$

Conclusion: The study requires 759 participants per group (1518 total) to detect a 5% increase in purchase rates due to free shipping, with 80% power and a 95% confidence level, accounting for a 10% dropout rate.

Continuous Outcome – Blood Pressure Reduction

Scenario: A clinical trial tests whether a new drug lowers systolic blood pressure (SBP) compared to a placebo. The required sample size needs to be calculated to detect a meaningful reduction in SBP.

Inputs:

- Expected difference in means (d): $d = 10$ mmHg (the drug is hypothesized to reduce SBP by 10 mmHg).
- Standard deviation (σ): $\sigma = 15$ mmHg (based on prior studies).
- Confidence level ($1 - \alpha$): 95% ($Z_{1-\alpha/2} = 1.96$).
- Power ($1 - \beta$): 80% ($Z_{1-\beta} = 0.84$).

Formula:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot \sigma^2}{d^2}$$

Calculation:

$$n = \frac{(1.96 + 0.84)^2 \cdot (15)^2}{(10)^2} \approx 18$$

Adjustment for Dropout: Assume a 10% dropout rate:

$$n^* = \frac{n}{1 - \text{Dropout Rate}} = \frac{18}{0.9} = 20$$

Conclusion: The study requires 20 participants per group (40 total) to detect a 10 mmHg reduction in systolic blood pressure with 80% power and a 95% confidence level, accounting for a 10% dropout rate.

3.5 Calculate the Requeired Sample Size in Gu et al. 2022

(Zhuojun Gu, Ravi Bapna, Jason Chan, and Alok Gupta. Measuring the impact of crowdsourcing features on mobile app user engagement and retention: A randomized field experiment. Management Science, 68(2):1297–1329, 2022)

This paper examines how crowdsourcing features in mobile apps impact user engagement and retention. The experiment involved four randomized groups of app users:

- Control Group: No crowdsourcing features enabled.
- Submission-Only Group: Users could submit content but not access it.
- Access-Only Group: Users could view community-generated content but not contribute.
- Full-Crowdsourcing Group: Users could both submit and access content.

Each group consisted of approximately 750 users, totaling 3,056 participants.

Sample Size Calculations For User Engagement (Continuous Outcome)

Inputs:

- Significance level $(1-\alpha)$: 95% ($Z_{1-\alpha/2} = 1.96$).
- Power $(1-\beta)$: 90% ($Z_{1-\beta} = 1.28$).
- Effect size (d): 0.12.
- Standard deviation (σ): Assume $\sigma = 1$ (This assumption works under the context of effect sizes that are expressed in standardized units, like Cohen's d, where σ is typically normalized to 1.).

Formula:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot \sigma^2}{d^2}$$

Calculation:

$$n = \frac{(1.96 + 1.28)^2 \cdot 1^2}{(0.12)^2} \approx 729 \text{ participants per group.}$$

3.6 Ethical Considerations in Determination of Sample Size

While we have to consider theoretical and statistical justifications to determine sample sizes, in practice, we may deviate from those calculations. Also, more is not always better when it comes to number of participants.

Underpowered Studies:

- An underpowered study has a high probability of *Type II errors*, meaning it may fail to detect a true effect.
- In clinical research, this could mean that a potentially life-saving treatment is dismissed due to lack of statistical significance.
- Ethical concerns arise because participants are exposed to risks and burdens without generating meaningful scientific knowledge.

- **Example:** A cancer drug trial with inadequate sample size may conclude "no significant benefit," even if a larger study would have found an effect.

Overpowered Studies:

- Increasing sample size reduces Type II errors, but excessive sample sizes raise concerns about unnecessary resource use and ethical considerations.
- In clinical trials, enrolling more participants than needed exposes them to potential risks without justification.
- Overpowered studies can detect statistically significant but clinically meaningless differences, leading to overinterpretation.
- **Example:** A study finds a new drug reduces blood pressure by 0.5 mmHg with $p < 0.001$. While statistically significant, the effect size is too small to be clinically meaningful.

Balancing Feasibility Constraints and Statistical Rigor

Researchers face *budget, time, and recruitment challenges*, making it impractical to always aim for high statistical power. Strategies for balance include:

- Using **pilot studies** to estimate variance and optimize sample size.
- Considering **adaptive designs** where sample size can be adjusted mid-study.
- Leveraging **meta-analysis** to aggregate findings from multiple smaller studies.
- **Pre-registering effect sizes** to prevent unnecessary increases in sample size.

When Would a Researcher Choose Lower Power (e.g., $< 80\%$)?

- *Rare diseases or hard-to-recruit populations* (e.g., genetic disorders, high-risk trials) Recruiting a large sample is infeasible, so researchers may accept lower power (e.g., 50-70%) to gain at least some evidence rather than none.
- *Exploratory or early-phase studies* Small sample sizes are acceptable when the goal is hypothesis generation rather than definitive conclusions.
- *When effect sizes are expected to be large* If prior evidence strongly suggests a large effect, researchers may accept lower power because even a small study is likely to detect it.

Required Readings

- M. O. Columb and M. S. Atkinson. Statistical analysis: Sample size and power estimations. *BJA Education*, 16(5):159–161, 2016. doi: 10.1093/bjaed/mkv034.

- Reena Bhardwaj, Usha Agrawal, Praveen Vashist, and Souvik Manna. Determination of sample size for various study designs in medical research: A practical primer. *Journal of Family Medicine and Primary Care*, 13(7):2555–2561, 2024. doi: 10.4103/jfmpc.jfmpc_1675_23.
- Zhuojun Gu, Ravi Bapna, Jason Chan, and Alok Gupta. Measuring the impact of crowdsourcing features on mobile app user engagement and retention: A randomized field experiment. *Management Science*, 68(2):1297–1329, 2022. doi: 10.1287/mnsc.2020.3943.

Part II

Optimizing Interventions

Chapter 4

The Multiphase Optimization Strategy and Factorial Designs

4.1 Summary of the Previous Session

In the previous session, we examined essential components of experimental design, particularly focusing on sample size determination and statistical power. These elements ensure that studies are designed with sufficient observations to detect meaningful effects while minimizing errors. This approach emphasized refining methodologies for analyzing experimental results and ensuring reliability through robust statistical practices.

In this session, we shift our focus to optimizing the intervention itself. The Multiphase Optimization Strategy (MOST) introduces a systematic framework for developing, optimizing, and evaluating interventions by iteratively testing individual components and their interactions. Unlike traditional randomized controlled trials (RCTs), which primarily assess the overall efficacy of an intervention, MOST emphasizes identifying the most effective, efficient, and scalable components of an intervention. By bridging these perspectives, we move from refining how we evaluate interventions to understanding how to design them for maximum impact in real-world applications.

4.2 Introduction to MOST

Suppose we designed an app for University of Miami students to reduce stress, but after running a randomized controlled trial (RCT), the results were weaker than expected. What are some reasons this might have happened?

- *Poor Fit of Components:* Some components of the intervention might not align well with the target audience's needs or preferences.
- *Lack of Engagement:* Users may not find the app engaging enough to utilize

its features consistently.
- *Unanticipated Interactions Between Components:* Certain components may interact in unexpected ways, either amplifying or offsetting each other's effects.

4.2.1 Key Conceptual Distinction Between MOST & RCT

- *RCTs:* Focus on testing the overall effectiveness of an intervention as a complete package *after* it has been implemented.
- *Multiphase Optimization Strategy (MOST):* Emphasizes *developing and optimizing* individual intervention components *before* testing the entire intervention.

4.2.2 Definition of a Component in MOST

In the *Multiphase Optimization Strategy (MOST)* framework, a *component* is defined as a distinct, modular element of an intervention that can be systematically tested, refined, or optimized to improve the overall effectiveness, efficiency, scalability, or affordability of the intervention.

Key Characteristics of a Component in MOST

1. **Independently Manipulable** – A component can be included, excluded, or modified without altering other components of the intervention.
2. **Functionally Meaningful** – Each component serves a specific purpose within the intervention (e.g., a reminder system in a behavioral health intervention, an incentive structure in an educational program).
3. **Optimizable** – Researchers can adjust the component's intensity, frequency, delivery method, or other characteristics to determine its most effective configuration.
4. **Evaluated for Contribution** – MOST evaluates whether each component meaningfully contributes to the intervention's overall effectiveness and cost-effectiveness.
5. **Can Interact with Other Components** – The effectiveness of one component may depend on the presence or absence of others, requiring factorial or other experimental designs to assess interactions.

4.2.3 Intervention Optimization

In MOST, the goal is to systematically test and optimize each component before assembling the final intervention package to ensure the most efficient and scalable intervention. In other words, intervention optimization is the process of identifying the intervention that provides the best expected outcome while adhering to key constraints. The first step towards optimizing the intervention is to define its

components, and identify their effects, costs, and interactions. Let's do these step by step. First, suppose in our example of the stress reduction app at University of Miami, we identify the following components:

- *Mindfulness Exercises*: Guided meditations and breathing exercises to manage stress in real time.
- *Physical Activity Goals*: Sets and tracks daily exercise goals with movement reminders.
- *Social Support Features*: Provides an anonymous platform for peer support, encouragement, and advice.

In the second step, we have to determine the effects of each component, as they each target a specific stress factor.

- *Mindfulness Exercises*: Immediate stress reduction during high-stress periods.
- *Physical Activity Goals*: Lower cortisol levels and long-term resilience against stress.
- *Social Support Features*: Reduces isolation and fosters connections for mental well-being.

After identifying the effects, we should identify the costs associated with each component. This could be for development, and or implementation of the component.

- *Mindfulness Exercises*: High cost (e.g., licensing and creating high-quality audio guides).
- *Physical Activity Goals*: Low-medium cost (e.g., smartphone-based step tracking).
- *Social Support Features*: High cost (e.g., moderation and secure servers).

The last step is to consider how these component interaction with eachother.

- *Mindfulness & Physical Activity Goals*: Regular mindfulness practices can enhance motivation for physical activity by reducing stress and improving focus.
- *Social Support & Physical Activity Goals*: Social accountability enhances adherence to physical activity goals, making users more likely to achieve their exercise objectives.
- *Mindfulness & Social Support*: Peer support can reduce the immediate need for mindfulness, as students feel less isolated and more connected.

4.3 The Four Desiderata of MOST: EASE

1. Effectiveness

- *Definition:* Effectiveness measures how well each intervention component achieves the intended outcomes. It is the degree to which the anticipated effects are realized.
- *Example:* In the stress reduction app, effectiveness would mean that students who engage with features like mindfulness exercises or physical activity goals report measurable stress relief. For high effectiveness, each component should reliably contribute to reduced stress as intended.

2. Affordability

- *Definition:* Affordability (previously referred to as economy in earlier publications) assesses whether the intervention can deliver the expected outcomes while staying within budget constraints. It emphasizes optimizing costs without compromising the quality of outcomes.
- *Example:* With a fixed budget for the app's development, affordability would involve choosing cost-effective components that still produce measurable stress relief. For instance, mindfulness exercises might offer high affordability if they are inexpensive to implement and significantly reduce stress.

3. Scalability

- *Definition:* Scalability is the capacity of the intervention to be implemented on a larger scale, effectively maintaining both affordability and efficiency.
- *Example:* In the app's context, scalability would mean that if the app successfully reduces stress among Miami students, it could be deployed to students at other universities with minimal additional cost or structural change. Features like automated mindfulness guides are more scalable than high-cost, personalized social support that may require extensive resources to expand.

4. Efficiency

- *Definition:* Efficiency is the relationship between outcomes achieved and resources used, essentially the intervention's productivity. It examines how much "output" (stress reduction) is gained per unit of input (e.g., time, cost, or developer effort).
- *Example:* For the app, efficiency could involve examining if the physical activity goals feature provides substantial stress reduction for a relatively low development cost compared to the high cost of social support features.

Efficient components are those that provide significant stress relief without excessive resource consumption.

With MOST, these objectives guide you to design an intervention that not only reduces stress (*effectiveness*) but also does so with minimal resources (*efficiency*), within budget (*affordable*), and is feasible to roll out on a larger scale (*scalability*).

Optimization Objective is to balance the four components of intervention EASE:

- *Effectiveness*: Achieving weight reduction.
- *Affordability*: Staying within budget constraints.
- *Scalability*: Expanding to larger populations with minimal adjustments.
- *Efficiency*: Maximizing outcomes relative to resources used.

Basiclly, which of the components or their combination should be used in the final intervention to maximize the effects while remaining within budget?

4.4 Steps in Optimization

Phase 1: Component Testing

- Evaluate each component's effectiveness in targeting upstream factors and intermediary goals.
- **Example:** Compare mindfulness exercises to social support features in reducing stress levels.

Phase 2: Interaction Analysis

- Analyze how combinations of components perform relative to **EASE** (Effectiveness, Affordability, Scalability, Efficiency) metrics.
- **Example:** Does combining physical activity goals with mindfulness exercises yield better stress reduction outcomes?

Phase 3: Optimization

- Identify the combination of components that best balances **effectiveness, affordability, scalability, and efficiency**.

Phase 4: Full RCT

- Validate the optimized intervention through an **RCT** to ensure it meets **EASE** criteria.

4.4.1 Component Costs and Effects

Mindfulness Exercises

- **Cost:** \$15 per participant.
- **Effectiveness:** Moderately reduces perceived stress and anxiety.
- **Scalability:** High scalability through guided meditation recordings or app-based interventions.

Physical Activity Goals

- **Cost:** \$30 per participant.
- **Effectiveness:** Strongly improves mood and reduces stress through regular movement.
- **Scalability:** Moderate scalability; automated goal tracking can be implemented, but sustained user engagement is required.

Social Support Features

- **Cost:** \$25 per participant.
- **Effectiveness:** Enhances emotional resilience and reduces stress through peer interactions.
- **Scalability:** High scalability when using community-based or AI-driven support features.

Example Outcome of Optimization

1. **Mindfulness Exercises** (\$15) + **Physical Activity Goals** (\$30) + **Social Support** (\$25):
   Total Cost: \$70 per participant.
   Effectiveness: High; addresses stress reduction through cognitive, behavioral, and social mechanisms.
2. **Mindfulness Exercises** (\$15) + **Gamification** (\$20) + **Social Support** (\$25):
   Total Cost: \$60 per participant.
   Effectiveness: Moderate; enhances adherence and engagement with stress management tools.
3. **Physical Activity Goals** (\$30) + **Data Visualization** (\$10) + **Social Support** (\$25):
   Total Cost: \$65 per participant.
   Effectiveness: Moderate; combines stress-relieving movement with self-monitoring and social accountability.

Selected Subset

Mindfulness Exercises + Physical Activity Goals + Social Support Features

Cost: $70 per participant.
Effectiveness: Maximized within budget, combining cognitive, behavioral, and social approaches to stress reduction.

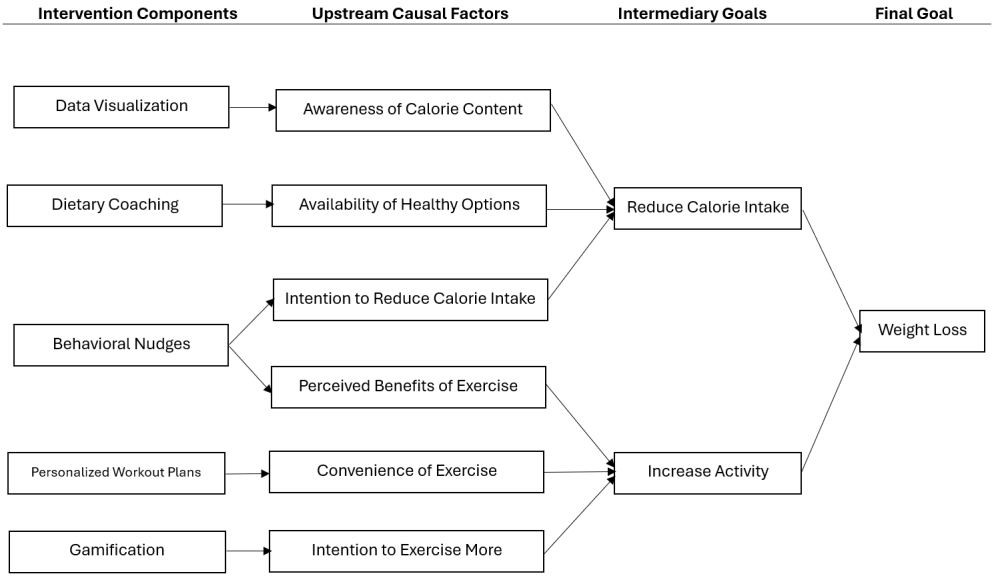

Figure 4.1: Intervention components and causal pathways in a weight loss experiment.

4.5 Factorial Designs

One of the experimental designs well-suited for optimization trials under the Multiphase Optimization Strategy (MOST) framework is the factorial design , a method introduced nearly a century ago by Sir Ronald A. Fisher, who was later knighted for his monumental contributions to the field of statistics. Factorial designs systematically test multiple components (factors) of an intervention and their interactions, providing an efficient and robust way to evaluate the effects of various intervention components.

A simple example of a 2×2 factorial design involves two factors, A and B , each having two levels: "Yes" (component included) or "No" (component excluded). This setup results in four experimental conditions , each representing a different combination of factor inclusion or exclusion:

	Factor B	
Factor A	**No (Excluded)**	**Yes (Included)**
No (Excluded)	μ_1	μ_3
Yes (Included)	μ_2	μ_4

Table 4.1: 2×2 Factorial Design: Experimental Conditions

Factorial designs allow researchers to estimate three key types of effects: main effects of individual factors and the interaction effect between them.

4.5.1 Main Effects

A main effect represents the average impact of a factor on the outcome, independent of the presence of the other factor.

1. Main Effect of Factor A (ME_A) To calculate the effect of Factor A , we compare conditions where A is included (μ_2, μ_4) versus conditions where A is excluded (μ_1, μ_3):

$$ME_A = \frac{\mu_2 + \mu_4}{2} - \frac{\mu_1 + \mu_3}{2}$$

This formula computes the average difference in outcomes between the inclusion and exclusion of Factor A , providing insight into whether adding this intervention component has a significant impact.

2. Main Effect of Factor B (ME_B) Similarly, the effect of Factor B is calculated by comparing conditions where B is included (μ_3, μ_4) versus conditions where B is excluded (μ_1, μ_2):

$$ME_B = \frac{\mu_3 + \mu_4}{2} - \frac{\mu_1 + \mu_2}{2}$$

This measures the independent contribution of Factor B to the outcome.

4.5.2 Interaction Effects

The interaction effect ($IE_{A\times B}$) measures whether the combined presence of Factor A and Factor B produces an effect that is greater or less than the sum of their individual effects . This is crucial because some intervention components may work synergistically , meaning their joint effect is greater than expected , or they may interfere with each other, reducing their effectiveness when combined.

The interaction effect is calculated as:

$$IE_{A\times B} = \frac{\mu_4 - \mu_3 - \mu_2 + \mu_1}{2}$$

This measures whether the combined influence of A and B deviates from what would be expected if their effects were simply additive. A positive interaction

effect suggests that the two components enhance each other , while a negative interaction effect suggests they diminish each other's effectiveness .

4.5.3 Interpretation & Explanation of Effects

Computing the Change in Factor B's Effect Due to Factor A

To see how Factor A influences the outcome when B is absent, we compute:

$$\mu_2 - \mu_1$$

Similarly, to see how Factor A influences the outcome when B is present, we compute:

$$\mu_4 - \mu_3$$

If there were no interaction, these two differences should be the same, meaning Factor A's effect should be the same regardless of whether B is present or absent.

Computing the Change in Factor A's Effect Due to Factor B

Now, we check if Factor B changes the effect of Factor A by taking the difference:

$$(\mu_4 - \mu_3) - (\mu_2 - \mu_1)$$

This measures how much the presence of B modifies the effect of A.

Averaging the Effect Across Both Levels

Since interaction effects represent a combined deviation, we take the average difference by dividing by 2:

$$IE_{A\times B} = \frac{(\mu_4 - \mu_3) - (\mu_2 - \mu_1)}{2}$$

Expanding the parentheses, we get the final formula:

$$IE_{A\times B} = \frac{\mu_4 - \mu_3 - \mu_2 + \mu_1}{2}$$

Intuitive Explanation

Think of Factor A as a dietary intervention (e.g., reduced sugar) and Factor B as an exercise program. The interaction effect tells us whether the benefit of dieting depends on whether the person exercises:

- If dieting helps people lose weight regardless of exercise, then there is no interaction.
- If dieting works much better when paired with exercise than we would expect based on their individual effects, then there is a positive interaction.

- If dieting actually becomes less effective when combined with exercise, then there is a negative interaction.

By using the formula, we capture whether the combined presence of A and B produces effects that are greater or less than expected based on their independent effects.

Required Readings

- Chapters 1–4 of Linda M. Collins. *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*. Statistics for Social and Behavioral Sciences. Springer, Cham, Switzerland, 2018. ISBN 978-3-319-72205-4. doi: 10.1007/978-3-319-72206-1.
- Ryan R. Landoll, Sara E. Vargas, Kristen B. Samardzic, Madison F. Clark, and Kate Guastaferro. The preparation phase in the multiphase optimization strategy (MOST): A systematic review and introduction of a reporting checklist. *Translational Behavioral Medicine*, 12(2):291–303, 2022. doi: 10.1093/tbm/ibab146.

Aspect	Optimization Trial	Evaluation Trial
Purpose	Identifies which components of the intervention work best and under what conditions.	Assesses the overall effectiveness of the finalized intervention package.
Stage in MOST Framework	Conducted during the optimization phase to refine and select intervention components.	Conducted during the evaluation phase after the intervention has been fully designed.
Focus	Builds knowledge about: • Effectiveness of individual components. • Interactions between components. • For whom and under what conditions the components are effective.	Tests whether the combined intervention package achieves its intended outcomes (e.g., weight reduction) in the target population.
Design	Experimental design that systematically manipulates and tests combinations of components.	A randomized controlled trial (RCT) or quasi-experimental design to measure the intervention's overall impact.
Outcome	Provides data to optimize the intervention package by balancing Effectiveness, Affordability, Scalability, and Efficiency (EASE).	Determines the overall effectiveness of the optimized intervention in achieving its primary outcome.
Iterative Process	Iterative; may involve testing multiple configurations of components.	Final; assesses the intervention as a fixed, fully-developed package.
Example Context	In the Weight Management Platform: Testing which components (e.g., Behavioral Nudges, Gamification, Dietary Coaching) or combinations work best within the $100 cost constraint.	In the Weight Management Platform: Testing whether the selected subset of components (e.g., Behavioral Nudges + Dietary Coaching + Gamification) reduces weight overall.

Factor A	**Factor B**	**Outcome**
Excluded (No)	Excluded (No)	μ_1
Included (Yes)	Excluded (No)	μ_2
Excluded (No)	Included (Yes)	μ_3
Included (Yes)	Included (Yes)	μ_4

Table 4.2: An easier way to show Experimental Conditions in a 2×2 Factorial Design

Effect	**Calculation Formula**
Main Effect of A (ME_A)	$\frac{(\mu_2+\mu_4)}{2} - \frac{(\mu_1+\mu_3)}{2}$
Main Effect of B (ME_B)	$\frac{(\mu_3+\mu_4)}{2} - \frac{(\mu_1+\mu_2)}{2}$
Interaction Effect ($IE_{A\times B}$)	$\frac{(\mu_4-\mu_3-\mu_2+\mu_1)}{2}$

Table 4.3: Calculation of Main and Interaction Effects in Factorial Design

Chapter 5

MOST in Practice: A Worked Example and Fractional Factorial Designs

5.1 Summary of the Previous Session

In our last session, we introduced the Multiphase Optimization Strategy (MOST) and explored how it differs from traditional randomized controlled trials (RCTs). While RCTs evaluate an intervention as a whole, MOST focuses on optimizing individual components before assembling the final intervention package. Using the example of a stress-reduction app for college students, we identified its three key components—mindfulness exercises, physical activity goals, and social support features—and examined their individual effects, costs, and interactions. We also introduced the four guiding principles of intervention optimization: Effectiveness, Affordability, Scalability, and Efficiency (EASE).

In today's session, we will focus on how to implement MOST principles in practice. Continuing with the stress-reduction app example, we will design an experiment to systematically test different combinations of intervention components, apply factorial design as a tool to analyze their effects, and determine the most effective, efficient, and scalable intervention. By the end of this session, we will have a clear understanding of how to use MOST to guide intervention development and optimization before moving to full-scale evaluation. Let's get started!

5.2 Implementing MOST for the Stress-Reduction App

This example demonstrates how to apply the Multiphase Optimization Strategy (MOST) to develop and refine an intervention aimed at reducing stress among college students. The intervention comprises three components:

- *Mindfulness Exercises* – Cost: \$15 per participant *Teaches meditation and breathing techniques to regulate emotions and reduce stress.*
- *Physical Activity Goals* – Cost: \$30 per participant *Uses endorphin release and improved well-being through regular exercise to lower stress.*
- *Social Support Systems* – Cost: \$25 per participant *Builds relationships and reduces isolation, which in turn buffers stress.*

5.2.1 Preparation Phase

The *Preparation Phase* establishes the intervention's theoretical framework, components, and specific optimization objectives.

A. Conceptual Model Each component addresses stress via distinct mechanisms:

- *Mindfulness*: Reduces acute stress responses.
- *Physical Activity*: Improves long-term resilience and mood.
- *Social Support*: Mitigates feelings of isolation and fosters peer encouragement.

B. Optimization Objective We aim to **maximize stress reduction** while keeping the per-participant cost at or below **\$50**. Including all three components would cost \$70 (\$15 + \$30 + \$25), so we must determine the most cost-effective combination.

C. Hypotheses

1. Mindfulness alone may yield moderate stress reduction, but combining it with Physical Activity or Social Support could *enhance* effectiveness.
2. Physical Activity and Social Support may have *interactive effects*, possibly increasing motivation and adherence.
3. There may be *diminishing returns* when adding multiple components, requiring careful cost-effectiveness analysis.

5.2.2 Optimization Phase

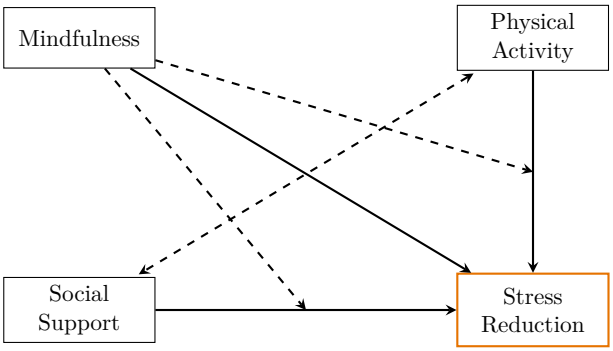

Figure 5.1: The direct and indirect effects of intervention components on the outcome, along with their interactions

In this phase, we conduct an *optimization trial* to identify the combination of components that best meets the targeted constraints and outcomes.

A. Experimental Design A *2×2×2 factorial design* efficiently evaluates multiple components and their potential interactions. Each of the three components (Mindfulness, Physical Activity, Social Support) can be either *included* (Yes) or *excluded* (No), resulting in eight experimental conditions:

Group	Mindfulness	Physical Activity	Social Support	Total Cost
1	Yes	Yes	Yes	\$70 (Over budget)
2	Yes	Yes	No	\$45
3	Yes	No	Yes	\$40
4	Yes	No	No	\$15
5	No	Yes	Yes	\$55 (Over budget)
6	No	Yes	No	\$30
7	No	No	Yes	\$25
8	No	No	No	\$0 (Control)

Table 5.1: 2×2×2 Factorial Conditions for the Stress-Reduction App

B. Determine the Requiered Sample Size To determine the required sample size per group, we follow standard power analysis procedures for factorial designs. The key factors that influence the sample size calculation are:

1. **Significance Level (α)**
   The probability of making a Type I error (false positive).
   Standard choice: $\alpha = 0.05$.

2. **Statistical Power ($1 - \beta$)**
   The probability of correctly detecting an effect if it exists.
   Standard choice: 0.80 (80%) or 0.90 (90%) for higher confidence.
3. **Expected Effect Size (f or d)**
   The magnitude of the effect you expect to observe.
   - Small: $f = 0.10$
   - Medium: $f = 0.25$
   - Large: $f = 0.40$ (Cohen, 1988)

   Since this is a behavioral intervention, a small to medium effect size ($f = 0.10$ to 0.25) is common.
4. **Number of Factors and Levels**
   The design consists of 3 factors, each with 2 levels (Yes/No):

$$2^3 = 8 \text{ groups}$$

The total sample size (N) for a factorial ANOVA is calculated using:

$$N = \frac{\lambda}{f^2} \tag{5.1}$$

where:

- λ is the noncentrality parameter (NCP)
- f is the effect size (Cohen's f)

We aim to calculate the required sample size for a factorial ANOVA with three factors, each having two levels, assuming:

- Significance level: $\alpha = 0.05$
- Power: 0.8
- Effect size (Cohen's f): 0.1
- Number of groups: $k = 2^3 = 8$

Step 1: Compute Degrees of Freedom The degrees of freedom for the between-group effect is:

$$df_1 = k - 1 = 8 - 1 = 7 \tag{5.2}$$

The within-group degrees of freedom (df_2) depends on the total sample size and is initially approximated as a large number (e.g., $10,000$) for estimating the critical F-value.

Step 2: Compute the Critical F-Value The critical F-value is determined from an F-distribution table at $\alpha = 0.05$:

$$F_{\text{crit}} = F_{(1-\alpha, df_1, df_2)} \tag{5.3}$$

Using statistical functions, we obtain:

$$F_{\text{crit}} = 2.0105 \tag{5.4}$$

```
data f_critical;
    alpha = 0.05;  /* Significance level */
    df1 = 7;       /* Degrees of freedom for the numerator */
    df2 = 10000;   /* Degrees of freedom for the denominator (
        large sample approximation) */
    f_crit = quantile("F", 1 - alpha, df1, df2);
    put f_crit=;
run;
```

Listing 5.1: SAS Code to Compute Critical F-Value

Step 3: Compute the Noncentrality Parameter (NCP) The noncentrality parameter (λ) is given by:

$$\lambda = F_{\text{crit}} \times df_1 \tag{5.5}$$

Substituting the values:

$$\lambda = 2.0105 \times 7 = 14.0735 \tag{5.6}$$

Step 4: Compute Total Sample Size The total sample size (N) is calculated using:

$$N = \frac{\lambda}{f^2} \tag{5.7}$$

Substituting the values:

$$N = \frac{14.0735}{(0.1)^2} == 1407.35 \tag{5.8}$$

Thus, the total sample size required is approximately:

$$N \approx 1407 \tag{5.9}$$

Step 5: Compute Sample Size Per Group The sample size per group is given by:

$$n = \frac{N}{k} = \frac{1407}{8} = 175.9 \approx 176 \tag{5.10}$$

Thus, each group should have approximately *176* participants. To compute the

sample size using G*Power, follow these steps:

1. Open G*Power software.

2. Select "F tests" as the test family.
3. Choose "ANOVA: Fixed effects, special, main effects and interactions" as the statistical test.
4. In the "Type of power analysis" dropdown, select "A priori: Compute required sample size – given α, power, and effect size".
5. Set the effect size (f) to 0.1.
6. Set α (Significance level) to 0.05.
7. Set the desired power level to 0.8.
8. Set the number of groups to 8 (since it is a 2^3 factorial design).
9. Set degrees of freedom to 7.
10. Click "Calculate" to obtain the total sample size.

5.2.3 Data Collection & Analysis

Through random assignment, ensure each participant is placed into one of the eight conditions described in Table 5.1. The *Primary Outcome* is stress reduction (measured by self-report scales, physiological markers, etc.). *Secondary Outcome* includes metrics like engagement (frequency of app use) and adherence (how consistently participants follow the intervention).

This SAS code generates synthetic data for a 2^3 factorial design study of a stress-reduction app, where the three factors (Mindfulness, Physical Activity, Social Support) each have two levels ("Yes"/"No") resulting in 8 unique experimental conditions. For each condition, 176 observations are simulated, yielding a total of 1,408 subjects. The code assigns binary factor levels based on the group index, computes a cell-specific mean response (starting from a baseline of 50 and adjusting for main effects—reductions of 5, 3, and 2 units for Mindfulness, Physical Activity, and Social Support respectively—and an additional interaction effect when both Mindfulness and Social Support are "Yes"), and then generates an individual response by adding normally distributed random error (with a standard deviation of 10). This dataset can be used for subsequent factorial analysis.

```
data synthetic;
   call streaminit(123); /* Set random seed for reproducibility */
   do group = 1 to 8;  /* 8 groups (cells) */
      do i = 1 to 176; /* 176 observations per group */
         /* For group 1 to 8, we determine factor levels by
             treating group-1 as a binary index (0 to 7) */
         b = group - 1;
         /* Convert b into binary digits for 3 factors:
            - The most significant bit for Mindfulness,
            - The next for Physical Activity,
            - The least for Social Support.
         */
```

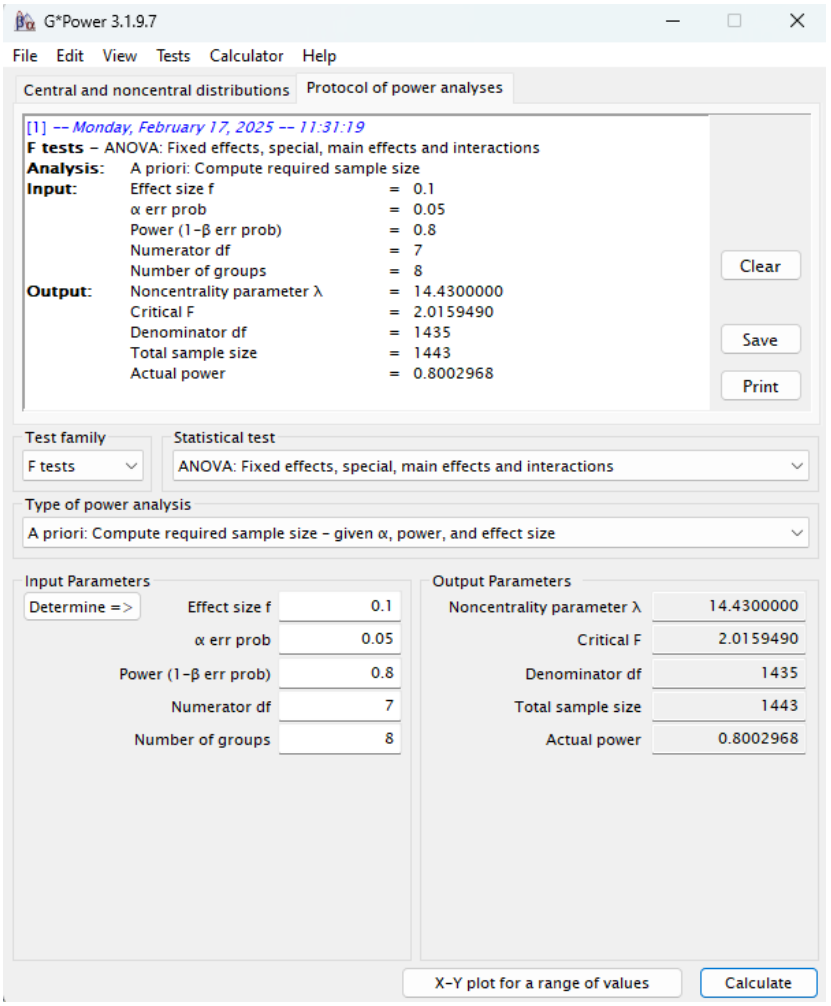

Figure 5.2: G*Power Sample Size Calculation

```
12          Mindfulness = ifc(mod(floor(b/4),2)=1, "Yes", "No");
13          PhysicalActivity = ifc(mod(floor(b/2),2)=1, "Yes", "No")
                ;
14          SocialSupport = ifc(mod(b,2)=1, "Yes", "No");
15
16          /* Compute the cell-specific mean response */
17          meanY = 50;  /* baseline stress level */
18          if Mindfulness = "Yes" then meanY = meanY - 5;
19          if PhysicalActivity = "Yes" then meanY = meanY - 3;
20          if SocialSupport = "Yes" then meanY = meanY - 2;
21          /* Example interaction: if both Mindfulness and
                SocialSupport are "Yes", further reduce by 1 unit */
22          if Mindfulness = "Yes" and SocialSupport = "Yes" then
                meanY = meanY - 1;
23
```

```
24          /* Generate the observed response with error variance =
                10^2 */
25          Response = meanY + rand("normal", 0, 10);
26
27          output;
28       end;
29    end;
30    drop group i b;
31 run;
```

Listing 5.2: SAS Code to Generate Synthetic Data

This SAS code performs a factorial ANOVA using PROC GLM on the synthetic dataset for our stress-reduction app study. The CLASS statement defines the three categorical factors—Mindfulness, Physical Activity, and Social Support—while the MODEL statement specifies the full factorial model including all main effects, two-way interactions, and the three-way interaction. Least-squares means are computed with Tukey's adjustment for multiple comparisons, and contrast statements are used to test the specific hypotheses for each main effect. This analysis is consistent with the Multiphase Optimization Strategy (MOST) approach to evaluating intervention components.

```
1
2
3  proc glm data=synthetic;
4     class Mindfulness (ref="No") PhysicalActivity (ref="No")
          SocialSupport (ref="No") ;
5     model Response = Mindfulness PhysicalActivity SocialSupport
6                      Mindfulness*PhysicalActivity
7                      Mindfulness*SocialSupport
8                      PhysicalActivity*SocialSupport
9                      Mindfulness*PhysicalActivity*SocialSupport /
                           solution;
10    /* Get overall tests of significance for the model */
11    lsmeans Mindfulness PhysicalActivity SocialSupport / adjust=
          tukey;
12    /* Optionally, request tests for interactions */
13    contrast 'Mindfulness main effect' Mindfulness 1 -1;
14    contrast 'Physical Activity main effect' PhysicalActivity 1
          -1;
15    contrast 'Social Support main effect' SocialSupport 1 -1;
16    run;
17 quit;
```

Listing 5.3: SAS Code to Perform Factorial Analyis

This code generates the following output.

Based on the GLM output, the overall model is highly significant ($F(7, 1400) = 27.12$, $p < 0.0001$), indicating that the combination of intervention components

Table 5.2: Overall ANOVA for Response

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	19631.55	2804.51	27.12	<0.0001
Error	1400	144758.15	103.40		
Corrected Total	1407	164389.70			

Table 5.3: Type I ANOVA for Main Effects and Interactions

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Mindfulness	1	13777.42	13777.42	133.25	<0.0001
PhysicalActivity	1	2127.80	2127.80	20.58	<0.0001
SocialSupport	1	3341.93	3341.93	32.32	<0.0001
Mindfulne*PhysicalAc	1	12.23	12.23	0.12	0.7309
Mindfulne*SocialSupp	1	291.93	291.93	2.82	0.0931
PhysicalA*SocialSupp	1	24.93	24.93	0.24	0.6235
Mindfu*Physic*Social	1	55.29	55.29	0.53	0.4647

explains a significant portion of the variance in stress reduction ($R^2 \approx 0.12$). Each main effect is robustly significant: Mindfulness ($F = 133.25$, $p < 0.0001$), Physical Activity ($F = 20.58$, $p < 0.0001$), and Social Support ($F = 32.32$, $p < 0.0001$) all contribute independently to reducing stress. Notably, the interaction between Mindfulness and Social Support shows partial significance ($F = 2.82$, $p = 0.0931$) at our $p < 0.1$ threshold, suggesting that when these components are combined, they may have an additional modest effect on stress reduction beyond their individual contributions. In contrast, the interactions between Mindfulness and

Table 5.4: Parameter Estimates from the Model

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	50.93	0.77	66.45	< .0001
Mindfulness	-5.93	1.08	-5.47	< .0001
PhysicalActivity	-3.31	1.08	-3.05	0.0023
SocialSupport	-2.83	1.08	-2.61	0.0091
Mindfulne*PhysicalAc	1.17	1.53	0.76	0.4472
Mindfulne*SocialSupp	-1.03	1.53	-0.67	0.5023
PhysicalA*SocialSupp	1.32	1.53	0.86	0.3876
Mindfu*Physic*Social	-1.59	2.17	-0.73	0.4647

Physical Activity ($F = 0.12$, $p = 0.7309$), and between Physical Activity and Social Support ($F = 0.24$, $p = 0.6235$), as well as the three-way interaction ($F = 0.53$, $p = 0.4647$), are not significant. Overall, the results imply that while each intervention component significantly reduces stress on its own, there is some evidence that the combined effect of Mindfulness and Social Support provides a small additional benefit.

For the interaction between Mindfulness and SocialSupport, the overall F-test in the ANOVA table may be picking up a slight joint effect (i.e., a small but nonzero collective impact across the levels of this interaction). However, when you look at the individual parameter estimates in the solution table, each coefficient for this interaction might not be statistically different from zero. This discrepancy happens because the overall test is more sensitive to a diffuse effect spread across several parameters, whereas the individual tests are more stringent on each separate parameter.

In short, the ANOVA table suggests that the interaction as a whole might have an effect (even if marginal), while the individual coefficients in the solution table do not reach significance on their own.

5.2.4 Evaluation Phase

The *Evaluation Phase* confirms the effectiveness of the optimized combination in real-world or routine conditions.

1. **Selecting the Final Combination**: The analysis shows *Mindfulness + Physical Activity* ($40) is the best option *that fits within our budget.*
2. **Confirmatory RCT**: Compare the optimized package against a control group or a standard intervention to validate real-world effectiveness.
3. **Assessing Scalability & Feasibility**: If *Mindfulness + Physical Activity* scales effectively, it can be introduced broadly. Otherwise, refine further as needed.

Key Takeaways

- **Systematic Component Testing**: MOST avoids an "all-or-nothing" approach, instead empirically determining which components (or combinations) deliver the best outcomes.
- **Cost-Benefit Balance**: Factoring cost constraints into the design ensures interventions are both *effective and affordable.*
- **Evidence-Based Optimization**: Factorial experiments provide detailed insight into main and interaction effects, guiding data-driven decisions about final component selection.
- **Scalability and Real-World Validation**: A confirmatory trial (RCT) checks whether the optimized intervention retains its effectiveness and feasi-

bility outside of controlled settings.

In sum, MOST offers a *systematic, data-driven* pathway to build and refine interventions that are not only **efficacious** but also **efficient, scalable, and affordable**—key considerations in modern experimental design and applied research.

5.3 Fractional Factorial Designs

Fractional factorial designs are experimental designs that use only a subset of the full factorial combinations. They are particularly useful when the number of factors is high, and testing all possible combinations would be prohibitively expensive or time-consuming. By carefully selecting a fraction of the full factorial set, researchers can still obtain valuable information about main effects and lower-order interactions, assuming that higher-order interactions are negligible.

5.3.1 Balanced vs. Unbalanced Fractional Factorial Designs

- **Balanced Designs:** In balanced fractional factorial designs, each treatment combination (or cell) is represented equally or nearly equally. This balance minimizes confounding and ensures that the estimates of main effects and interactions are unbiased. Balanced designs are generally preferred when the cost per experimental unit is uniform and when it is feasible to maintain equal group sizes.
- **Unbalanced Designs:** Unbalanced designs occur when the number of subjects or experimental units across cells is not equal. Although unbalanced designs may introduce some bias or loss of efficiency in the estimates, they are sometimes unavoidable due to practical constraints (e.g., recruitment difficulties, missing data, or varying resource availability). Special statistical techniques are often required to analyze unbalanced data properly.

5.3.2 Scenarios and Advantages of Fractional Factorial Designs

- **Resource Constraints:** When it is impractical to test all combinations of factors due to limited resources (e.g., time, budget, or available participants), fractional factorial designs provide a cost-effective alternative.
- **Screening Studies:** These designs are especially advantageous during early phases of research or intervention development, where the primary objective is to identify the most influential factors from a large set of potential components.
- **Efficiency in Intervention Optimization:** By reducing the number of experimental conditions, fractional factorial designs allow researchers to rapidly and efficiently optimize interventions. This approach aligns well

with the Multiphase Optimization Strategy (MOST), where the goal is to fine-tune intervention components while balancing effectiveness, affordability, scalability, and efficiency.

In summary, while balanced fractional factorial designs offer the advantage of equal representation across treatment combinations and generally lead to more straightforward analysis, unbalanced designs may be necessary in real-world scenarios. Both approaches, however, are valuable tools for optimizing interventions efficiently under resource constraints and are integral to modern experimental design strategies such as MOST.

Required Readings

- Chapter 5 of Linda M. Collins. *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*. Statistics for Social and Behavioral Sciences. Springer, Cham, Switzerland, 2018. ISBN 978-3-319-72205-4. doi: 10.1007/978-3-319-72206-1.
- David L. Wyrick, Amanda E. Tanner, Jeffrey J. Milroy, Kate Guastaferro, Sandesh Bhandari, Kari C. Kugler, Shemeka Thorpe, Samuella Ware, Alicia M. Miller, and Linda M. Collins. itMatters: Optimization of an online intervention to prevent sexually transmitted infections in college students. *Journal of American College Health*, 70(4):1212–1222, 2022. doi: 10.1080/07448481.2020.1790571.
- Siobhan M. Phillips, Linda M. Collins, Frank J. Penedo, Kerry S. Courneya, Whitney Welch, Alison Cottrell, Gillian R. Lloyd, Kara Gavin, David Cella, Ronald T. Ackermann, Juned Siddique, and Bonnie Spring. Optimization of a technology-supported physical activity intervention for breast cancer survivors: Fit2Thrive study protocol. *Contemporary Clinical Trials*, 66:9–19, 2018. doi: 10.1016/j.cct.2018.01.001.
- Amanda L. Graham, George D. Papandonatos, Sarah Cha, Michael S. Amato, Megan A. Jacobs, Amy M. Cohn, Lorien C. Abroms, and Robyn Whittaker. Effectiveness of an optimized text message and Internet intervention for smoking cessation: A randomized controlled trial. *Addiction*, 117(4):1035–1046, 2022. doi: 10.1111/add.15677.

Chapter 6

Sequential Multiple Assignment Randomized Trials (SMART)

6.1 Summary of the Previous Session

In the previous session, we explored the Multiphase Optimization Strategy (MOST). The focus was on designing, optimizing, and evaluating an experiment before testing it through a randomized controlled trial (RCT). MOST is an excellent approach for developing experiments that are efficient, affordable, scalable, and effective. However, its primary limitation is that once the experiment is designed and optimized, it remains fixed—it is essentially a "one-shot" experiment that cannot be adjusted at least in the *short term*. While MOST provides insights on the *optimal combination* of components, it does not allow us to examine the *optimal sequence* of treatment components.

This rigidity can be a drawback in situations where interventions might benefit from adaptation over time. Since participants may respond differently to an intervention, it can be advantageous to modify the treatment based on early feedback. For instance, if a participant responds well, the same treatment can be continued; if not, switching to an alternative might yield better outcomes.

To address this need for flexibility, we will be covering the SMART (**Sequential Multiple Assignment Randomized Trial**) methodology in this session. SMART offers a structured way to adapt interventions over time, tailoring treatments to individual responses and potentially enhancing overall effectiveness.

6.2 What is SMART?

SMART is an experimental design used to develop and evaluate **adaptive** interventions—strategies that change based on how people respond over time. Unlike a traditional RCT, which tests one fixed approach against another in a single go, SMART involves *multiple stages of randomization*. At each stage, you assess how

participants are doing (via *intermediate outcomes*) and use pre-set decision rules to adapt the intervention for the next step. It's perfect for situations where one-size-fits-all doesn't work—like in business, where customer or employee responses can vary widely.

6.3 SMART Implementation of the Stress Reduction App at University of Miami

SMART builds on optimization frameworks like MOST. While MOST enables us to determine which components—such as *physical activity, social support*, and *mindfulness*—are most effective, SMART takes the next step by testing **how to sequence and adapt these components** based on individual responses during the intervention. Following our example from the previous two sessions, in our stress reduction app, this means we can identify the optimal adaptive path to reduce student stress, ensuring that each participant receives the most effective intervention according to their unique needs. Let's take the stress reduction app as an example to demonstrate how we can implement SMART methodology.

Step 1: Define the Problem and Goal

Problem: University of Miami students experience varying levels of stress, and a single fixed intervention (e.g., just mindfulness) may not work for all—some need different support as they progress.

Goal: Develop an adaptive stress reduction app that adjusts its features over 8 weeks to minimize self-reported stress scores (e.g., on a 0–10 scale).

Components (from MOST): We'll use the three optimized components. Here we can assume that the following three components were determined to be the most effective ones based on the prior factorial analysis in MOST.

- Physical Activity: Daily step goals or exercise prompts.
- Social Support: Group chats or buddy check-ins.
- Mindfulness Exercises: Guided meditation sessions.

Step 2: Set Up the First Stage (Initial Randomization)

Intervention Options: Start with two initial app versions. The selection of these two components is based on the assumption that our MOST optimization results indicate that these are top two most effectuve components.

- Option A: Physical Activity (e.g., "Walk 5,000 steps daily").
- Option B: Mindfulness Exercises (e.g., "10-minute meditation daily").

Randomization: Randomly assign 200 students between the two treatment options A and B. Run this stage for 4 weeks. Here, we're testing two promising

components from MOST, but we expect some students won't respond well initially and need adaptation. (Social support will come into play later.)

We have a synthetic dataset that represents the data collected through a SMART design. The following SAS code shows the number of observations in each option along with their average level of baseline stress. We also run a simple t-test to examine the equality of means between the two groups.

```
/* Tabulate stress_base and stress_week4 by stage1 using PROC
    TABULATE */
PROC TABULATE DATA=stress_app_synthetic;
    CLASS stage1;
    VAR stress_base stress_week4;
    TABLE stage1 ALL,
          N='Sample Size'
          stress_base * (MEAN='Mean' STD='Std Dev')
          stress_week4 * (MEAN='Mean' STD='Std Dev') / RTS=20;
    FORMAT stress_base stress_week4 6.2;

RUN;

/* Independent t-test to compare baseline stress between stage1
    groups */
PROC TTEST DATA=stress_app_synthetic;
    CLASS stage1;
    VAR stress_base;
    TITLE "T-Test: Baseline Stress by Stage 1 Treatment";
RUN;
```

Listing 6.1: SAS Code to Tabulate the Data and Test Randomization

Table 6.1: Baseline Stress by Stage 1 Treatment

Stage 1 Treatment	Sample Size	Mean	Std Dev
Physical Activity (A)	88	6.86	1.61
Mindfulness Exercises (B)	112	6.83	1.43

In the synthetic data, we can observe that 88 people were assigned to option A, and the remaining 112 people were assigned to option B. The randomization is implemented correctly, because we can see that there is no statistically significant difference between the groups when it comes to baseline stress.

Step 3: Measure an Intermediate Outcome

What to Measure: Stress reduction after 4 weeks, via self-reported stress scores (0–10). We are going to define a threshold based on the level of improvement in the stress. If an individual has shown significant improvement, we call them

Table 6.2: T-Test Results for Baseline Stress by Stage 1 Treatment

Method	Variances	DF	t Value	$Pr > \|t\|$
Pooled	Equal	198	0.11	0.9087
Satterthwaite	Unequal	175.4	0.11	0.91

responder, otherwise, they are non-responder because they have not responded well to our treatment.

- **Responders:** Stress reduction ≥ 2 points (e.g., from 8 to 6—they're improving).
- **Non-Responders:** Stress reduction < 2 points (e.g., from 8 to 7 or no change—they need a tweak).

In the synthetic data, we have the following

- Option A (Physical Activity): 65 responders (stress dropped 2+), 23 non-responders.
- Option B (Mindfulness): 70 responders, 42 non-responders.

```
/* Calculate stress reduction and summarize by path */
DATA stress_paths;
    SET stress_app_synthetic;
    stress_reduction = stress_base - stress_week4; /* Positive =
       reduction */
RUN;
    PROC TABULATE DATA=stress_paths;
    CLASS path;
    VAR stress_reduction;
    TABLE path ALL,
          N='Sample Size'
          stress_reduction * (MEAN='Mean Reduction' );
    TITLE "SMART Paths: Stress Reduction by Path";
RUN;
```

Listing 6.2: SAS Code to Show the SMART Treatment Allocations

Step 4: Define Decision Rules for Adaptation

Here we should either rely on theory, analytical optimization models, results from prior studies, or pragmatic considerations to define a set of rules that determine what we should do based on the response levels of participants to our initial intervention components. Here is an example:

- **If Responder:** Keep their current intervention (A or B) for weeks 5–8—it's working.

Table 6.3: SMART Paths: Responder Status and Treatment Assignment

Path	Stage 1	Stage 2	Responder	Sample Size	Stress Reduction		
					Mean	Min	Max
A_A	A	None	Yes	**65**	2.97	2.00	3.97
A_C	A	C	No	9	0.67	0.04	1.38
A_D	A	D	No	14	0.83	0.08	1.41
B_B	B	None	Yes	**70**	3.07	2.05	4.00
B_C	B	C	No	27	0.66	0.03	1.41
B_D	B	D	No	15	0.91	0.22	1.47
All				200	2.29	0.03	4.00

- **If Non-Responder:** Re-randomize to a new option incorporating the third component (social support) or a combination:
  - Option C: Social Support (e.g., "Join a stress buddy group").
  - Option D: Combo of their Stage 1 option + Social Support (e.g., Physical + Social or Mindfulness + Social).

Step 5: Second Stage (Adaptive Randomization)

Responders (135 total):

- 65 from Option A stay on Physical Activity.
- 70 from Option B stay on Mindfulness Exercises.

Non-Responders (65 total):

- 23 from Option A:
  - Randomize 9 to Option C (Social Support).
  - Randomize 14 to Option D (Physical + Social).
- 42 from Option B:
  - Randomize 27 to Option C (Social Support).
  - Randomize 15 to Option D (Mindfulness + Social).

Duration: Run this stage for 4 more weeks (weeks 5–8).

Step 6: Measure the Final Outcome

What to Measure: Stress score at week 8 (with the goal of achieving the lowest possible score).

- **A → A (Physical kept):** Average stress = 2.30.

- **B → B (Mindfulness kept):** Average stress = 2.20.
- **A → C (Physical to Social):** Average stress = 4.80.
- **A → D (Physical to Physical + Social):** Average stress = 5.13.
- **B → C (Mindfulness to Social):** Average stress = 5.34.
- **B → D (Mindfulness to Mindfulness + Social):** Average stress = 4.85.

The following SAS code calculates the descriptive statistics per each path.

```
1
2 PROC MEANS DATA=stress_paths MEAN STD MIN MAX N;
3     CLASS path;
4     VAR stress_week8;
5     TITLE "Descriptive Statistics for Stress at Week 8 by Path";
6 RUN;
```

Listing 6.3: SAS Code to Caluclate Stress at Week 8

Table 6.4: Stress at Week 8 by SMART Path

Path	Sample Size	Stress at Week 8			
		Mean	Std Dev	Min	Max
A_A	65	2.30	1.57	0.00	5.68
A_C	9	4.80	1.85	1.46	6.68
A_D	14	5.13	1.62	3.08	7.71
B_B	70	2.20	1.63	0.00	6.66
B_C	27	5.34	1.61	2.73	8.58
B_D	15	4.85	1.33	1.74	6.32

Step 7: Interpret the Adaptive Paths

Key Findings:

- **Physical Responders:** Best to continue with Physical Activity (Avg. stress = 2.30).
- **Mindfulness Responders:** Best to continue with Mindfulness Exercises (Avg. stress = 2.20).
- **Physical Non-Responders:** Benefit most from switching to Social Support (Avg. stress = 4.8).
- **Mindfulness Non-Responders:** Benefit most from switching to a combination of Mindfulness and Social Support (Avg. stress = 4.85).

6.4 Empirical Methods for Analyzing SMART Data

Analyzing data from a SMART design involves methods that account for its multi-stage, adaptive structure:

1. **Descriptive Analysis:** Summarize the outcomes for each adaptive path using measures such as means, standard deviations, and visualizations (e.g., bar charts). This is what we have done so far already.
2. **Regression-Based Methods:** Model the final stress scores as a function of baseline stress and treatment path. For example:

$$\text{stress}_{\text{week8},i} = \beta_0 + \sum_{j=1}^{5} \beta_j \text{Path}_{j,i} + \beta_6 \text{stress}_{\text{base},i} + \epsilon_i \tag{6.1}$$

where $\text{stress}_{\text{week8},i}$ is the outcome for individual i, β_0 is the intercept (path A_A at zero baseline stress), $\text{Path}_{j,i}$ are dummies for paths (A_A as reference), β_j are their coefficients, β_6 is the baseline stress effect, and $\epsilon_i \sim N(0, \sigma^2)$.

```
1
2  proc glm data=stress_paths;
3      class path (ref="A_A");
4      model stress_week8 = path stress_base /  solution;
5
6      /* Contrast A-based vs. B-based paths */
7      contrast "A vs. B Paths" path 1 1 1 -1 -1 -1; /* A_A, A_
           C, A_D vs. B_B, B_C, B_D */
8      /* Contrast Responders vs. Non-Responders */
9      contrast "Responders vs. Non-Responders" path 1 0 0 1 -1
            -1; /* A_A, B_B vs. B_C, B_D */
10
11     lsmeans path / adjust=tukey pdiff;  /* Adjusted means
           and pairwise comparisons */
12     title "GLM Analysis of Stress at Week 8 by Treatment
           Path, Adjusted for Baseline Stress";
13 run;
```

Listing 6.4: SAS Code to run the GLM model

The intercept (-4.213, $p < 0.0001$) in this GLM model represents the predicted stress at week 8 for path A_A when baseline stress is zero. Adjusted for baseline stress (0.949 per unit, $p < 0.0001$), A_A and B_B (-0.049, $p = 0.6755$) are the most effective paths, with significantly lower stress than A_C (3.024, $p < 0.0001$), A_D (2.520, $p < 0.0001$), B_C (3.079, $p < 0.0001$), and B_D (2.443, $p < 0.0001$), highlighting the superiority of keeping responders on their initial treatment.
The GLM contrast analysis reveals no significant difference between A-based (A_A, A_C, A_D) and B-based (B_B, B_C, B_D) paths in terms of stress at week 8 ($F = 0.01$, $p = 0.9435$), suggesting that the initial treatment

Table 6.5: GLM Analysis of Stress at Week 8 by Treatment Path (Reference: A_A)

Parameter	Estimate	Standard Error	t Value	$Pr > \|t\|$
Intercept	−4.213	0.235	−17.89	<0001
path A_C	3.024	0.242	12.49	<0001
path A_D	2.520	0.200	12.59	<0001
path B_B	−0.049	0.117	−0.42	0.676
path B_C	3.079	0.155	19.81	<0001
path B_D	2.443	0.194	12.57	<0001
path A_A	0.000	0	0	0
stress_base	0.949	0.032	29.64	<0001

Table 6.6: Contrast Results for GLM Analysis of Stress at Week 8

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
A vs. B Paths	1	0.00232	0.00232	0.01	0.9435
Responders vs. Non-Responders	1	57.87872	57.87872	125.65	<0.0001

choice (A or B) does not substantially affect outcomes when averaged across all adaptations. However, a highly significant difference exists between responders (A_A, B_B) and non-responders switched from B to C or D (B_C, B_D) ($F = 125.65$, $p < 0.0001$), with responders exhibiting markedly lower stress, underscoring the effectiveness of maintaining responders on their initial treatment. This contrast excludes non-responders from A (A_C, A_D), limiting its scope, but aligns with prior findings that A_A and B_B outperform most non-responder paths.

Understanding LSMeans vs. Model Estimates in PROC GLM

```
/* 3. Summarize LSMEANS for Optimal Strategy */
proc glm data=stress_paths;
    class path (ref="A_A");
    model stress_week8 = path stress_base;
    lsmeans path / adjust=tukey stderr; /* Get LSMEANS with
        standard errors */
    ods output lsmeans=lsmeans_table; /* Save LSMEANS to
        dataset */
    title "LSMEANS for Ranking Treatment Paths";
run;
proc sort data=lsmeans_table;
    by lsmean; /* Sort by adjusted mean stress */
run;
```

```
proc print data=lsmeans_table;
    var path lsmean stderr;
    title "Ranked LSMEANS of Stress at Week 8 by Path";
run;
```

Listing 6.5: SAS Code to summerize the optimal strategy

Table 6.7: LSMeans Estimates for Treatment Paths from PROC GLM

Path	LSMean	Standard Error	Pr $>$ $\|t\|$	LSMean Number
A_C	5.3087	0.2269	< 0.0001	1
A_D	4.8049	0.1817	< 0.0001	2
B_B	2.2358	0.0811	< 0.0001	3
B_C	5.3635	0.1306	< 0.0001	4
B_D	4.7281	0.1753	< 0.0001	5
A_A	2.2849	0.0842	< 0.0001	6

LSMeans (Least Squares Means) are the adjusted average stress levels at week 8 for each path, factoring in baseline stress (`stress_base`). They tell us what stress we'd expect if everyone started with the same baseline, making it a breeze to compare paths directly. For ranking strategies—like finding the best path to reduce stress—LSMeans are our go-to numbers.
Here's what we get from our `proc glm` run:

Table 6.8: LSMeans Estimates for Treatment Paths from PROC GLM

Path	LSMean	Standard Error	Pr $>$ $\|t\|$	LSMean Number
`A_C`	5.3087	0.2269	< 0.0001	1
`A_D`	4.8049	0.1817	< 0.0001	2
`B_B`	2.2358	0.0811	< 0.0001	3
`B_C`	5.3635	0.1306	< 0.0001	4
`B_D`	4.7281	0.1753	< 0.0001	5
`A_A`	2.2849	0.0842	< 0.0001	6

Look at these: `B_B` (2.2358) and `A_A` (2.2849) show the lowest stress—our top picks! Meanwhile, `A_C` (5.3087) and `B_C` (5.3635) are the highest, suggesting they're less effective. The small standard errors (e.g., 0.0811 for `B_B`) mean these numbers are precise, and the p-values (< 0.0001) confirm they're real effects—though we care more about comparing them to each other than to zero.

How They Differ from Model Estimate Coefficients

Model estimate coefficients show how each path's stress shifts compared to our reference, `A_A`. They're like the building blocks of the model—slopes telling us the difference from `A_A`'s baseline stress, adjusted for `stress_base`. For example, if `A_C`'s coefficient is 3.0239, it means `A_C` is 3.0239 units higher than `A_A`'s 2.2849, landing close to 5.3088—matching its LSMean.
So why pick LSMeans over coefficients?

- **Ready-to-Use**: LSMeans are the final averages—like 5.3087 for `A_C`—no extra math needed. Coefficients (e.g., 3.0239 for `A_C`) require you to add them to `A_A`'s value to get the full story.
- **Big Picture**: LSMeans give us each path's stress level outright, perfect for ranking. Coefficients are just the "shift" from `A_A`, less intuitive for quick comparisons.
- **Extra Perks**: LSMeans come with standard errors and can pair with Tukey-adjusted differences (e.g., `B_B` vs. `A_C`), showing which paths win significantly—coefficients don't do that alone.

LSMeans are our shortcut to the answers we want: which path reduces stress best? They skip the behind-the-scenes details of coefficients and deliver clear, adjusted averages.

3. **Marginal Structural Models (MSM) to Address the Confounding Issue:**
In a SMART design, Stage 1 randomly assigns `A` or `B`, but Stage 2 depends on week 4 response: responders stay on `None` (e.g., `A_A`, `B_B`), while non-responders get `C` or `D` (e.g., `A_C`, `A_D`). This adaptive process causes confounding because intermediate (`stress_week4`) and baseline (`stress_base`) factors affect both the path and outcome (`stress_week8`), biasing the GLM estimates. For example, `A_D` might seem worse due to higher stress in non-responders, not `D` itself.
Picture this: in our SMART experiment, some people end up on *A_D* because they're struggling at week 4—high stress, non-responders. Meanwhile, *A_A* gets the "easier" crowd—low stress, responders. When we compare their week 8 stress, *A_D* looks worse, but is it really *D*'s fault? Or is it just that *A_D* got stuck with tougher cases? That's the confounding mess we need to fix.
Inverse Probability Weighting (IPW) is like a magic scale adjuster. It looks at each person and asks, "How likely were you to land on your path, given your starting stress and week 4 stress?"

- If someone had high week 4 stress and was almost guaranteed to get *A_D*, IPW says, "You're too common in this group—let's shrink your influence a bit" and gives them a small weight (like 0.5).
- If someone had low stress and still ended up on *A_A* when they could've

switched, IPW says, "You're rare—let's boost your voice" and gives them a big weight (like 2).

By tweaking everyone's "volume" this way, IPW creates a pretend world where paths aren't stacked with certain types of people. It's like giving everyone an equal chance to be on any path, no matter their stress levels. Now, when we compare A_A and A_D, we're seeing what each path actually does to stress—not just who they got assigned to.

Step 1: Figure Out the Odds (Assigning Probabilities)

We start by figuring out how likely each person is to end up on their actual path—like `A_D` or `B_B`—based on their situation. For Stage 1, everyone gets `A` or `B`, and since it's random, we assume a 50% chance for each. In Stage 2, it depends: responders stick with `None` (e.g., `A_A`), so their probability is 100%. Non-responders switch to `C` or `D` (e.g., `A_C`, `A_D`), and that chance hinges on their starting stress (`stress_base`), week 4 stress (`stress_week4`), and initial treatment (`stage1`).
It's like guessing someone's next move in a game—knowing their odds helps us spot who's typical or surprising on their path.

```
/* Stage 1: Probability of A vs. B (assuming equal
    randomization) */
data stress_paths;
    set stress_paths;
    if stage1 = "A" then prob_stage1 = 0.5;
    else if stage1 = "B" then prob_stage1 = 0.5;
run;
```

Listing 6.6: SAS Code to calculate and assign probabilities to each observationl

Step 2: Let Logistic Regression Do the Guesswork

For non-responders, we use `proc logistic` to predict their chance of getting `D` over `C`. We feed it `stress_base`, `stress_week4`, and `stage1`, asking: "Given this person's vibe, how likely are they to land on `D`?"

```
/* Stage 2: Probability of treatment conditional on
    responder status */
proc logistic data=stress_paths;
    where responder = 0; /* Non-responders only */
    class stage2 (ref="C") stage1 (ref="A");
    model stage2 (event="D") = stress_base stress_week4
        stage1;
    output out=stage2_probs pred=prob_stage2;
run;

```

```
proc sort data=stress_paths;
    by id;
run;

proc sort data=stage2_probs;
    by id;
run;

/* Now merge the sorted datasets */
data stress_paths_weighted;
    merge stress_paths stage2_probs (keep=id prob_stage2);
    by id;
    if responder = 1 then prob_stage2 = 1; /* Assign 1 for
        responders */
    if prob_stage2 = . then prob_stage2 = 1; /* Fill missing
         values */
run;
```

Listing 6.7: SAS Code to use logistic regression to calculate probabilities

Step 3: Flip the Odds (Inverse Probabilities)

Next, we flip those probabilities upside down to make weights. If someone's got a 70% chance (0.7) of `D`, their weight is $1/0.7 \approx 1.43$. For their full path (Stage 1 $\times$ Stage 2), say $0.5 \times 0.7 = 0.35$, the weight becomes $1/0.35 \approx 2.86$. Responders? Their Stage 2 probability is 1, so their weight is just $1/0.5 = 2$. This is the fairness trick—big weights boost rare cases, small weights quiet down the usual suspects, leveling the field. In the code below, stabilized IPW is used to ensure that the weights do not get too large.

```
data stress_paths_weighted;
    set stress_paths_weighted;
    ipw = 1 / (prob_stage1 * prob_stage2);
    if stage1 = "A" then stab_ipw = 0.5 * ipw;
    else if stage1 = "B" then stab_ipw = 0.5 * ipw;
run;
```

Listing 6.8: SAS Code to calculate stabilized IPWs

Step 4: Use the Weights to Rebalance Everything

We plug these weights into `proc genmod` with `weight stab_ipw`, tweaking how much each person's stress score counts. A rare `A_A` person (weight = 3) counts triple, while a common `A_D` person (weight = 0.5) counts half.
Why it matters: It's like remixing a playlist—weights turn down overplayed cases and crank up the rare ones, so every path gets a fair shot in our analysis.

```
1
2  proc genmod data=stress_paths_weighted;
3      class path (ref="A_A");
4      model stress_week8 = path stress_base / dist=normal link
           =identity;
5      weight stab_ipw;
6      lsmeans path / diff adjust=tukey;
7      ods output lsmeans=msm_lsmeans;
8      title "MSM Estimates of Path Effects with IPW";
9  run;
```

Listing 6.9: SAS code to estimate the smart coefficients using stabilized weights

Table 6.9: Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.2918	0.2378	[-4.7580, -3.8257]	325.62	< 0.0001
path A_C	1	3.0851	0.2190	[2.6558, 3.5144]	198.36	< 0.0001
path A_D	1	2.5802	0.1968	[2.1945, 2.9659]	171.90	< 0.0001
path B_B	1	-0.0484	0.1391	[-0.3211, 0.2243]	0.12	0.7280
path B_C	1	3.2477	0.1304	[2.9922, 3.5033]	620.40	< 0.0001
path B_D	1	2.5763	0.1604	[2.2619, 2.8907]	257.98	< 0.0001
path A_A	0	0.0000	0.0000	[0.0000, 0.0000]	–	–
stress_base	1	0.9604	0.0314	[0.8989, 1.0220]	934.94	< 0.0001
Scale	1	0.8077	0.0404	[0.7323, 0.8908]	–	–

Table 6.10: Path Least Squares Means

Path	Estimate	Standard Error	z Value	Pr > |z|
A_C	5.3697	0.1947	27.58	< 0.0001
A_D	4.8648	0.1694	28.72	< 0.0001
B_B	2.2362	0.0965	23.16	< 0.0001
B_C	5.5323	0.0834	66.36	< 0.0001
B_D	4.8609	0.1253	38.80	< 0.0001
A_A	2.2846	0.1002	22.80	< 0.0001

The `proc genmod` LSMeans, adjusted with IPW (e.g., `B_B` = 2.2362, `B_C` = 5.5323), show each path's true `stress_week8`, fixing confounding that `proc glm` (e.g., `B_B` = 2.2358, `B_C` = 5.3635) didn't catch. Both rank `B_B` and `A_A` best (around 2.23–2.28) and `A_C` and `B_C` worst (5.3–5.5), but IPW nudges `C` and `D` paths up (e.g., `B_C` +0.1688), showing they're less effective than `proc glm` hinted, due to biased assignments. Parameter estimates (e.g., `A_C` = 3.0851) shift from `A_A` like in `proc glm`, but IPW refines them.

4. **Power Analysis:**
   Simulating power for an MSM in a SMART.
   For MSM with IPW in our SMART—comparing paths like `A_A` vs. `B_B`—simulation beats formulas due to adaptive splits and weights. We assume `A_A` stress = 2.5, `B_B` = 2.0 ($\Delta = 0.5$), $\sigma = 2$, and $N = 200$ (100 `A`, 100 `B`; 70% responders, 30% split 1:2 `C`:`D`). Steps: (1) Generate `stress_base` ($N(6.8, 2)$), assign `stage1` (50/50), model `stress_week4` ($0.8 \times$ stress_base $+ \epsilon$), set `responder` (70% via logit), assign `stage2`, and simulate `stress_week8`. (2) Apply IPW (e.g., $1/(0.5 \times 0.7)$ for `A_A`), fit `proc genmod`, test `A_A` vs. `B_B`. (3) Repeat 1000 times—power is the % of $p < 0.05$. Our 200 might give 60% power ($p = 0.1193$ aligns), needing 400 for 80%. Simulation captures our design's quirks—no simple n formula fits!
5. **Q-Learning:** It is a reinforcement learning trick that's like teaching a coach to pick the best plays in a game. For our SMART experiment, it figures out the optimal treatment path—starting with `A` or `B` at Stage 1, then choosing `C`, `D`, or `None` at Stage 2—to minimize stress at week 8 (`stress_week8`). It works backward: first, it predicts Stage 2 outcomes based on week 4 stress (`stress_week4`) and responder status, then uses those to pick Stage 1 based on baseline stress (`stress_base`). The "Q" stands for quality—how good each choice is, tailored to where a patient's at. Unlike fixed rankings, it builds a flexible rule, adapting to each individual person's journey.
   The following SAS code implements Q-learning to optimize treatment in our SMART data, minimizing `stress_week8` across two stages.

   - **Stage 2 Q-Function**: `proc glm` fits `stress_week8` for non-responders (`responder = 0`) using `stage2` (`C`, `D`), `stress_week4`, `stress_base`, and `stage1` as predictors. The `/ solution` option details coefficients, and `output out=stage2_preds predicted=q2_pred` saves predicted stress as `q2_pred`.
   - **Assign Q2**: The `data` step merges predictions with the full dataset, setting `q2` to observed `stress_week8` for responders (`responder = 1`), or `q2_pred` for non-responders, with a fallback to `stress_week8` if predictions are missing.
   - **Stage 2 LSMeans**: `proc glm` re-models `stress_week8` for non-responders, computing LSMeans for `stage2` at mean `stress_week4` and `stress_base`. `adjust=tukey` adjusts p-values, and `ods output` saves results to `stage2_lsmeans`.
   - **Stage 1 Q-Function**: `proc glm` models `q2` (optimal Stage 2 stress) with `stage1` (`A`, `B`) and `stress_base`, outputting predictions as `q1` to `stage1_preds`.
   - **Stage 1 LSMeans**: Another `proc glm` calculates LSMeans for `stage1` at mean `stress_base`, using `adjust=tukey` for significance, saving to `stage1_lsmeans`.

- **Output**: `proc print` displays `stage2_lsmeans` and `stage1_lsmeans`, showing predicted stress for `C` vs. `D` and `A` vs. `B`, guiding our policy.

This backward process—Stage 2 to Stage 1—builds a dynamic treatment rule based on Q-values, leveraging regression to predict and compare outcomes at each decision point.

```
1
2  /* Stage 2 Q-Function for Non-Responders */
3  proc glm data=stress_paths;
4      where responder = 0;
5      class stage2 stage1;
6      model stress_week8 = stage2 stress_week4 stress_base
           stage1 / solution;
7      output out=stage2_preds predicted=q2_pred;
8  run;
9
10 /* Assign Q2 values */
11 data stress_paths_q2;
12     merge stress_paths stage2_preds (keep=id q2_pred);
13     by id;
14     if responder = 1 then q2 = stress_week8; /* Responders:
           Q2 is observed */
15     else if q2_pred = . then q2 = stress_week8; /* Fallback
           for missing preds */
16     else q2 = q2_pred; /* Non-responders: Q2 is predicted */
17 run;
18
19 /* Stage 2 LSMeans for Non-Responders  */
20 proc glm data=stress_paths_q2;
21     where responder = 0;
22     class stage2 stage1;
23     model stress_week8 = stage2 stress_week4 stress_base
           stage1 / solution;
24     lsmeans stage2 / adjust=tukey; /* No 'at', uses mean
           values */
25     ods output lsmeans=stage2_lsmeans;
26 run;
27
28 /* Stage 1 Q-Function */
29 proc glm data=stress_paths_q2;
30     class stage1;
31     model q2 = stage1 stress_base / solution;
32     output out=stage1_preds predicted=q1;
33 run;
34
35 /* Stage 1 LSMeans  */
36 proc glm data=stress_paths_q2;
37     class stage1;
38     model q2 = stage1 stress_base / solution;
39     lsmeans stage1 / adjust=tukey; /* No 'at', uses mean
           stress_base */
```

```
    ods output lsmeans=stage1_lsmeans;
run;

/* Output Results for Policy */
proc print data=stage2_lsmeans;
    title "Stage 2 LSMeans for Non-Responders (C vs D)";
run;

proc print data=stage1_lsmeans;
    title "Stage 1 LSMeans (A vs B)";
run;
```

Listing 6.10: SAS code to implement Q-Learning

Here's the output:

Table 6.11: Stage 2 LSMeans for Non-Responders (C vs. D)

Stage 2	LSMean	Std Err	Pr > |t|
C	5.2660	0.15	0.0033
D	4.8963	0.14	–

Table 6.12: Stage 1 LSMeans (A vs. B)

Stage 1	LSMean	Std Err	Pr > |t|
A	2.9950	0.10	0.1193
B	3.3236	0.09	–

At Stage 2, for non-responders, switching to D predicts a lower stress (4.8963) than C (5.2660), with a significant difference (p = 0.0033)—so D wins. At Stage 1, starting with A gives a lower expected stress (2.9950) than B (3.3236), though it's not a slam dunk (p = 0.1193). Our policy: "Start with A; if a responder at week 4, stay (A_A); if a non-responder, switch to D (A_D)." This averages over our data's mean `stress_base` (around 6.8), so it's a broad rule.

Our IPW-adjusted `proc genmod` gave fixed LSMeans: A_A = 2.2846, B_B = 2.2362, A_D = 4.8648, B_D = 4.8609, A_C = 5.3697, B_C = 5.5323. It ranked B_B and A_A as best, with C paths lagging. Q-learning aligns somewhat: A_A fits its Stage 1 pick of A (2.9950) for responders, and D over C matches A_D (4.8648) beating A_C (5.3697). But Q-learning flips Stage 1, favoring A over B, unlike IPW's slight edge to B_B. Why? IPW averages whole paths, while Q-learning optimizes step-by-step, suggesting A sets up better flexibility. The real difference: IPW ranks static paths; Q-learning crafts a dynamic "if-then" plan—perfect for personalizing treatment!

Class Activity Take a few minutes to review *"An adaptive behavioral intervention for weight loss management: a randomized clinical trial." JAMA, 332(1):21–30, 2024.* Given what you have learned in class, present the study and draw parallels between the concepts discussed in class and what has been implemented in the paper.

Required Readings

- Theophile Bigirumurame, Germaine Uwimpuhwe, and James Wason. Sequential multiple assignment randomized trial studies should report all key components: A systematic review. *Journal of Clinical Epidemiology*, 142:152–160, 2022. doi: 10.1016/j.jclinepi.2021.11.007.
- Kelley M. Kidwell and Daniel Almirall. Sequential, multiple assignment, randomized trial designs. *JAMA*, 329(4):336–337, 2023. doi: 10.1001/jama.2022.24324.
- Amy S. Ruppert, Jun Yin, Marie Davidian, Anastasios A. Tsiatis, John C. Byrd, Jennifer A. Woyach, and Sumithra J. Mandrekar. Application of a sequential multiple assignment randomized trial (SMART) design in older patients with chronic lymphocytic leukemia. *Annals of Oncology*, 30(4):542–550, 2019. doi: 10.1093/annonc/mdz053.
- Bonnie Spring, Angela F. Pfammatter, Laura Scanlan, Elyse Daly, Jean Reading, Sam Battalio, H. Gene McFadden, Don Hedeker, Juned Siddique, and Inbal Nahum-Shani. An adaptive behavioral intervention for weight loss management: A randomized clinical trial. *JAMA*, 332(1):21–30, 2024. doi: 10.1001/jama.2024.0821.

Part III

Adaptive and Real-Time Designs

Chapter 7

Just-in-Time Adaptive Interventions (JITAIs)

7.1 Summary of the Previous Session

In our previous sessions, we explored the Multiphase Optimization Strategy (MOST) and Sequential Multiple Assignment Randomized Trials (SMART) in great detail. MOST is a framework that helps us build and fine-tune interventions by identifying their most effective **components**. Meanwhile, SMART takes a dynamic approach, using multiple stages of randomization to adapt intervention **sequences** based on how individuals respond over time. Together, these tools provide a solid foundation for creating interventions that are both efficient and personalized to meet diverse needs.

Now, we're ready to dive into Just-in-Time Adaptive Interventions (JITAIs), which take adaptability to a whole new level. JITAIs use **real-time data**—like information from wearable devices or smartphone sensors—to deliver **tailored interventions exactly when it's needed**, such as a reminder to take medication or a boost of encouragement during a tough moment. Building on what we've learned about MOST and SMART, this session will explore how these frameworks can help us design and refine JITAIs, ensuring they're not only effective but also seamlessly integrated into everyday life.

7.2 Introduction to JITAIs

They are an innovative class of interventions that deliver **personalized** support to individuals precisely when they need it, **adapting in real-time to their changing circumstances**. Unlike traditional interventions, which are often fixed and uniform, JITAIs leverage technologies such as mobile apps, wearable sensors, or self-reports to collect real-time data and dynamically adjust the type, timing, and content of the support provided.

For example, a JITAI might send a motivational prompt to encourage exercise when a person has been inactive or offer stress-relief strategies when elevated stress is detected. As another example, In an e-commerce context, a JITAI can drive purchases by delivering personalized offers—like a discount or free shipping—exactly when customers show intent, such as lingering on a product page or abandoning a cart, using real-time data and decision rules to convert hesitation into action.

JITAIs represent a paradigm shift in experimental research. Their key advantages include:

- **Optimal Timing:** Interventions are delivered at moments of greatest need or receptivity, enhancing effectiveness.
- **Personalization:** Support is tailored to an individual's current state, context, and preferences.
- **Efficiency:** By targeting interventions strategically, JITAIs minimize unnecessary actions, reducing participant burden.
- **Scalability:** Technology enables JITAIs to reach large populations cost-effectively.

7.3 Components of JITAIs

7.3.1 Decision Points

Definition: Specific moments when a decision is made about whether and how to intervene.

Examples:

- Time-based (e.g., every morning)
- Event-based (e.g., entering a high-risk location)
- State-based (e.g., increased heart rate)

Research Consideration: Identifying relevant decision points requires understanding the behavior or outcome of interest.

7.3.2 Intervention Options

Definition: The range of support strategies available at each decision point.

Examples:

- Sending a text reminder
- Offering educational content
- Prompting a specific action like deep breathing

Research Consideration: Options should be evidence-based and feasible within the delivery platform (e.g., a smartphone).

7.3.3 Tailoring Variables

Definition: Data-driven factors that determine which intervention option is delivered and when.

Examples:

- Demographic traits (e.g., age)
- Behavioral history (e.g., past activity levels)
- Real-time metrics (e.g., location, stress levels)

Research Consideration: Selecting tailoring variables involves balancing granularity with practicality.

7.3.4 Decision Rules

Definition: The logic or algorithms connecting tailoring variables to intervention options.

Examples:

- Simple rules (e.g., "If stress > threshold, send relaxation prompt")
- Complex models (e.g., machine learning predictions)

Research Consideration: Rules must be tested and refined to ensure they effectively target the intended outcomes.

7.4 Extending the Stress-Reduction App Example with JITAIs

7.4.1 Background

In prior sessions, we explored a stress-reduction app for University of Miami students, optimized using the Multiphase Optimization Strategy (MOST) to identify effective components—mindfulness exercises, social support prompts, and physical activity prompts. The intervention was then imporved via a Sequential Multiple Assignment Randomized Trial (SMART) to adapt interventions based on response and figure out the best sequence of components.

Now, we extend this into a JITAI framework, focusing on real-time delivery, implementation, and analysis, using a synthetic dataset to demonstrate key concepts.

7.4.2 Designing the JITAI

The JITAI aims to reduce stress by delivering the right intervention at the right time. Its design builds on three core components:

Decision Points

Moments of elevated stress, detected via:

- **Self-reported stress levels:** Collected via app check-ins, using a 1-10 scale.
- **Physiological data:** Monitored through wearable devices (e.g., heart rate variability from a Fitbit).
- **Contextual triggers:** Detected through app usage patterns (e.g., late-night study hours).

Intervention Options

- **Mindfulness exercise:** A 5-minute guided breathing session.
- **Social support prompt:** A suggestion such as "Text a friend for a quick chat."
- **Physical activity prompt:** A notification such as "Take a 10-minute walk."

Tailoring Variables

The intervention is tailored based on:

- Stress level at the decision point.
- Time of day.
- The student's past response to interventions (e.g., prefers mindfulness over activity).

Decision Rules

A set of predefined logic determines which intervention to deliver:

- **Example Rule:** If stress > 6 and it is after 8 PM, deliver a mindfulness exercise.
- **Example Rule:** If stress > 4 and the student has been inactive for 2 hours, suggest physical activity.

7.4.3 Implementation

The implementation leverages mobile technology and real-time data:

Platform

The app integrates with wearable devices (e.g., Fitbit) to track physiological data and uses push notifications for intervention delivery.

Data Collection

- Continuous monitoring of stress indicators:
  - Self-reports every 4 hours.
  - Wearable sensor data recorded every minute.
- Tracking user responses to interventions (e.g., accepted, declined, completed).

Execution

At each decision point, the app's algorithm:

- Assesses the relevant tailoring variables.
- Selects an appropriate intervention based on decision rules.
- Triggers intervention delivery with minimal latency (e.g., <30 seconds from stress detection to notification).

Variables in the Synthetic Dataset

The synthetic dataset, `jitai_stress_data.csv`, contains 550 observations generated for the JITAI implementation of a stress-reduction app for University of Miami students. The dataset spans March 1–7, 2025, with 550 observations (50 students, ∼11 decision points each).Below is a description of each variable, reflecting its design and role in the real-time intervention framework.

1. **Student_ID**
   - *Description*: Unique identifier for each participating student.
   - *Type*: Character (e.g., "S001", "S002", ..., "S050").
   - *Purpose*: Distinguishes individual students to track stress, interventions, and responses.
   - *Generation*: Assigned sequentially to 50 hypothetical students (S001 to S050).
   - *Role in JITAI*: Enables individual-level analysis, capturing student-specific patterns.
2. **Timestamp**
   - *Description*: Date and time of the decision point when stress is assessed and an intervention is delivered.
   - *Type*: Character in CSV (e.g., "01MAR2025:08:23:45"), convertible to datetime.
   - *Purpose*: Marks the exact moment of intervention, crucial for real-time adaptation.
   - *Generation*: Starts March 1, 2025, with random offsets over 7 days (March 1–7), simulating 2–4 check-ins per day.

- *Role in JITAI*: Links interventions to timing, enabling evaluation of temporal effects (e.g., night vs. morning).

3. **Stress_Level**
   - *Description*: Student's stress level at the decision point, on a 1–10 scale.
   - *Type*: Numeric (integer, 1 to 10).
   - *Purpose*: Primary tailoring variable to determine intervention need and type.
   - *Generation*: Drawn from a skewed distribution (higher probability for 4–7), with a +2 increase during Night hours.
   - *Role in JITAI*: Triggers decision rules (e.g., Stress_Level > 6 prompts Mindfulness at night).
4. **Time_of_Day**
   - *Description*: Categorizes the decision point into Morning (6 AM–12 PM), Afternoon (12 PM–6 PM), Evening (6 PM–12 AM), or Night (12 AM–6 AM).
   - *Type*: Character (e.g., "Morning", "Night").
   - *Purpose*: Tailoring variable to adapt interventions to circadian or contextual factors.
   - *Generation*: Derived from `Timestamp` hour using conditional logic.
   - *Role in JITAI*: Influences decision rules (e.g., Mindfulness prioritized in Evening/Night for high stress).
5. **Hours_Since_Active**
   - *Description*: Hours since the student's last recorded physical activity, inferred from wearable data.
   - *Type*: Numeric (float, 0 to 5, one decimal place).
   - *Purpose*: Tailoring variable to detect inactivity, triggering physical activity prompts.
   - *Generation*: Randomly assigned from a uniform distribution (0 to 5 hours).
   - *Role in JITAI*: Drives rule: if > 2 hours inactive and Stress_Level > 4, deliver Physical intervention.
6. **Past_Preference**
   - *Description*: Student's historically preferred intervention (Mindfulness, Social, Physical, or None).
   - *Type*: Character (e.g., "Mindfulness", "Social").
   - *Purpose*: Tailoring variable reflecting individual tendencies for personalization.
   - *Generation*: Randomly assigned with equal probability across options

(simplified; ideally dynamic).

- *Role in JITAI*: Guides intervention choice for moderate stress or when rules don't apply.

7. **Intervention**
   - *Description*: Intervention delivered: Mindfulness (5-minute breathing), Social (friend prompt), Physical (10-minute walk), or None.
   - *Type*: Character (e.g., "Mindfulness", "None").
   - *Purpose*: Action tested for stress reduction efficacy.
   - *Generation*: Assigned via decision rules: Stress_Level $>$ 6 and Evening/Night $\rightarrow$ Mindfulness; Stress_Level $>$ 4 and Hours_Since_Active $>$ 2 $\rightarrow$ Physical; else Social or None (weighted by Past_Preference).
   - *Role in JITAI*: Core component evaluated for real-time effectiveness.
8. **Response**
   - *Description*: Whether the student completed the intervention (1 = yes, 0 = no/declined).
   - *Type*: Numeric (binary, 0 or 1).
   - *Purpose*: Measures engagement, a proximal indicator of acceptability.
   - *Generation*: 70% completion if matches `Past_Preference`, 50% otherwise (Bernoulli).
   - *Role in JITAI*: Links delivery to behavior, impacting stress reduction likelihood.
9. **Stress_Reduced**
   - *Description*: Outcome indicating if stress decreased post-intervention (1 = yes, 0 = no/increase).
   - *Type*: Numeric (binary, 0 or 1).
   - *Purpose*: Primary outcome assessing JITAI effectiveness (akin to "purchase").
   - *Generation*: 80% success if `Response` = 1, 20% if `Response` = 0 (Bernoulli with noise).
   - *Role in JITAI*: Drives evaluation and optimization of decision rules.

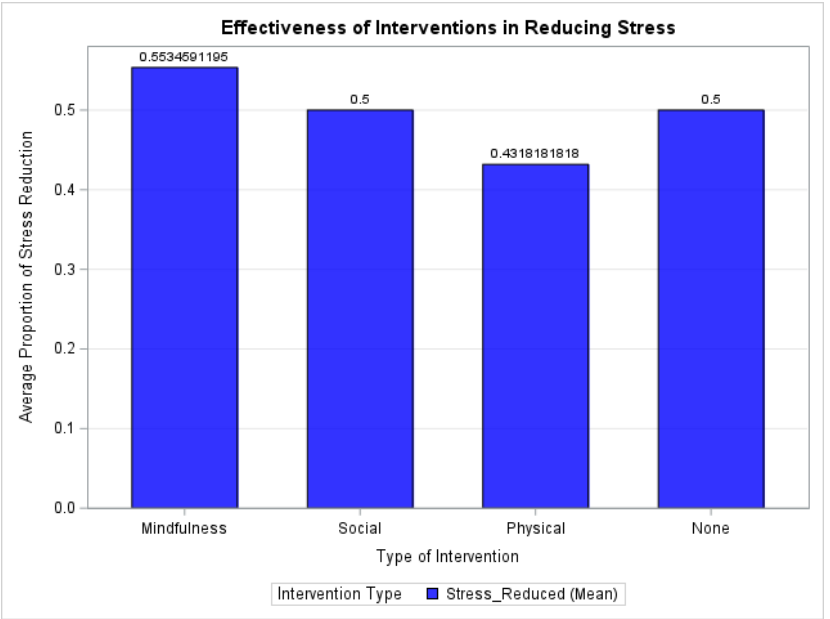

Figure 7.1: Effectivness of Interventions

```
proc logistic data=jitai_data descending;
    class Intervention (ref="None") Time_of_Day (ref="Morning")
        Past_Preference (ref="None") / param=ref;
    model Stress_Reduced = Intervention Stress_Level Time_of_Day
        Hours_Since_Active Past_Preference / link=logit firth;
    where Response = 1;
    oddsratio Intervention;
run;
```

Listing 7.1: SAS code to estimate the effects of interventions for those who responded

When we run the code without the firth adjustment, we get a warning that says "*The maximum likelihood estimate may not exist*". This is due to the nature of the data that we are using here, which leads to uasi-Complete Separation.

This occurs when predictors (e.g., `Intervention`, `Stress_Level`) perfectly or nearly perfectly predict the outcome (`Stress_Reduced`) in a data subset, leading to infinite or unstable maximum likelihood estimates (MLEs). In logistic regression, the probability of the outcome is modeled as:

$$P(\text{Stress_Reduced} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{Intervention} + \beta_2 \cdot \text{Stress_Level} + \dots)}}$$

where β_0 is the intercept, and $\beta_1, \beta_2, \dots$ are coefficients estimated via MLE to maximize the likelihood of observing the data. Normally, a mix of 1s and 0s across predictor values allows the algorithm to find finite coefficients. However, if

a predictor (e.g., `Intervention` = Mindfulness) perfectly splits the outcome (all 1s, while `None` = all 0s), the model seeks an infinite β_1 to achieve $P = 1$ (since $e^{-\text{large number}} \to 0$), but no finite value suffices. This causes convergence failure, as the algorithm cannot settle on a solution.

Visually, imagine a scatterplot with `Intervention` (None = 0, Mindfulness = 1) on the x-axis and `Stress_Reduced` (0 or 1) on the y-axis. If all Mindfulness points are 1 and None points are 0, a vertical line separates them perfectly, requiring an infinite slope—an impossibility in logistic regression. Thus, quasi-complete separation (near-perfect prediction) destabilizes MLEs, necessitating adjustments like Firth's penalized likelihood to stabilize estimates.

The Firth method addresses quasi-complete separation by adding a penalty to logistic regression's likelihood, preventing infinite coefficients. Standard logistic regression predicts the probability of the outcome using maximum likelihood to maximize the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

When separation occurs (e.g., all Mindfulness = 1, None = 0), β_1 grows infinitely, stalling convergence. The Firth method modifies this to a penalized log-likelihood:

$$\ell^*(\beta) = \ell(\beta) + \frac{1}{2} \log |\mathbf{I}(\beta)|$$

Here, $\mathbf{I}(\beta)$ is the Fisher information matrix, reflecting data precision about β. The penalty, $\frac{1}{2} \log |\mathbf{I}(\beta)|$, increases as β becomes extreme, pulling coefficients back from infinity. For the JITAI data, this ensures finite estimates (e.g., Mindfulness odds = 20 vs. None) even with near-perfect separation, allowing the model to converge and produce usable results.

We could also look at the interaction effects to see how the effectiveness of an intervention changes with other variables.

```
proc logistic data=jitai_data descending;
    class Intervention (ref="None") / param=ref;
    model Stress_Reduced = Intervention Hours_Since_Active
        Intervention*Hours_Since_Active / link=logit firth;
run;
```

Listing 7.2: SAS code to estimate the intercation effects

Required Readings

- Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. Just-in-time adaptive interventions

Table 7.1: Penalized Maximum Likelihood Estimates

Parameter	Level	Estimate	Std. Error	Wald Chi-Square	Pr > ChiSq
Intercept		2.2754	1.8082	1.5835	0.2083
Intervention	Mindfulness	-1.1500	1.8046	0.4061	0.5240
Intervention	Physical	-1.6071	1.7450	0.8482	0.3570
Intervention	Social	-1.6230	1.7660	0.8446	0.3581
Stress_Level		-0.00702	0.00740	0.9011	0.3425
Time_of_Day	Afternoon	0.2246	0.3990	0.3168	0.5735
Time_of_Day	Evening	0.4311	0.6540	0.4345	0.5098
Time_of_Day	Night	-0.8690	0.7209	1.4533	0.2280
Hours_Since_Active		0.2005	0.1074	3.4871	0.0619
Past_Preference	Mindfulness	-0.1948	0.4380	0.1978	0.6565
Past_Preference	Physical	0.2138	0.4288	0.2486	0.6180
Past_Preference	Social	0.4167	0.4483	0.8639	0.3526

Table 7.2: Penalized Maximum Likelihood Estimates

Parameter	Level	Estimate	Std. Error	Wald Chi-Square	Pr > ChiSq
Intercept		-4.3106	3.5488	1.4754	0.2245
Intervention	Mindfulness	4.2700	3.5550	1.4427	0.2297
Intervention	Physical	4.1415	3.5670	1.3481	0.2456
Intervention	Social	4.7174	3.5870	1.7296	0.1885
Hours_Since_Active		5.5345	4.3258	1.6369	0.2008
Hours_Since_Active * Intervention	Mindfulness	-5.4257	4.3265	1.5727	0.2098
Hours_Since_Active * Intervention	Physical	-5.5696	4.3273	1.6566	0.1981
Hours_Since_Active * Intervention	Social	-5.9669	4.3502	1.8814	0.1702

(JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018. doi: 10.1007/s12160-016-9830-8.

- Suat Gonul, Tuncay Namli, Sasja Huisman, Gokce Banu Laleci Erturkmen, Ismail Hakki Toroslu, and Ahmet Cosar. An expandable approach for design and personalization of digital, just-in-time adaptive interventions. *Journal of the American Medical Informatics Association*, 26(3):198–210, 2019. doi: 10.1093/jamia/ocy160.

Chapter 8

Micro-Randomized Trials (MRTs)

8.1 Summary of the Previous Session

In the previous session we learned how SMART optimizes the sequence of interventions and how JITAIs take this further by delivering real-time, context-sensitive interventions (e.g., via mobile devices), often relying on SMART-like decision rules to determine 'when' and 'what' to deliver. MRTs, the focus of today's session, zoom in on the proximal effects of these intervention components, randomizing them repeatedly within individuals to assess their immediate impact and refine JITAIs. MRTs provide the experimental backbone to test and perfect JITAI delivery—offering a pipeline from design to real-time adaptation in mobile health interventions.

8.2 Micro Randomized Trial (MRT)

Micro-Randomized Trials (MRTs) are an **experimental design framework** used to assess the **proximal** (short-term) effects of intervention components in real time. In MRTs, *participants are* ***randomized repeatedly****—often multiple times per day* to different intervention options (e.g., receiving a push notification vs. no notification) at specific **decision points** over the course of a study. This approach allows researchers to evaluate how these components impact immediate outcomes, such as behavior or mood, and to understand how effects may vary across time or contexts. MRTs are particularly valuable in developing adaptive interventions, such as those in mobile health (mHealth), by providing data to optimize the timing, dosage, and type of support delivered to individuals.

Purpose: MRTs help us assess how specific intervention components—like mindfulness, physical activity, or peer support in our University of Miami stress-reduction app—impact students' stress levels in real time.

Key Idea: By testing these components moment-to-moment, MRTs reveal what works, when, and for whom, enabling us to build smarter, adaptive interventions.

8.3 How Does MRT Work?

- **Repeated Randomizations:** Participants are randomized multiple times (e.g., daily or hourly) to intervention options or a control (no intervention). Note that in many cases, we should first determine if an individual is eligible for randomization.
- **Decision Points:** These are specific moments when the app decides what to deliver—like 8 AM, 12 PM, or 4 PM when stress might spike for UM students.
- **Proximal Outcomes:** The immediate effects we measure—like a student's self-reported stress level 30 minutes after a physical activity prompt.

For our app, we might randomize a student to a 5-minute breathing exercise at 12 PM, then check their stress level. Tomorrow at 12 PM, the exact same student might get a 'walk break' prompt instead

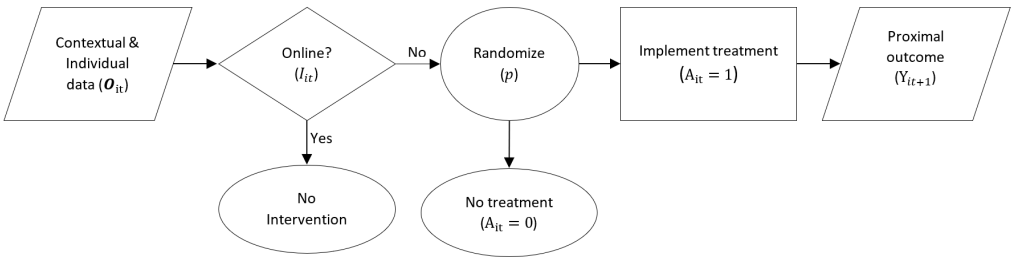

Figure 8.1: MRT Flowchart From Yaraghi et al. 2025

8.4 Modeling Time-Varying Effects and Endogeneity in MRTs

In MRTs, **time-varying effects** refer to how the impact of an intervention component (e.g., a mindfulness prompt) might change over time or across decision points—due to factors like habituation, fatigue, or cumulative effects. **Endogeneity** arises when the decision to deliver an intervention at a given time depends on prior states or interventions (e.g., a student's low stress at 3 PM might result from a 12 PM prompt), making the treatment assignment non-random in a causal sense. This complicates the estimation of true proximal effects because the current intervention's delivery is entangled with past outcomes or actions. MRTs address

this through advanced statistical modeling, ensuring that causal effects are isolated despite these dependencies.

8.4.1 Why It's a Problem in MRTs?

MRTs involve repeated randomizations at multiple decision points (e.g., 8 AM, 12 PM, 3 PM daily). If a student's state at 3 PM (e.g., low stress or already walking) is influenced by a prior intervention (e.g., a walk prompt at 12 PM), then excluding them from randomization at 3 PM (due to unavailability) introduces a feedback loop. This feedback makes the treatment assignment endogenous—it's no longer purely random but conditional on prior effects, violating the assumptions of standard causal inference models (e.g., independence of treatment assignment).

8.4.2 Goal of Modeling

The aim is to estimate the average proximal effect of an intervention (e.g., how much a 3 PM mindfulness prompt reduces stress by 3:30 PM) across all participants, while accounting for:

- Time-varying effects: Does the prompt's effect diminish over days as students habituate?
- Endogeneity: Does the prompt's delivery depend on prior prompts' success?

Without adjustment, estimates could be biased—overstating or understating the true effect.

8.4.3 How to Address the Problem Using GEE and IPW Method

MRTs use statistical techniques like generalized estimating equations (GEE) with inverse probability weighting (IPW) to adjust for these dependencies. These methods:

- Treat prior interventions and states as time-varying covariates.
- Reweight the data to mimic a scenario where randomization is independent of past effects.

Model the proximal effect of a 3 PM intervention (A_t) on stress at 3:30 PM (Y_t):

$$Y_t = \beta_0 + \beta_1 A_t + \beta_2 T + \beta_3 (A_t \times T) + \beta_4 S_t + \beta_5 A_{t-1} + \epsilon_t$$

- Y_t: Stress at 3:30 PM (0-10).
- A_t: 3 PM intervention (e.g., Walk = 1, None = 0).
- T: Day (1 to 14).
- $A_t \times T$: Time-varying effect.
- S_t: Stress at 3 PM.
- A_{t-1}: 12 PM intervention (e.g., Mindfulness = 1).

Fit with:

- *IPW*: Weight $w_t = 1/P(A_t|H_t)$, where $H_t = \{S_t, A_{t-1}, T\}$.
- *GEE*: AR(1) correlation for repeated measures.

Interpretation

- β_1: Average effect of walk at 3 PM (e.g., -0.9 points).
- β_3: Effect change per day (e.g., -0.05 if fading).
- Adjusted for A_{t-1} to correct endogeneity.

This ensures unbiased estimates of intervention effects, even with prior influences or time trends.

8.4.4 How to Address the Problem Using Weighted and Centered Least-Squares (WCLS) Method

The **Weighted and Centered Least-Squares (WCLS)** method estimates causal excursion effects in Micro-Randomized Trials (MRTs), focusing on proximal intervention effects in real-time digital interventions (Boruvka et al., 2018).

How It Works

- *Key Idea*: WCLS measures the difference in `Stress_t30` when an intervention (e.g., Mindfulness) is delivered versus not, among available students. It uses MRT's repeated randomizations.
- *Steps*:
  1. **Weighting**: Uses `Available` (I_t) as a binary weight (1 if available, 0 if not), including only randomized time points.
  2. **Centering**: Centers treatment indicators around the randomization probability ($p = 0.25$). E.g., `Mindfulness_c = Mindfulness - 0.25`.
  3. **Modeling**: Fits $\text{Stress}_t30 = \alpha_1 \text{Prev_Intervention} + \beta_1 \text{Mindfulness_c} + \beta_2 \text{Walk_c} + \beta_3 \text{Peer_c} + \text{time effects}$.
  4. **Estimation**: Solves via GEE with independence correlation matrix.

Example

At 3 PM, a student is available (`Available = 1`) and randomized to Mindfulness (`Mindfulness_c = 0.75`) or None (`Mindfulness_c = -0.25`). WCLS estimates the change in `Stress_t30`, using `Prev_Intervention` to reduce noise.

Advantages

- *Simplicity*: Uses known $p = 0.25$, no complex weight modeling.
- *Robustness*: Consistent even if controls are misspecified (Qian et al., 2020).

Limitations

- Assumes randomization eliminates confounding within available cases.
- Best for continuous outcomes (e.g., stress).

Implementation

In SAS, use `PROC GENMOD` with `weight Available`, centered predictors, and `type=ind`, as in our app's analysis.

Required Readings

- Ashley Walton, Inbal Nahum-Shani, Lori Crosby, Predrag Klasnja, and Susan Murphy. Optimizing digital integrated care via micro-randomized trials. *Clinical Pharmacology & Therapeutics*, 104(1):53–58, 2018. doi: 10.1002/cpt.1079.
- Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A. Murphy. Sample size calculations for micro-randomized trials in mHealth. *Statistics in Medicine*, 35(12): 1944–1971, 2016. doi: 10.1002/sim.6847.
- Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A. Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018. doi: 10.1080/01621459.2017.1305274.
- Timothy NeCamp, Srijan Sen, Elena Frank, Maureen A. Walton, Edward L. Ionides, Yu Fang, Ambuj Tewari, and Zhenke Wu. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: Micro-randomized trial. *Journal of Medical Internet Research*, 22(3):e15033, 2020. doi: 10.2196/15033.
- Niam Yaraghi, Ramaswamy Ramesh, and Giri Kumar Tayi. Pay, pat, and clawback: Incentivizing service providers' participation in on-demand digital platforms. *Management Science*, 71(9):7579–7599, 2025. doi: 10.1287/mnsc.2023.02603.

Part IV

Preferences, Scale, and New Tools

Chapter 9

Conjoint Analysis and Preference Elicitation

9.1 Summary of the Previous Session

Imagine you're designing a stress reduction app with three components: mindfulness exercises, social and peer support, and physical exercise. Randomized Controlled Trials (RCTs) can tell you if the app reduces stress, Multiphase Optimization Strategy (MOST) can optimize which components work best together, Micro-Randomized Trials (MRTs) can determine the best time to send mindfulness prompts, and Sequential Multiple Assignment Randomized Trials (SMART) can adapt the app's features based on user progress. But how do you know if users prefer a $5/month app with mindfulness exercises alone or a $15/month app that includes physician-guided exercise? That's where conjoint analysis comes in. Unlike these experimental methods, conjoint analysis reveals *what people value* and *how they choose*—critical insights for designing a product users will adopt before you even build it.

In the last session, each one of you presented an excepriment that was using one of these methods to examine a rwsearch question of your interest.

Your familiarity with RCTs, MOST, MRT, and SMART equips you to test and refine interventions. Here's a quick recap:

- **RCTs**: Measure the causal impact of one intervention versus another.
- **MOST**: Optimize intervention components for efficiency and effect.
- **MRTs**: Assess just-in-time effects of micro-interventions.
- **SMART**: Adapt strategies over time based on response.

Conjoint analysis, however, fills a unique gap by focusing on preference elicitation and trade-off measurement—key for business decisions in strategy, marketing, and product development.

9.2 Types of Questions Conjoint Analysis Addresses

Below are five types of questions conjoint analysis can tackle that RCTs, MOST, MRT, and SMART are not well suited for, with examples tied to our stress reduction app.

1. **How do users weigh trade-offs between multiple attributes when making a choice?**
   Conjoint Strength: Quantifies the relative importance and utility of attributes (e.g., cost vs. mindfulness vs. exercise) and reveals user priorities in a single decision.
   Why Others Fall Short: RCTs test if the app works but don't dissect why one version is preferred; MOST optimizes components but doesn't uncover preferences organically; MRT focuses on timing, not trade-offs; SMART adapts over time but doesn't map static preferences.
   Example: "When choosing a stress app, how much more do users value physical exercise feature over social support feature, and what's the tipping point for price?"
2. **What combination of features maximizes user preference or market share without real-world testing?**
   Conjoint Strength: Simulates market scenarios to predict preferences (e.g., via choice-based conjoint) before launch.
   Why Others Fall Short: RCTs require real-world tests; MOST evaluates existing components, not untested combinations; MRT and SMART focus on interventions, not static predictions.
   Example: "Which bundle (e.g., $10 with mindfulness and support vs. $20 with all three) will attract the most users?"
3. **What are the psychological or emotional benefits driving consumer choices beyond observable features?**
   Conjoint Strength: Benefit-based conjoint digs into motivations (e.g., "calmness" from mindfulness vs. "trust" from peer support) rather than just attributes.
   Why Others Fall Short: RCTs, MOST, MRT, and SMART measure outcomes or behaviors, not latent preferences or emotions.
   Example: "Do users prefer the app for the peace of mindfulness or the trust of peer support, and how much are they willing to pay for each?"
4. **How sensitive are consumers to changes in specific attributes (e.g., price elasticity)?**
   Conjoint Strength: Isolates attribute impacts (e.g., price changes) and calculates willingness-to-pay or sensitivity.
   Why Others Fall Short: RCTs test price effects but not in multi-attribute contexts; MOST and SMART optimize but don't quantify sensitivity; MRT targets timing, not static impacts.

Example: "How much will a $5 price hike affect preference compared to adding physical exercise?"

5. **What is the optimal product positioning or segmentation strategy based on preference heterogeneity?**
   Conjoint Strength: Segments users by preference patterns (e.g., cost-sensitive vs. peer support focused) for tailored strategies.
   Why Others Fall Short: RCTs test average effects, not individual variations; MOST and SMART refine interventions, not market segments; MRT focuses on micro-decisions, not positioning.
   Example: "Which users value social support over cost, and how should we position the app to target them?"

These questions don't replace RCTs, MOST, MRT, or SMART—they complement them by providing the "why" and "what" to design better experiments or strategies. Conjoint analysis equips you to understand user preferences in a way that experimental methods alone cannot, making it a vital tool for your business research toolkit.

9.3 Foundations of Conjoint Analysis

9.3.1 Explanation of Principles

Definition: Conjoint analysis is a quantitative research method that assesses how individuals value different attributes of a product or service by observing their choices or rankings in simulated decision tasks. Unlike direct surveys that rely on stated preferences—where users explicitly report what they like—conjoint analysis infers preferences indirectly by analyzing trade-offs made during these tasks, a process rooted in revealed preference theory. This distinction is critical: insights derived from revealed preferences are generally more reliable than stated ones, as they reflect actual decision-making behavior rather than hypothetical or biased self-reports, providing a robust framework for understanding consumer priorities. For our stress reduction app, it helps us determine how much users value *physical exercise*, *mindfulness*, and *social support* when deciding to subscribe.

Core Principles:

1. *Users Evaluate Bundles*: Participants assess the whole package, not features in isolation.
   Example: A user might weigh mindfulness alongside an affordable price, not just mindfulness alone.
2. *Trade-Offs Matter*: Choices reveal what people are willing to sacrifice.
   Example: Would users pay $15/month for all three features (physical exercise, mindfulness, social support) or opt for $5/month with just mindfulness?
3. *Utility Scores Quantify Preferences*: Conjoint assigns numerical values (utilities) to each feature level, showing their relative importance.

Example: If physical exercise scores 4, mindfulness 3, and social support 2, physical exercise drives choice most.

9.4 Implementing Conjoint Analysis in the Stress Reduction App

Steps in Conjoint Analysis:

1. *Design Attributes and Levels*: Define key features and their variations.

   - **Physical Exercise**: None, Basic (videos), Premium (personalized plans).
   - **Mindfulness**: None, Basic (guided meditations), Premium (live sessions).
   - **Social Support**: None, Basic (chat groups), Premium (peer coaching).
   - **Price**: \$5/month, \$10/month, \$15/month.

2. *Design the Choice Profiles*
   In conjoint analysis, the experimental design determines the combinations of attribute levels (profiles) that respondents evaluate. For our stress reduction app, we have four attributes—Physical Exercise, Mindfulness, Social Support, and Price—each with three levels (e.g., None, Basic, Premium for features; \$5, \$10, \$15 for Price). A full factorial design, which includes every possible combination, results in $3^4 = 81$ profiles. While this captures all main effects and interactions, presenting 81 options to respondents is impractical due to fatigue and time constraints.
   To address this, we use a fractional factorial design, selecting a smaller subset of profiles that still allows us to estimate key effects, typically the main effects (i.e., the independent influence of each attribute on choice). The minimum number of profiles needed depends on the degrees of freedom: with four attributes, each having three levels (2 degrees of freedom per attribute), we require at least $4 \times 2 = 8$ plus 1 (for the overall mean), or 9 profiles, to estimate main effects. In practice, a slightly larger set—such as 12 profiles—improves balance and robustness, ensuring reliable preference estimates.
   The goal is to create an efficient design that is orthogonal (attributes vary independently, minimizing correlation) and balanced (each level appears roughly equally often). For example, a set of 12 profiles might include combinations like Basic Physical, Premium Mindfulness, No Social, \$10, alongside others, carefully chosen to represent the attribute space without redundancy. These profiles are then organized into choice sets—typically pairs or small groups of options, often with a "none" choice—for respondents to evaluate. This reduced design preserves the statistical power to infer

preferences while making the experiment manageable, forming the foundation of our conjoint analysis.

3. *Collect Data*: Present options and ask respondents to choose or rank. Example Choice Set:
   - Option 1: \$5/month, Basic Mindfulness, No Physical Exercise, No Social Support.
   - Option 2: \$15/month, Premium Physical Exercise, Premium Mindfulness, Basic Social Support.

   Respondents pick their preferred option, revealing trade-offs.
4. *Analyze Preferences*: Use statistical models (e.g., regression or logit) to estimate utility scores.

9.4.1 Analyzing Simulated Data

The table below shows the 12 profiles as they are assigned to the 6 choice sets, with each profile defined by its attribute levels. The 'Option' column indicates the option number within the choice set (1 or 2), and 'OptionID' links to the profile's original position in the design.

Table 9.1: All Options in Each Choice Set Before Simulation

ChoiceSet	Option	Physical	Mindful	Social	Price
1	1	1	1	2	3
	2	1	2	3	2
2	1	1	3	1	1
	2	2	1	1	1
3	1	2	2	1	2
	2	2	2	3	1
4	1	2	3	2	2
	2	2	3	3	3
5	1	3	1	3	2
	2	3	2	1	3
6	1	3	2	2	1
	2	3	3	1	2

In this design:

- Each of the 6 choice sets contains 2 unique profiles, resulting in 12 profiles total (e.g., ChoiceSet 1 pairs profiles with levels 1-1-2-3 and 1-2-3-2).
- The profiles are a subset of the full factorial, selected to efficiently estimate

main effects of the attributes on preferences.

- During simulation, a "none" option (all attributes at level 1) is added to each set as Option 3, making 3 options per set for respondents to choose from.
- The final dataset reflects choices from these sets, though not all profiles may be chosen due to utility differences and random variation.

This structure allows us to test preferences across a manageable number of options while covering a broad range of attribute combinations.

```
proc sql;
    create table profile_summary as
    select
        Physical,
        Mindful,
        Social,
        Price,
        count(*) as COUNT, /* Count occurrences of each profile */
        (count(*) / 600.0) * 100 as PERCENT format=8.4 /*
            Percentage out of 600, rounded to 4 decimals */
    from conjoint_data
    group by Physical, Mindful, Social, Price
    order by COUNT desc; /* Sort by count in descending order for
        readability */
quit;
```

Listing 9.1: SAS code to summerize the actual choices

After designing the conjoint experiment with 12 unique profiles paired across 6 choice sets (as shown in the previous table), we simulate choices for 100 respondents, each evaluating all 6 sets. Each choice set offers 3 options: 2 profiles from the original design plus a "none" option (Physical=1, Mindful=1, Social=1, Price=1). Respondents select the option with the highest utility, adjusted by random noise, resulting in 600 total choices (6 sets $\times$ 100 respondents). The table below lists the unique profiles chosen, their frequency (COUNT), and percentage of total choices (PERCENT).

This output shows 11 unique profiles chosen out of the 600 simulated choices, including the "none" option (Obs 1: 1-1-1-1). Compared to the original 12 profiles designed (see previous table):

- **Profile Coverage**: The original design included 12 distinct profiles, but only 10 of these appear here (Obs 2–11), plus "none." Two profiles from our original 12—1 1 2 3 (ChoiceSet 1) and 2 2 1 2 (ChoiceSet 3)—were not chosen. Their utilities were outmatched by competitors in their sets, leaving 10 of the 12 profiles plus "none" in the final 600 choices

We can examine how respondents selected among the three options in each

Table 9.2: Unique Profiles Chosen by Respondents

Obs	Physical	Mindful	Social	Price	COUNT	PERCENT
1	1	1	1	1	8	1.3333
2	1	2	3	2	92	15.3333
3	1	3	1	1	92	15.3333
4	2	1	1	1	8	1.3333
5	2	2	3	1	100	16.6667
6	2	3	2	2	33	5.5000
7	2	3	3	3	67	11.1667
8	3	1	3	2	97	16.1667
9	3	2	1	3	3	0.5000
10	3	2	2	1	47	7.8333
11	3	3	1	2	53	8.8333

of the 6 choice sets. Table 9.3 shows the frequency of chosen options within each choice set, based on 100 respondents per set, totaling 600 choices. These counts reflect respondent preferences driven by utility values and random noise, revealing which options were favored in each set.

```
proc sql;
    proc freq data=conjoint_data;
    tables ChoiceSet * ChosenOption / nopercent norow nocol out=
        choices_by_set;
    title "Frequency of Chosen Options Within Each Choice Set";
run;
quit;
```

Listing 9.2: SAS code examine the chosen profile within each set

Table 9.3: Frequency of Chosen Options Within Each Choice Set

ChoiceSet	Option 1	Option 2	Option 3	Total
1	0	92	8	100
2	92	8	0	100
3	0	100	0	100
4	33	67	0	100
5	97	3	0	100
6	47	53	0	100
Total	269	323	8	600

In our conjoint analysis, we should expand the original simulated dataset

from 600 to 1800 observations by adding 1200 additional rows to include all possible options within each choice set. Initially, the dataset contained only the 600 chosen options (one per respondent per choice set), reflecting the selections of 100 respondents across 6 sets. However, choice-based conjoint models, such as those fitted by `PROC MDC` or `PROC PHREG`, require data on all alternatives—chosen and unchosen—to estimate part-worth utilities by comparing preferences across options. By adding the 1200 unchosen options (2 per set $\times$ 600 sets), we created a full choice set structure with 3 options per set (2 profiles + "none"), enabling the model to assess the relative utility of each attribute level within the context of every decision, thus ensuring accurate and robust estimation of respondent preferences.

```
/* Recreate choice_sets_full (6 sets x 3 options) */
data choice_sets_full;
    input ChoiceSet Option Physical Mindful Social Price;
    datalines;
    1 1 1 1 2 3  /* ChoiceSet 1: Option 1 */
    1 2 1 2 3 2  /* ChoiceSet 1: Option 2 */
    1 3 1 1 1 1  /* ChoiceSet 1: Option 3 (none) */
    2 1 1 3 1 1  /* ChoiceSet 2: Option 1 */
    2 2 2 1 1 1  /* ChoiceSet 2: Option 2 */
    2 3 1 1 1 1  /* ChoiceSet 2: Option 3 (none) */
    3 1 2 2 1 2  /* ChoiceSet 3: Option 1 */
    3 2 2 2 3 1  /* ChoiceSet 3: Option 2 */
    3 3 1 1 1 1  /* ChoiceSet 3: Option 3 (none) */
    4 1 2 3 2 2  /* ChoiceSet 4: Option 1 */
    4 2 2 3 3 3  /* ChoiceSet 4: Option 2 */
    4 3 1 1 1 1  /* ChoiceSet 4: Option 3 (none) */
    5 1 3 1 3 2  /* ChoiceSet 5: Option 1 */
    5 2 3 2 1 3  /* ChoiceSet 5: Option 2 */
    5 3 1 1 1 1  /* ChoiceSet 5: Option 3 (none) */
    6 1 3 2 2 1  /* ChoiceSet 6: Option 1 */
    6 2 3 3 1 2  /* ChoiceSet 6: Option 2 */
    6 3 1 1 1 1  /* ChoiceSet 6: Option 3 (none) */
    ;
run;

/* Merge to create full choice set data (1800 rows) */
proc sql;
    create table full_choices as
    select
        a.ChoiceSet,
        a.Respondent,
        b.Option,
        b.Physical,
        b.Mindful,
        b.Social,
        b.Price,
        case when a.ChosenOption = b.Option then 1 else 0 end as
            Chosen
```

```
    from conjoint_data a
    cross join choice_sets_full b
    where a.ChoiceSet = b.ChoiceSet;
quit;

/* Confirm totals */
proc freq data=full_choices;
    tables Chosen;
    title "Verification: Count of Chosen vs. Unchosen Options";
run;
```

Listing 9.3: SAS code to prepare the data for analysis

In choice-based conjoint (CBC) analysis, we estimate part-worth utilities to understand how attribute levels influence respondent preferences across multiple options. Two SAS procedures, `PROC MDC` and `PROC PHREG`, can achieve this, though they originate from different econometric traditions: discrete choice modeling and survival analysis, respectively. Despite their distinct purposes, they are equivalent for CBC under specific conditions, as demonstrated below using econometric specifications.

Consider a choice set C_{ir} for respondent i in choice set r, containing $J = 3$ options (e.g., two profiles and "none"). Each option j has attributes $X_{jir} = (X_{jir1}, X_{jir2}, \ldots, X_{jirk})$ (e.g., Physical, Mindful, Social, Price levels), and the respondent chooses one option, denoted by $y_{ir} = j$.

Conditional Logit Model (PROC MDC) In `PROC MDC`, the conditional logit model assumes the utility of option j for respondent i in choice set r is:

$$V_{jir} = X_{jir}\beta = \beta_1 X_{jir1} + \beta_2 X_{jir2} + \cdots + \beta_k X_{jirk}$$

where β is a vector of part-worth utilities. The probability of choosing option j is:

$$P(y_{ir} = j | C_{ir}) = \frac{e^{V_{jir}}}{\sum_{k \in C_{ir}} e^{V_{kir}}}$$

The log-likelihood is maximized over all choices:

$$LL = \sum_{i=1}^{N} \sum_{r=1}^{R} \log \left(\frac{e^{X_{y_{ir}}\beta}}{\sum_{k \in C_{ir}} e^{X_{kir}\beta}} \right)$$

Here, $N = 100$ respondents, $R = 6$ choice sets, and $J = 3$ options per set. `PROC MDC` directly estimates β using this formulation, requiring data with all alternatives (1800 rows: $6 \times 100 \times 3$).

Stratified Cox Model (PROC PHREG) In `PROC PHREG`, designed for survival analysis, the hazard rate for an event (e.g., death) at time t is:

$$h(t|X) = h_0(t)e^{X\beta}$$

where $h_0(t)$ is the baseline hazard, and $X\beta$ is the linear predictor. For CBC, we adapt this by:

- Treating the choice event as instantaneous (no time dimension).
- Defining "chosen" (`Chosen`=1) as the event, and unchosen options (`Chosen`=0) as censored.
- Stratifying by `ChoiceSet` and `Respondent` (600 strata), ensuring comparisons within each choice set.

The partial likelihood for a stratum (choice set C_{ir}) with one chosen option j is:

$$L_{ir} = \frac{e^{X_{jir}\beta}}{\sum_{k \in C_{ir}} e^{X_{kir}\beta}}$$

The total log-likelihood across all strata is:

$$LL = \sum_{i=1}^{N} \sum_{r=1}^{R} \log \left(\frac{e^{X_{jir}\beta}}{\sum_{k \in C_{ir}} e^{X_{kir}\beta}} \right)$$

where j is the chosen option in C_{ir}.

Equivalence Comparing the log-likelihoods:

- `PROC MDC`: $LL = \sum_{i=1}^{N} \sum_{r=1}^{R} \log \left(\frac{e^{X_{y_{ir}}\beta}}{\sum_{k \in C_{ir}} e^{X_{kir}\beta}} \right)$
- `PROC PHREG`: $LL = \sum_{i=1}^{N} \sum_{r=1}^{R} \log \left(\frac{e^{X_{jir}\beta}}{\sum_{k \in C_{ir}} e^{X_{kir}\beta}} \right)$

Both are identical, as $y_{ir} = j$ represents the chosen option. In `PROC PHREG`, the baseline hazard $h_0(t)$ cancels out within each stratum (since only one event occurs), leaving the same utility form as conditional logit. Thus, the β estimates (part-worth utilities) from `PROC PHREG` match those from `PROC MDC`'s conditional logit model, adjusted for parameterization (e.g., reference coding with `ref='1'`).

```
1
2  /*Estimate Part-Worth Utilities with PROC MDC */
3  /*This does not seem to work*/
4  proc mdc data=full_choices;
5      class Physical  Mindful Social Price ;
6      model Chosen = Physical Mindful Social Price /
7            type=clogit
8            nchoice=3; /* 3 alternatives per choice set */
9      id SetRespID_num; /* Numeric ID for each ChoiceSet-Respondent
           pair */
10     title "Conditional Logit Model for Utility Estimation";
11 run;
12 /*Estimate Part-Worth Utilities with PROC PHREG */
13 proc phreg data=full_choices;
14     class Physical (ref='1') Mindful (ref='1') Social (ref='1')
          Price (ref='1') / param=reference;
```

```
    model Chosen = Physical Mindful Social Price / ties=breslow;
    strata ChoiceSet Respondent;
    title "Conditional Logit Model for Utility Estimation";
run;
```

Listing 9.4: SAS code run the regression model

9.4.2 Interpreting Part-Worth Utilities from PROC PHREG

Table 9.4 below presents the maximum likelihood estimates, with reference levels set to 1 (None for features, $5 for Price), reflecting utilities relative to these baselines. Each parameter's significance is assessed via chi-square tests, and hazard ratios indicate the relative likelihood of selection compared to the reference level.

Table 9.4: Part-Worth Utility Estimates from PROC PHREG (Reference Level = 1)

Parameter	DF	Estimate	Std. Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Physical 2	1	0.41933	0.11965	12.2819	0.0005	1.521	Physical 2
Physical 3	1	-0.11551	0.15362	0.5654	0.4521	0.891	Physical 3
Mindful 2	1	-0.47330	0.13331	12.6048	0.0004	0.623	Mindful 2
Mindful 3	1	-0.87108	0.14165	37.8158	<0.0001	0.418	Mindful 3
Social 2	1	-0.33045	0.10411	10.0739	0.0015	0.719	Social 2
Social 3	1	-2.17274	0.16658	170.1171	<0.0001	0.114	Social 3
Price 2	1	0.13174	0.13815	0.9095	0.3403	1.141	Price 2
Price 3	1	0.68197	0.12117	31.6778	<0.0001	1.978	Price 3

The results reveal respondent preferences relative to the absence of features (None, level 1) and a $5 price point (Price level 1). For Physical Exercise, Basic (2) significantly increases utility by 0.419 (p=0.0005) over None, while Premium (3) slightly reduces it by -0.116 (p=0.4521), suggesting a preference for moderate exercise over none or premium options, though the latter is not significant.

Mindfulness shows a significant aversion, with Basic (2) decreasing utility by -0.473 (p=0.0004) and Premium (3) by -0.871 (p<0.0001) compared to None, indicating respondents disfavor mindfulness features.

Social Support follows a similar trend, with Basic (2) reducing utility by -0.330 (p=0.0015) and Premium (3) sharply decreasing it by -2.173 (p<0.0001), reflecting a strong rejection of social features, especially at premium levels.

For Price, $10 (2) slightly increases utility by 0.132 (p=0.3403), but not significantly, while $15 (3) significantly boosts it by 0.682 (p<0.0001), a counterintuitive result suggesting higher prices are preferred over $5. This anomaly likely stems from a simulation coding error (e.g., Price levels flipped: $5 as 3, $15 as 1), inconsistent with typical price sensitivity. At the end of the day, I am only a human and at this point, I really do not want to go back and fix my simulation code ;)

Required Readings

- Vithala R. Rao. Conjoint analysis. In Jagdish N. Sheth and Naresh K. Malhotra, editors, *Wiley International Encyclopedia of Marketing*. John Wiley & Sons, Chichester, UK, 2010. doi: 10.1002/9781444316568.wiem02019.
- Dong Soo Kim, Roger A. Bailey, Nino Hardt, and Greg M. Allenby. Benefit-based conjoint analysis. *Marketing Science*, 36(1):54–69, 2017. doi: 10.1287/mksc.2016.1003.
- Mohamed A. Hussein, Courtney Lee, and S. Christian Wheeler. How do consumers react to ads that meddle in out-party primaries? *Journal of Consumer Research*, 51 (6):1186–1208, 2025. doi: 10.1093/jcr/ucae039.
- Simon J. D. Schillebeeckx, Sankalp Chaturvedi, Gerard George, and Zella King. What do I want? The effects of individual aspiration and relational capability on collaboration preferences. *Strategic Management Journal*, 37(7):1493–1506, 2016. doi: 10.1002/smj.2396.
- Niam Yaraghi, Weiguang Wang, Guodong Gao, and Ritu Agarwal. How online quality ratings influence patients' choice of medical providers: Controlled experimental survey study. *Journal of Medical Internet Research*, 20(3):e99, 2018. doi: 10.2196/jmir.8986.

Chapter 10

Scaling Experiments and Heterogeneous Treatment Effects

10.1 Summary of the Previous Session

Throughout this course, we have explored a comprehensive range of experimental designs, building a robust foundation for conducting rigorous scientific research. We began with randomized controlled trials (RCTs), the gold standard for establishing causality, and progressed to more complex frameworks such as Just-in-Time Adaptive Interventions (JITAIs), Micro-Randomized Trials (MRTs), and Sequential Multiple Assignment Randomized Trials (SMARTs). Our primary focus has been on methodological rigor—ensuring reliable, valid results through careful design, randomization, and control of confounding variables to support causal inference.

Having covered the spectrum of experimental design methodologies, we are now well-equipped to shift our focus. In today's session, we will explore practical challenges in implementing experiments, such as logistical constraints and real-world applicability. This session bridges the gap between theoretical design and actionable, innovative research practice.

10.2 Scalability in Experiments

According to Al-Ubaydli et al. (2019), many experiments fail to scale because of what they call the **Voltage Effect**: the idea that the impact (or "voltage") of an intervention diminishes as it moves from a controlled trial to broader implementation.

Scaling is hard not just because of logistical complexity, but because the effect size observed in pilot studies is often inflated, unstable, or conditional on specific circumstances that don't generalize. The authors highlight three critical threats

to scalability:

1. **Statistical Inference** – *Are the results even real?*
    - False positives due to small sample sizes or noisy data
    - Publication bias and selective reporting inflate effect sizes
    - Winner's curse: the largest effects may come from random noise
2. **Representativeness of the Population** – *Does this work for the general public?*
    - Pilot subjects may be more motivated, educated, or incentivized
    - Results don't generalize to broader, more heterogeneous populations
3. **Representativeness of the Situation** – *Does the setting match real-world conditions?*
    - Implementation fidelity often drops at scale
    - Admin quality, delivery agents, or context vary widely
    - Network effects and diseconomies of scale can distort results

The **Head Start home visiting program** was an early childhood intervention targeting at-risk families. In its initial trials, the program showed strong effects on child development and parenting outcomes.

However, when scaled up, the results diminished significantly. Studies found that:

- Home visits to "at-risk" families were often interrupted by distractions
- Less time was spent on child-focused activities during visits
- Program delivery was inconsistent across facilitators

This is a textbook case of the Voltage Effect: the situation changed (threat #3), the fidelity declined, and the population scaled up in ways the original study didn't account for.

How can we design experiments that actually work when taken beyond the lab or pilot setting?

Scaling is not just about increasing size — it's about maintaining impact across broader populations, messier environments, and lower control. To design experiments that truly work at scale, we must anticipate threats and build solutions directly into the design phase.

Five Strategies for Designing Scalable Experiments

1. **Anticipate threats to scale in the design**
   - Use List and Suskind's three-bin framework to evaluate:
     - Statistical inference risk
     - Population representativeness
     - Situational fidelity
2. **Use large, diverse, real-world samples**
   - Avoid "WEIRD" subject pools
   - Design for generalizability from the outset
3. **Leverage digital platforms for experimentation**
   - Conduct A/B tests, and adaptive trials online
   - Use real-time dashboards for iterative learning and deployment
4. **Measure and monitor implementation fidelity**
   - Track delivery agents, dosage, and adherence at scale
   - Prevent program drift through standardization or tech-enabled delivery
5. **Combine traditional methods with predictive modeling**
   - Use machine learning to estimate heterogeneous treatment effects
   - Improve counterfactual estimates and policy targeting

Case Example: Amazon Cart Recommendations

Amazon tested personalized cart recommendations despite leadership opposition. Rather than relying on expert opinion, they ran a large-scale online experiment.

- The test occurred on live traffic — high external validity
- Digital platform ensured fidelity and consistent delivery
- The effect was robust and profitable — led to full-scale rollout

This example demonstrates how modern experimentation infrastructure can reduce scalability threats and allow evidence-based product decisions.

10.3 Application of Machine Learning to Estimate Heterogeneous Treatment effects

Traditional RCTs focus on estimating the **average treatment effect (ATE)**:

$$\text{ATE} = E[Y(1) - Y(0)]$$

However, in many real-world settings, the treatment effect may vary across individuals or subgroups. The goal is to estimate:

$$\tau(x) = E[Y(1) - Y(0) \mid X = x]$$

which is known as the **Conditional Average Treatment Effect (CATE)**. Machine learning (ML) can help estimate this function, enabling personalized policy decisions and improving external validity.

Step-by-Step Framework

1. **Run an RCT.** Randomly assign units to treatment and control groups to ensure internal validity.
2. **Collect covariates and outcomes.** For each unit i, observe:

   $$(X_i, D_i, Y_i)$$

   where X_i are pre-treatment covariates, $D_i \in \{0, 1\}$ is treatment assignment, and Y_i is the outcome.
3. **Use ML to estimate treatment effects conditional on X_i.**

Two Approaches to Estimating CATE

1. T-Learner (Two-Model Approach)

The T-learner separately estimates potential outcomes under treatment and control.

1. **Split the Data:**
   - Divide your dataset into treatment ($W = 1$) and control ($W = 0$) groups.
   - Each group has outcomes Y and covariates X.
2. **Fit Two Models:**
   - Train a model $\mu_1(X)$ on the treatment group to predict Y given X.
   - Train a separate model $\mu_0(X)$ on the control group to predict Y given X.
   - Any supervised learning model can be used (e.g., linear regression, random forests, neural networks), depending on the data complexity.

3. **Estimate CATE:**
    - For any individual with covariates $X = x$, compute:

$$\text{CATE}(x) = \mu_1(x) - \mu_0(x)$$

    - This gives the estimated treatment effect conditional on X.

Advantages:

- Simple and modular: any regression model can be used for $\hat{\mu}_1$ and $\hat{\mu}_0$
- Flexible for high-dimensional covariates

Limitations:

- If the treated and control groups differ greatly in X, each model is trained on different support $\rightarrow$ extrapolation becomes risky
- No built-in mechanism for prioritizing treatment effect heterogeneity during training

2. Causal Forests (Wager & Athey, 2018)

Causal Forests, part of the Generalized Random Forest framework, directly estimate CATE by adapting random forests to focus on treatment effect heterogeneity. They partition the covariate space to find regions where treatment effects differ, optimizing for causal inference rather than just prediction.

Before discussing Causal Forests, it is useful to understand standard Random Forests. A Random Forest is an ensemble learning method that constructs multiple decision trees, each trained on a random subset of the data and features. It predicts outcomes by averaging the predictions of all trees, excelling at capturing complex, non-linear relationships in data. Random Forests are widely used for prediction tasks but do not inherently focus on estimating causal effects or treatment heterogeneity.

Causal Forests build on the Random Forest framework by modifying the splitting criterion to prioritize heterogeneity in treatment effects rather than predictive accuracy. This adaptation links the two methods: while Random Forests optimize for outcome prediction, Causal Forests optimize for estimating the conditional average treatment effect, making them a specialized tool for causal inference.

Steps to Implement

1. **Prepare Data:**
    - Ensure you have X (covariates), W (treatment indicator, binary), and Y (outcome).
    - Data should be from a randomized experiment or observational study with unconfoundedness ($\{Y(1), Y(0)\} \perp W \mid X$), meaning that the data used to estimate causal effects must satisfy conditions ensuring

that the treatment assignment is not systematically biased in a way that distorts the relationship between the treatment and outcomes.

2. **Fit Causal Forest:**
   - Use a Causal Forest algorithm (e.g., from the `grf` package in R or `econml` in Python).
   - The algorithm builds trees that split the covariate space to maximize heterogeneity in treatment effects, estimating:

$$\text{CATE}(x) = E[Y(1) - Y(0) \mid X = x]$$

     Imagine studying a tutoring program's effect on test scores ((Y), with covariates like prior grades and study hours (X). A Causal Forest might split students by prior grades, finding that tutoring boosts scores more for low-performing students than high-performing ones, estimating different CATEs for each group
   - It accounts for propensity scores ($P(W = 1 \mid X)$) to handle confounding. Causal Forests estimate propensity scores (e.g., using a separate model or internally) to understand how X influences treatment assignment. When estimating CATE, they use these scores to reweight or adjust comparisons between treated and control groups, ensuring that individuals with similar X are compared fairly. This adjustment mimics randomization, reducing bias from confounding under the unconfoundedness assumption ($\{Y(1), Y(0)\} \perp W \mid X$), meaning all confounders are captured in X.
3. **Predict CATE:**
   - For any $X = x$, the fitted Causal Forest outputs the estimated $\text{CATE}(x)$.
4. **Evaluate:**
   - Check model calibration (e.g., using out-of-bag predictions).
   - Analyze variable importance to understand which covariates drive heterogeneity.
   - Validate on a holdout set if feasible.

Advantages:

- Built for estimating heterogeneity directly
- Uses *honest estimation*: splits are trained on one half of the data, effect sizes are estimated on the other
- Provides valid confidence intervals for $\hat{\tau}(x)$ under regularity assumptions

Limitations:

- Less interpretable than simple models

- Requires careful tuning and validation

Use Cases

- **Policy targeting:** Deliver interventions only to subgroups with high $\hat{\tau}(x)$
- **Resource allocation:** Prioritize high-impact populations
- **Generalization:** Simulate effects for new populations by applying $\hat{\tau}(x)$ to new X values

Key References

- Susan Athey and Stefan Wager. "Estimating Treatment Effects with Causal Forests," *JASA*, 2018.
- Kunzel, Sekhon, Bickel, and Yu. "Meta-learners for Estimating Heterogeneous Treatment Effects," *PNAS*, 2019.

Required Readings

- Hal R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016. doi: 10.1073/pnas.1510479113.
- Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009. doi: 10.1007/s10618-008-0114-1.
- Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981, 2019. doi: 10.1287/mnsc.2018.3093.
- Omar Al-Ubaydli, John A. List, and Dana Suskind. The science of using science: Towards an understanding of the threats to scaling experiments. Working Paper 25848, National Bureau of Economic Research, 2019. Published as the 2017 Klein Lecture, *International Economic Review* 61(4):1387–1409, 2020.

Chapter 11

Large Language Models in Experimental Research

11.1 Summary of the Previous Session

So far, we've been discussing various experimental designs intended to study human behaviors and decision-making directly. However, with the rapid advancements and widespread adoption of Large Language Models (LLMs) and Generative AI, we're now at an exciting turning point. In this session, we will explore how these cutting-edge AI tools can significantly enhance the way we design, implement, and analyze experiments, opening new avenues for more efficient, scalable, and insightful experimental research.

11.2 Why LLMs Matter for Experiments

- LLMs can process economic scenarios, simulate participant responses, and generate experimental materials (e.g., instructions, code) with minimal human effort.
- Their scalability allows researchers to run large-scale simulations at low cost, unlike human-based experiments, which are resource-intensive.
- They address challenges in online experiments, such as participant comprehension and compliance, by providing tailored interactions.

11.2.1 Homo Silicus as a Computational Model

Definition: John Horton introduces *Homo Silicus* as an analogy to the classical *Homo Economicus*. While *Homo Economicus* assumes perfect rationality to model human behavior, *Homo Silicus* represents Large Language Models (LLMs) acting as computational models of humans. These models simulate behavior by drawing

from vast, real-world training data—capturing not only rational choices but also heuristic-driven and sometimes irrational decisions.

Significance: *Homo Silicus* bridges theoretical and empirical research. Like *Homo Economicus*, it helps economists explore behavior in abstracted scenarios—but with a critical advantage: LLMs simulate responses that are dynamic, data-driven, and more reflective of the messy realities of human decision-making.

Relevance to Business: In business contexts, *Homo Silicus* enables rapid and low-cost piloting of strategies—such as pricing or incentive structures—by simulating consumer and employee responses.

11.2.2 Mechanism: Endowing LLMs for Simulation

How It Works: LLMs can be "endowed" with specific characteristics—like preferences or beliefs—via prompt engineering, allowing them to act as agents in economic scenarios. Unlike static models, LLMs use computational simulation, grounded in their exposure to vast textual data.

Technical Details:

- **Prompt Engineering:** Prompts serve as experimental treatments. For instance, "You are a low-income consumer" frames the model's role, much like assigning participants in a behavioral lab .
- **Stochastic Responses:** LLMs produce varied outputs using parameters like temperature (e.g., 0.8), enabling simulations of heterogeneous agents. Repeated API calls generate a sample of responses.
- **Training Data:** LLMs are trained on discussions of economic behavior (e.g., bargaining, consumption), allowing them to simulate realistic decision patterns based on embedded social norms and knowledge.

Here is an example of how we can use these tools:
`https://colab.research.google.com/drive/13S7eM4Gh94paBgHsQCubA_dudLoo3UcV?usp=sharing`

11.2.3 Advantages Over Homo Economicus

- **Realism:** *Homo Silicus* simulates behaviors like fairness, loss aversion, and status quo bias—features absent in perfectly rational models.
- **Flexibility:** LLMs can instantly switch roles (e.g., manager, worker, customer) through prompts, enabling diverse scenario testing.
- **Scalability and Cost-Efficiency:** Experiments using LLMs are inexpensive and fast. Horton ran experiments costing $\sim$ \$50 in minutes. A ride-sharing simulation with 5 API calls cost less than \$0.10.
- **Business Applications:** Firms can simulate consumer reactions to dynamic pricing or test employee incentives before real-world implementation, aligning with the tech industry's shift toward data-driven experimentation.

11.2.4 Limitations and Considerations

- **Performativity:** LLMs might echo textbook answers if prompted with familiar scenarios.
- **Training Biases:** Models trained on Western-centric data may not generalize globally.
- **Hallucinations:** LLMs can generate convincing yet inaccurate outputs, requiring human oversight .
- **Prompt Dependency:** Outcomes vary with prompt quality, necessitating thoughtful design .

Mitigation: Design clear prompts, compare against real data, and log interactions to ensure transparency and replicability. Also read the paer by Yang et al. (2024) [Are large language models (llms) good social predictors? arXiv preprint arXiv:2402.12620, 2024] as it seems to pour cold water on many of these expectations.

11.3 How LLMs Improve the Experiments

11.3.1 Enhancing Experiment Design

Research Question Formulation

LLMs assist researchers in identifying promising hypotheses by analyzing vast bodies of academic literature and economic trends. They detect gaps, suggest variables of interest, and recommend experiment types such as economic games or randomized controlled trials.

- **Prompt Engineering:** Prompts like "Based on recent literature on consumer pricing, suggest a hypothesis for a field experiment" elicit hypotheses grounded in prior studies and contextual knowledge.
- **Trend Extraction:** LLMs infer emerging topics and behavioral patterns from training data such as journal articles, online discussions, and working papers.

Treatment and Setting Selection

LLMs guide the selection of experimental settings (e.g., lab, field, online) and help ensure appropriate randomization strategies to satisfy key assumptions like SUTVA and exclusion restrictions. They also recommend sample sizes, stratification methods, and ethical considerations.

- **Algorithmic Suggestions:** LLMs propose sample sizes, assess statistical power and effect sizes, and optimize randomization strategies under given resource constraints.

- **Simulation:** Researchers can simulate treatment effects using LLMs to test robustness and fine-tune design parameters.

Example: A researcher studying employee incentives might prompt, "Suggest an experimental setting and randomization strategy for testing bonus structures," and receive detailed, feasible recommendations.

Instruction Wording

LLMs generate clear, concise, and accessible instructions tailored to diverse participant demographics. By accounting for variations in education and language proficiency, they reduce measurement error and improve participant compliance.

- **Natural Language Processing:** LLMs adjust tone and complexity to match target populations, using techniques from tools like QualiGPT.
- **Iterative Refinement:** LLMs generate multiple instruction drafts, integrating feedback from pilot runs to improve clarity and engagement.

Example: In a ride-sharing simulation, GPT-3.5-turbo was prompted: "You are a low-income commuter. Rate a 50% ride-sharing price increase as Completely Fair, Acceptable, Unfair, or Very Unfair, and explain why." The model produced clear instructions and comprehension checks that helped detect ambiguity.

Code and Documentation

LLMs accelerate the implementation of experiments by generating clean, usable code and standardized documentation. They support multiple platforms like Python, Qualtrics, and oTree, and can create IRB documentation and replication packages for full transparency.

- **Code Generation:** LLMs convert plain-language task descriptions into executable code across different platforms.
- **Documentation:** They produce structured replication packages, including pre-registration templates and IRB documents.

Example: A researcher prompts, "Generate Python code for a Qualtrics survey on consumer pricing preferences." GPT-4 generates both the logic and documentation, reducing human error and promoting reproducibility.

11.3.2 Enhancing the Implementation of Experiments

Conversational Support

AI chatbots enhance online experiments by providing real-time, personalized support. They clarify instructions and reduce confusion, improving participant comprehension and compliance—addressing common challenges such as measurement errors and dropout.

- **Conversational AI:** LLMs use natural language processing to simulate human-like interactions and tailor responses to individual participant profiles (e.g., language proficiency).
- **API Integration:** Real-time API calls allow chatbots to handle large volumes of simultaneous queries, making them scalable for large experiments.

Example: In an online survey on employee incentives, a participant asks, "What does 'performance-based bonus' mean?" The AI chatbot responds: "It's a bonus tied to your task performance, like meeting sales targets." This ensures clarity and reduces attrition.

Simulation of Social Interactions

LLMs can simulate realistic social dynamics—such as bargaining, group decision-making, or negotiations—thereby enhancing the ecological validity of experimental designs.

- **Role-Playing:** LLMs adopt specific roles (e.g., buyer, seller) via prompts and generate contextually grounded dialogue based on behavioral patterns in their training data.
- **Multi-Agent Simulation:** Multiple API calls simulate interactions among distinct agents, enabling the modeling of strategic behavior and social preferences.

Example: In a dictator game simulation, GPT-3 agents endowed with different preferences (e.g., fairness, efficiency) produced allocation choices that mirrored real human behavior, enabling a controlled study of social norms.

Engagement and Monitoring

AI tools monitor participant engagement in real-time, reduce attrition, and detect non-compliant behaviors (e.g., cheating). Interactive prompts keep participants focused and ensure high-quality data.

- **Behavioral Tracking:** LLMs integrate with browser-based plugins to detect tab-switching or inactivity and flag potential disengagement.
- **Engagement Prompts:** Dynamic reminders such as "Please focus on the task" help re-engage participants and maintain data quality.

Example: In an online study, the system detects repeated tab-switching. The AI prompts: "Please stay on the survey page to ensure accurate responses." This intervention helps reduce noise and preserve data integrity.

Here is an example of how we can use AI tools to simulate a negotiation:
`https://colab.research.google.com/drive/13S7eM4Gh94paBgHsQCubA_dudLoo3UcV?usp=sharing`

Required Readings

- Gary Charness, Brian Jabarian, and John A. List. Generation next: Experimentation with AI. Working Paper 31679, National Bureau of Economic Research, 2023.
- John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, 2023.
- Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, and Hui Liu. Are large language models (LLMs) good social predictors? arXiv preprint arXiv:2402.12620, 2024.

Appendix A

Research Ethics and IRB Training

A.1 Ethical and Practical Challenges in Experimental Research

The course devotes a full session to the ethical complexities of experimental research, particularly in field settings: the tension between scientific integrity and participant autonomy, informed consent, harm minimization, and societal impact; frameworks for managing ethical dilemmas; and case studies of how researchers have handled them. There are no lecture notes for that session — it runs as a discussion of the readings below — so they are collected here rather than at the end of a chapter.

Required Readings

- Scott Desposato. Ethical challenges and some solutions for field experiments. Working paper, University of California, San Diego. `http://www.desposato.org/ethicsfieldexperiments.pdf`, 2014.
- Trisha Phillips. Ethics of field experiments. *Annual Review of Political Science*, 24(1): 277–300, 2021. doi: 10.1146/annurev-polisci-041719-101956.
- Jonathan Lewis. Experimental design: Ethics, integrity, and the scientific method. In Ron Iphofen, editor, *Handbook of Research Ethics and Scientific Integrity*, pages 459–474. Springer, Cham, Switzerland, 2020. doi: 10.1007/978-3-030-16759-2_19.

A.2 Registering for Required IRB Training at the University of Miami

University of Miami (UM) students must complete the **Collaborative Institutional Training Initiative (CITI)** program before obtaining IRB approval for any human subjects experiment. Below are the steps:

1. **Understand the Requirement**: All team members conducting human subjects research must complete CITI training for IRB approval, covering ethical and regulatory standards.
2. **Access CITI**: Go to `https://www.citiprogram.org`.
3. **Register**:
   - Click "Register" (top right).
   - Select "**University of Miami Jackson Health Systems (SSO)**" as your institution.
   - It will take you to UM's SSO login page.
   - Agree to terms, then use your UM email (e.g., `yourname@miami.edu`) and password.
4. **Select Course**: Under "Add a Course," pick "Human Subjects Research" and choose:
   - "Social & Behavioral Research – Basic/Refresher" for behavioral studies (4-6 hours).
   - you will be taken through a series of options for other courses, you should skip them all until your reach the final step.

   Then submit.
5. **Complete Training**: Finish 11-15 modules with quizzes (80% to pass), then download your certificate.
6. **Submit to IRB**: Attach the certificate to your IRB application in IBISResearch (`https://ibis-research.miami.edu/IRB`).
7. **Get Help**: Email `HSRO@miami.edu` or call 305-243-3195 for assistance.

Tip: Complete early (valid for 3 years) and save your certificate. You can always access it through your account on the CITI website.

Bibliography

Omar Al-Ubaydli, John A. List, and Dana Suskind. The science of using science: Towards an understanding of the threats to scaling experiments. Working Paper 25848, National Bureau of Economic Research, 2019. Published as the 2017 Klein Lecture, *International Economic Review* 61(4):1387–1409, 2020.

Omar Al-Ubaydli, Min Sok Lee, John A. List, Claire L. Mackevicius, and Dana Suskind. How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, 5(1): 2–49, 2021. doi: 10.1017/bpp.2020.17.

Reena Bhardwaj, Usha Agrawal, Praveen Vashist, and Souvik Manna. Determination of sample size for various study designs in medical research: A practical primer. *Journal of Family Medicine and Primary Care*, 13(7):2555–2561, 2024. doi: 10.4103/jfmpc.jfmpc_1675_23.

Theophile Bigirumurame, Germaine Uwimpuhwe, and James Wason. Sequential multiple assignment randomized trial studies should report all key components: A systematic review. *Journal of Clinical Epidemiology*, 142:152–160, 2022. doi: 10.1016/j.jclinepi.2021.11.007.

Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A. Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018. doi: 10.1080/01621459.2017.1305274.

Francesco Cecchi, Robert Lensink, and Edwin Slingerland. Ambiguity attitudes and demand for weather index insurance with and without a credit bundle: Experimental evidence from Kenya. *Journal of Behavioral and Experimental Finance*, 41:100885, 2024. doi: 10.1016/j.jbef.2023.100885.

Gary Charness, Brian Jabarian, and John A. List. Generation next: Experimentation with AI. Working Paper 31679, National Bureau of Economic Research, 2023.

Linda M. Collins. *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*. Statistics for

Social and Behavioral Sciences. Springer, Cham, Switzerland, 2018. ISBN 978-3-319-72205-4. doi: 10.1007/978-3-319-72206-1.

M. O. Columb and M. S. Atkinson. Statistical analysis: Sample size and power estimations. *BJA Education*, 16(5):159–161, 2016. doi: 10.1093/bjaed/mkv034.

Scott Desposato. Ethical challenges and some solutions for field experiments. Working paper, University of California, San Diego. `http://www.desposato.org/ethicsfieldexperiments.pdf`, 2014.

Yue Katherine Feng, Jennifer L. Claggett, Elena Karahanna, and Kar Yan Tam. A randomized field experiment to explore the impact of herding cues as catalysts for adoption. *MIS Quarterly*, 46(2):1135–1164, 2022. doi: 10.25300/MISQ/2022/16141.

Guillaume R. Fréchette, Kim Sarnoff, and Leeat Yariv. Experimental economics: Past and future. *Annual Review of Economics*, 14(1):777–794, 2022. doi: 10.1146/annurev-economics-081621-124424.

Suat Gonul, Tuncay Namli, Sasja Huisman, Gokce Banu Laleci Erturkmen, Ismail Hakki Toroslu, and Ahmet Cosar. An expandable approach for design and personalization of digital, just-in-time adaptive interventions. *Journal of the American Medical Informatics Association*, 26(3):198–210, 2019. doi: 10.1093/jamia/ocy160.

Amanda L. Graham, George D. Papandonatos, Sarah Cha, Michael S. Amato, Megan A. Jacobs, Amy M. Cohn, Lorien C. Abroms, and Robyn Whittaker. Effectiveness of an optimized text message and Internet intervention for smoking cessation: A randomized controlled trial. *Addiction*, 117(4):1035–1046, 2022. doi: 10.1111/add.15677.

Zhuojun Gu, Ravi Bapna, Jason Chan, and Alok Gupta. Measuring the impact of crowdsourcing features on mobile app user engagement and retention: A randomized field experiment. *Management Science*, 68(2):1297–1329, 2022. doi: 10.1287/mnsc.2020.3943.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Mohamed A. Hussein, Courtney Lee, and S. Christian Wheeler. How do consumers react to ads that meddle in out-party primaries? *Journal of Consumer Research*, 51(6):1186–1208, 2025. doi: 10.1093/jcr/ucae039.

Kelley M. Kidwell and Daniel Almirall. Sequential, multiple assignment, randomized trial designs. *JAMA*, 329(4):336–337, 2023. doi: 10.1001/jama.2022.24324.

Dong Soo Kim, Roger A. Bailey, Nino Hardt, and Greg M. Allenby. Benefit-based conjoint analysis. *Marketing Science*, 36(1):54–69, 2017. doi: 10.1287/mksc.2016.1003.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009. doi: 10.1007/s10618-008-0114-1.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981, 2019. doi: 10.1287/mnsc.2018.3093.

Ryan R. Landoll, Sara E. Vargas, Kristen B. Samardzic, Madison F. Clark, and Kate Guastaferro. The preparation phase in the multiphase optimization strategy (MOST): A systematic review and introduction of a reporting checklist. *Translational Behavioral Medicine*, 12(2):291–303, 2022. doi: 10.1093/tbm/ibab146.

Moran Lazar, Ella Miron-Spektor, Gilad Chen, Brent Goldfarb, Miriam Erez, and Rajshree Agarwal. Forming entrepreneurial teams: Mixing business and friendship to create transactive memory systems for enhanced success. *Academy of Management Journal*, 65(4):1110–1138, 2022. doi: 10.5465/amj.2020.0393.

Jonathan Lewis. Experimental design: Ethics, integrity, and the scientific method. In Ron Iphofen, editor, *Handbook of Research Ethics and Scientific Integrity*, pages 459–474. Springer, Cham, Switzerland, 2020. doi: 10.1007/978-3-030-16759-2_19.

Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A. Murphy. Sample size calculations for micro-randomized trials in mHealth. *Statistics in Medicine*, 35(12):1944–1971, 2016. doi: 10.1002/sim.6847.

John A. List. Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25(3):3–16, 2011. doi: 10.1257/jep.25.3.3.

Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018. doi: 10.1007/s12160-016-9830-8.

Timothy NeCamp, Srijan Sen, Elena Frank, Maureen A. Walton, Edward L. Ionides, Yu Fang, Ambuj Tewari, and Zhenke Wu. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: Micro-randomized trial. *Journal of Medical Internet Research*, 22(3):e15033, 2020. doi: 10.2196/15033.

Siobhan M. Phillips, Linda M. Collins, Frank J. Penedo, Kerry S. Courneya, Whitney Welch, Alison Cottrell, Gillian R. Lloyd, Kara Gavin, David Cella, Ronald T. Ackermann, Juned Siddique, and Bonnie Spring. Optimization of a technology-supported physical activity intervention for breast cancer survivors: Fit2Thrive study protocol. *Contemporary Clinical Trials*, 66:9–19, 2018. doi: 10.1016/j.cct.2018.01.001.

Trisha Phillips. Ethics of field experiments. *Annual Review of Political Science*, 24(1):277–300, 2021. doi: 10.1146/annurev-polisci-041719-101956.

Omid Rafieian, Anuj Kapoor, and Amitt Sharma. Multiobjective personalization of marketing interventions. *Marketing Science*, 44(2):457–477, 2025. doi: 10.1287/mksc.2023.0122.

Vithala R. Rao. Conjoint analysis. In Jagdish N. Sheth and Naresh K. Malhotra, editors, *Wiley International Encyclopedia of Marketing*. John Wiley & Sons, Chichester, UK, 2010. doi: 10.1002/9781444316568.wiem02019.

Amy S. Ruppert, Jun Yin, Marie Davidian, Anastasios A. Tsiatis, John C. Byrd, Jennifer A. Woyach, and Sumithra J. Mandrekar. Application of a sequential multiple assignment randomized trial (SMART) design in older patients with chronic lymphocytic leukemia. *Annals of Oncology*, 30(4):542–550, 2019. doi: 10.1093/annonc/mdz053.

Simon J. D. Schillebeeckx, Sankalp Chaturvedi, Gerard George, and Zella King. What do I want? The effects of individual aspiration and relational capability on collaboration preferences. *Strategic Management Journal*, 37(7):1493–1506, 2016. doi: 10.1002/smj.2396.

Bonnie Spring, Angela F. Pfammatter, Laura Scanlan, Elyse Daly, Jean Reading, Sam Battalio, H. Gene McFadden, Don Hedeker, Juned Siddique, and Inbal Nahum-Shani. An adaptive behavioral intervention for weight loss management: A randomized clinical trial. *JAMA*, 332(1):21–30, 2024. doi: 10.1001/jama.2024.0821.

Hal R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016. doi: 10.1073/pnas.1510479113.

Ashley Walton, Inbal Nahum-Shani, Lori Crosby, Predrag Klasnja, and Susan Murphy. Optimizing digital integrated care via micro-randomized trials. *Clinical Pharmacology & Therapeutics*, 104(1):53–58, 2018. doi: 10.1002/cpt.1079.

David L. Wyrick, Amanda E. Tanner, Jeffrey J. Milroy, Kate Guastaferro, Sandesh Bhandari, Kari C. Kugler, Shemeka Thorpe, Samuella Ware, Alicia M. Miller, and Linda M. Collins. itMatters: Optimization of an online intervention to

prevent sexually transmitted infections in college students. *Journal of American College Health*, 70(4):1212–1222, 2022. doi: 10.1080/07448481.2020.1790571.

Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, and Hui Liu. Are large language models (LLMs) good social predictors? arXiv preprint arXiv:2402.12620, 2024.

Niam Yaraghi. An empirical analysis of the financial benefits of health information exchange in emergency departments. *Journal of the American Medical Informatics Association*, 22(6):1169–1172, 2015. doi: 10.1093/jamia/ocv068.

Niam Yaraghi, Weiguang Wang, Guodong Gao, and Ritu Agarwal. How online quality ratings influence patients' choice of medical providers: Controlled experimental survey study. *Journal of Medical Internet Research*, 20(3):e99, 2018. doi: 10.2196/jmir.8986.

Niam Yaraghi, Darrell M. West, Ram D. Gopal, and Ram Ramesh. (How) did attack advertisements increase Affordable Care Act enrollments? *PLOS ONE*, 15(2):e0228185, 2020. doi: 10.1371/journal.pone.0228185.

Niam Yaraghi, Ramaswamy Ramesh, and Giri Kumar Tayi. Pay, pat, and clawback: Incentivizing service providers' participation in on-demand digital platforms. *Management Science*, 71(9):7579–7599, 2025. doi: 10.1287/mnsc.2023.02603.