Design II: Inference

Ryan Oprea

University of California, Santa Barbara

Economics 276a,
Lectures 3

What Makes a Good Design?

The first characteristic of a good design is that it is completely shaped by the goal of answering a specific question or set of questions.

What Makes a Good Design?

The first characteristic of a good design is that it is completely shaped by the goal of answering a specific question or set of questions.

The second characteristic of a good design is that it enables a maximally clear test of these questions. It should:

- Minimize noise by avoiding unnecessary complication, uncontrolled preferences and superfluous behaviors.
- Avoid confounds by clearly mapping how the design enables hypothesis tests prior to data collection.
- Avoids causal ambiguity by using exogenous treatment and parameter variation to directly test as many of the questions as possible.

What Makes a Good Design?

The first characteristic of a good design is that it is completely shaped by the goal of answering a specific question or set of questions.

The second characteristic of a good design is that it enables a maximally clear test of these questions. It should:

- Minimize noise by avoiding unnecessary complication, uncontrolled preferences and superfluous behaviors.
- Avoid confounds by clearly mapping how the design enables hypothesis tests prior to data collection.
- Avoids causal ambiguity by using exogenous treatment and parameter variation to directly test as many of the questions as possible.

Finally, a good design accomplishes all of this efficiently without wasting time and money.

Some standard terminology

Subject: An individual being observed in the experiment.

Session: A single meeting for data collection using a single set of subjects.

Period: A repetition of the decision task (or tasks) within a session (the meaning of the term differs from experiment type to experiment type).

Parameter: A (usually numerical) variable in a model that may or may not be changed in the experiment (the connotation in using the term is usually that it will be changed from treatment to treatment).

Treatment Variable: A set of conditions in the experiment that will take on more than one level over the course of the investigation. Similar in meaning to parameter.

Treatment: A collection of settings of the treatment variables.

Experiment: A collection of treatments designed to answer a question.

What Does a Design Consist Of?

A design is basically a set of choices about the following items:

1. Model Translation: A decision about the basic economic environment to be implemented in the lab including details on the elements that will not be varied in the design.

What Does a Design Consist Of?

A design is basically a set of choices about the following items:

1. Model Translation: A decision about the basic economic environment to be implemented in the lab including details on the elements that will not be varied in the design.
2. Treatments and Calibration: A choice of a set of variables you plan to vary (the treatment variables); calibration of these treatment variables if they are numeric.

What Does a Design Consist Of?

A design is basically a set of choices about the following items:

1. Model Translation: A decision about the basic economic environment to be implemented in the lab including details on the elements that will not be varied in the design.
2. Treatments and Calibration: A choice of a set of variables you plan to vary (the treatment variables); calibration of these treatment variables if they are numeric.
3. Statistical Design: A choice about how treatment variables will be applied and arrayed with respect to one another (the factorial design) and with respect to sessions and subjects.

What Does a Design Consist Of?

A design is basically a set of choices about the following items:

1. Model Translation: A decision about the basic economic environment to be implemented in the lab including details on the elements that will not be varied in the design.
2. Treatments and Calibration: A choice of a set of variables you plan to vary (the treatment variables); calibration of these treatment variables if they are numeric.
3. Statistical Design: A choice about how treatment variables will be applied and arrayed with respect to one another (the factorial design) and with respect to sessions and subjects.
4. Power Considerations and Sample Size: A choice about how many sessions to run, how many subjects will be in the session and how many periods to run.

The rest of this lecture is mostly about the considerations that go into these choices.

An Experiment is a Living Model

Building an experiment is a lot like building a model.

- List players, actions space, payoff functions, time, information, information sets etc.
- Big difference: "solution" comes from actual behavior

An Experiment is a Living Model

Building an experiment is a lot like building a model.

- List players, actions space, payoff functions, time, information, information sets etc.
- Big difference: "solution" comes from actual behavior

The model you implement in the lab will probably come from one of these places:

- You start with a model from the experimental literature.
- You start with a model from the theoretical literature.
- You start with a question and search for the right model from the theoretical literature.
- You start with a question and build the right model from scratch.

Designing a Video Game

The key to figuring out how to implement an experiment is to decide in exact detail how you would transform your model into (essentially) a video game (or a hand run game):

- What choices will subjects face?
- What feedback will they get.
- How will subjects interact?
- How will the decision problem, preferences etc. be described to subs?

Designing a Video Game

The key to figuring out how to implement an experiment is to decide in exact detail how you would transform your model into (essentially) a video game (or a hand run game):

- What choices will subjects face?
- What feedback will they get.
- How will subjects interact?
- How will the decision problem, preferences etc. be described to subs?

Might seem silly but thinking about this in every detail early on can be hugely clarifying:

- Especially important for complex experiments.
- Helps you determine feasibility of project.
- Helps make tough design decisions very early on.
- Thinking through subject experience often raises important alternative hypotheses.

Real Options Video Game

Subject choices: Subjects choose in real time when to make an investment

- Other options emerge...could have subs choose a threshold value or time at the beginning of period.
- Called strategy method, discussed in next lecture.

Real Options Video Game

Subject choices: Subjects choose in real time when to make an investment

- Other options emerge...could have subs choose a threshold value or time at the beginning of period.
- Called strategy method, discussed in next lecture.

Feedback After Decision: Definitely must show at least how much they subjects earned.

- Could end the period immediately on investment or
- could show counterfactual of where the value would have evolved to if sub had waited longer.

Real Options Video Game

Subject choices: Subjects choose in real time when to make an investment

- Other options emerge...could have subs choose a threshold value or time at the beginning of period.
- Called strategy method, discussed in next lecture.

Feedback After Decision: Definitely must show at least how much they subjects earned.

- Could end the period immediately on investment or
- could show counterfactual of where the value would have evolved to if sub had waited longer.

Subject Interaction: Single agent model so no real need for interaction. Still

- Social Learning: Could show other subjects' decisions from the same session.
- Ballinger et al. (2003) show this improves decisions a lot in other dynamic stochastic games.
- Cost of this is I lose statistical independence from subject to

Continuous Video Game

Subject choices: Subjects choose a time to enter market.

- Could do this in real time or have them give a time at beginning of game.
- Former seems more real, latter avoids nasty censoring problems.

Continuous Video Game

Subject choices: Subjects choose a time to enter market.

- Could do this in real time or have them give a time at beginning of game.
- Former seems more real, latter avoids nasty censoring problems.

Feedback: Tell subjects when counterpart exited and how much they each earned.

- Could deny them this information till the end to stamp out learning.

Continuous Video Game

Subject choices: Subjects choose a time to enter market.

- Could do this in real time or have them give a time at beginning of game.
- Former seems more real, latter avoids nasty censoring problems.

Feedback: Tell subjects when counterpart exited and how much they each earned.

- Could deny them this information till the end to stamp out learning.

Subject Interaction: Model spells out how strategy profile impacts earnings.

- Could give them tools outside the model that model says shouldn't matter (like chat communication).
- Again, could create scope for social learning by showing decisions from other pairs.

Boil it Down

Common to be shocked by complexity of your data the first time you collect it.

Boil it Down

Common to be shocked by complexity of your data the first time you collect it.

Important to take your complex model and make it as simple as possible **to answer your question**.

- Does my question require decisions over the whole model?
- Does the variation necessary to answer a question come from something that happens endogenously in the experiment?

Boil it Down

Common to be shocked by complexity of your data the first time you collect it.

Important to take your complex model and make it as simple as possible **to answer your question**.

- Does my question require decisions over the whole model?
- Does the variation necessary to answer a question come from something that happens endogenously in the experiment?

It is common to

- Cut out stages of games that don't matter
- Automate decisions you're not interested in
- Automate players you're not interested in

Boil it Down

Common to be shocked by complexity of your data the first time you collect it.

Important to take your complex model and make it as simple as possible **to answer your question**.

- Does my question require decisions over the whole model?
- Does the variation necessary to answer a question come from something that happens endogenously in the experiment?

It is common to

- Cut out stages of games that don't matter
- Automate decisions you're not interested in
- Automate players you're not interested in

Big Idea: Complex equilibria or optimization problems often have several interdependent parts. If uninteresting parts can be automated, inference over the interesting parts becomes much easier.

Classic Boiling Examples

Ultimatum game:

- Grew out of questions about willingness to punish inequitable distributions in alternating offer bargaining games.
- The repeated structure of these games makes for complicated inference.
- Guth et al. (1982) noticed the game could be boiled down to a simple 2-stage game, focusing crisply on a single central question.

Classic Boiling Examples

Ultimatum game:

- Grew out of questions about willingness to punish inequitable distributions in alternating offer bargaining games.
- The repeated structure of these games makes for complicated inference.
- Guth et al. (1982) noticed the game could be boiled down to a simple 2-stage game, focusing crisply on a single central question.

Oligopoly problems:

- Early experiments examined posted offer markets by having one set of subjects induced as firms and the other as buyers.
- But all interesting action comes from the price setters (buyers do fairly predictable things in these games).
- Common now to use "robotic" buyers.
- Standard in fact in most contemporary IO experiments (i.e. Cournot experiments).

Straying from the Model

Simplify, but be careful about how you simplify.

Straying from the Model

Simplify, but be careful about how you simplify.

Especially: translation of **continuous elements** of the model into simplifying discretizations. These can wreak havoc, destabilizing old equilibria and creating new ones:

- Continuous time
- Continuous action spaces
- Continuous populations
- Brownian motion

Straying from the Model

Simplify, but be careful about how you simplify.

Especially: translation of **continuous elements** of the model into simplifying discretizations. These can wreak havoc, destabilizing old equilibria and creating new ones:

- Continuous time
- Continuous action spaces
- Continuous populations
- Brownian motion

Always examine a complicated model via simulation.

- Program robot agents to make decisions in your exact implementation.
- Have robots make range of decisions including optimum.
- Are reward functions / best responses maximized as in the model?

Straying in Our Examples

The **Continuous Example** is actually probing one way of "straying" (using discrete approximations) so by virtue of the research questions we're being careful about straying.

Straying in Our Examples

The **Continuous Example** is actually probing one way of "straying" (using discrete approximations) so by virtue of the research questions we're being careful about straying.

The **Real Options Example** is more troublesome. Can't actually generate a true Brownian motion on a computer.

- Whenever forced to stray, let theory be a guide.
- Well known binomial approximations already in the literature that can come arbitrarily close to true Brownian motion.

Straying in Our Examples

The **Continuous Example** is actually probing one way of "straying" (using discrete approximations) so by virtue of the research questions we're being careful about straying.

The **Real Options Example** is more troublesome. Can't actually generate a true Brownian motion on a computer.

- Whenever forced to stray, let theory be a guide.
- Well known binomial approximations already in the literature that can come arbitrarily close to true Brownian motion.

Let V iterate up or down by a fixed proportion h with probability p over a tiny time interval Δt. Then we have approximate Brownian parameters α and β approximated well with small Δt

$$\alpha = \lim_{\Delta t \to 0} \frac{(2p-1)h}{\Delta t}. \quad (1)$$

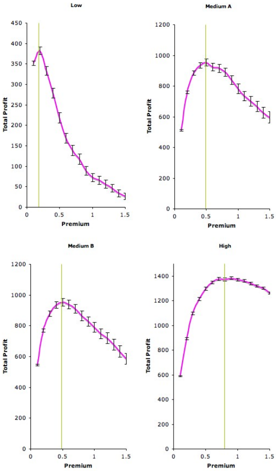
$$\sigma^2 = \lim_{\Delta t \to 0} \frac{4p(1-p)h^2}{\Delta t} \quad (2)$$

Simulating Real Options

Real Options Example: Are my approximations “close enough” to recover theoretical optima?

Simulating Real Options

Real Options Example: Are my approximations "close enough" to recover theoretical optima?



Ask Yourself These Questions

- Is there a simpler model that allows me to get at my question? Use it instead.
- Are there players in the model whose decisions have nothing to do with my question(s)? Automate them.
- Are there parts of players' decisions that fail to illuminate my problem? Automate them.
- Once I've reduced my model, is the player's decision problem trivial? Backtrack.
- In order to turn this model into an experiment, do I have to deviate from the model and do these deviations matter?
- Does the resulting model, if brought to the lab, avoid the key Inducement problems discussed in the last lecture set?

All boil down to the (slightly) **unhelpful advice**: Make the model as simple as possible but no simpler.

Choosing Treatment Variables

A treatment variable is just something you plan to vary during the experiment. What motivates treatment variables? Some of these items overlap:

- Comparative statics from a model
- Suspicious theoretical isomorphisms from theory
- Payoff irrelevant elements that seem to matter in the field but not in theory (like frames, focal points etc.)
- Other things that seem to have effects in the field but not in theory
- Variation needed for structural identification
- Need to show robustness (stay out of a narrow corner of the parameter space)

Most importantly, treatment variables flow directly from your central question(s).

Treatment Variables

Continuous Example:

- Time: Discrete vs. "Freeze Time" Continuous
- Epsilon: Vary π_M, c and number subperiods (T) to vary the costs of playing Continuous strategies in Discrete time.

Treatment Variables

Continuous Example:

- Time: Discrete vs. "Freeze Time" Continuous
- Epsilon: Vary π_M, c and number subperiods (T) to vary the costs of playing Continuous strategies in Discrete time.

Real Options Example

- Binomial parameters: p, h and q.
- Could also vary the nature of feedback to subjects.

Calibrating Parameter Values

Most models feature a lot of greek symbols that will need to be filled in before you take it to the lab. How do you choose?

- Make sure you're looking at a representative portion of the parameter space of interest.
- If this isn't possible, consider studying your experiment under a wide set of parameters, perhaps varying them over periods.
- Too much parameter variation however can interfere with inference by thinning data.
- If you're going to be varying parameters to test a comparative static predictions, make sure the predicted effect is large enough to be observable in your noisy data.
- If possible, consider studying different parameter sets that generate the same prediction (but for different reasons) to sharply test theory.
- Make sure deviations in action space represent substantial deviations in payoff space under your parameters.

Parameters in Real Options Example

Fixed Parameters: Main one to choose is Δt – picked 0.003 minutes, largely because it “looks” continuous to human eyes.

Parameters in Real Options Example

Fixed Parameters: Main one to choose is Δt – picked 0.003 minutes, largely because it “looks” continuous to human eyes.

Treatment Parameters: Vary (h, p, q) to fulfill a number of the criteria above.

- Low: $(0.0155, 0.513, 0.007)$ generates a $w = 0.18$
- Medium A: $(0.0155, 0.524, 0.003)$ generates a $w = 0.49$
- Medium B: $(0.03, 0.524, 0.007)$ also generates a $w = 0.49$
- High: $(0.03, 0.513, 0.003)$ generates a $w = 0.80$

Parameters in Real Options Example

Fixed Parameters: Main one to choose is Δt – picked 0.003 minutes, largely because it "looks" continuous to human eyes.

Treatment Parameters: Vary (h, p, q) to fulfill a number of the criteria above.

- Low: $(0.0155, 0.513, 0.007)$ generates a $w = 0.18$
- Medium A: $(0.0155, 0.524, 0.003)$ generates a $w = 0.49$
- Medium B: $(0.03, 0.524, 0.007)$ also generates a $w = 0.49$
- High: $(0.03, 0.513, 0.003)$ generates a $w = 0.80$

Get a lot of parameter coverage, get relatively steep payoff hills, get large predicted effects (Low vs. High) and two very different parameter vectors that yield same prediction.

Parameters in Continuous Example

We have to choose a parameter vector (π_R, π_M, c) and values of T to take to the lab. Simon and Stinchombe give us a nice starting point of $(3, 2, 1)$ but has really flat payoffs.

Parameters in Continuous Example

We have to choose a parameter vector (π_R, π_M, c) and values of T to take to the lab. Simon and Stinchombe give us a nice starting point of $(3, 2, 1)$ but has really flat payoffs.

Fixed Parameters: π_R doesn't impact steepness so we'll probably fix it for the whole design. Will probably also fix c (but probably not at 1).

Parameters in Continuous Example

We have to choose a parameter vector (π_R, π_M, c) and values of T to take to the lab. Simon and Stinchombe give us a nice starting point of $(3, 2, 1)$ but has really flat payoffs.

Fixed Parameters: π_R doesn't impact steepness so we'll probably fix it for the whole design. Will probably also fix c (but probably not at 1).

Treatment Parameters: Probably will vary the vector (π_M, T) in the design.

- Note for any $T <$ inf discrete Nash predictions are invariant to (π_M, T)
- but ϵ threshold needed to play continuous strategy instead varies in both parameters.
- **Idea:** Fix two ϵ-threshold values, ϵ_L and ϵ_H.
- Choose several combinations (π_M, T) that generate these ϵ_i in different ways.
- Test (i) whether ϵ matters and (ii) whether ϵ is all that matters.

Factorial Design

The cardinal rule of making inferences between treatments is to change only one thing at a time.

Factorial Design

The cardinal rule of making inferences between treatments is to change only one thing at a time.

Factorial designs are classic inference tools in experiments and they have a central place in experimental economics.

Factorial Design

The cardinal rule of making inferences between treatments is to change only one thing at a time.

Factorial designs are classic inference tools in experiments and they have a central place in experimental economics.

Idea: observe subjects under a set of interactions between treatment variables

- **Factor:** A set of variables that will be varied in the experiment.
- **Level:** A setting for a variable.
- **Cell:** A combination of level settings across all factors.

Full Factorial Designs

A full factorial design consists of the full combination of factors and levels.

- The convention is to describe the size of the design as a {factor} x {level} or $k \times n$.
- This will lead to n^k cells of the design (a cell in this context is a treatment).
- The 2x2 experimental design is very common in experimental economics.

Full Factorial Designs

A full factorial design consists of the full combination of factors and levels.

- The convention is to describe the size of the design as a {factor} x {level} or $k \times n$.
- This will lead to n^k cells of the design (a cell in this context is a treatment).
- The 2x2 experimental design is very common in experimental economics.

Factorial designs can quickly blow up. If you have more than two treatment variables it is often necessary to avoid the full factorial.

Full Factorial Designs

A full factorial design consists of the full combination of factors and levels.

- The convention is to describe the size of the design as a {factor} x {level} or $k \times n$.
- This will lead to n^k cells of the design (a cell in this context is a treatment).
- The 2x2 experimental design is very common in experimental economics.

Factorial designs can quickly blow up. If you have more than two treatment variables it is often necessary to avoid the full factorial.

The most important thing is to observe level changes with all else held constant on the treatments you are most interested in.

Factorial Designs and Main Effects

Factorial Designs let you make two types of inferences:

Factorial Designs and Main Effects

Factorial Designs let you make two types of inferences:

- **Main effects:** The average effect of changing a level of a factor averaged across all the levels of the other factors.

Factorial Designs and Main Effects

Factorial Designs let you make two types of inferences:

- **Main effects:** The average effect of changing a level of a factor averaged across all the levels of the other factors.
- **Interaction effects:** Effect of a second factor on the effect of changing levels on a first factor (2nd order interaction effect)

Factorial Designs and Main Effects

Factorial Designs let you make two types of inferences:

- **Main effects:** The average effect of changing a level of a factor averaged across all the levels of the other factors.
- **Interaction effects:** Effect of a second factor on the effect of changing levels on a first factor (2nd order interaction effect)

Add factors, can consider higher order interaction effects: how does the interaction of a third factor impact the interaction effect of a second factor on the level changes of a first factor (try saying that 10 times fast).

Factorial Designs and Main Effects

Factorial Designs let you make two types of inferences:

- **Main effects:** The average effect of changing a level of a factor averaged across all the levels of the other factors.
- **Interaction effects:** Effect of a second factor on the effect of changing levels on a first factor (2nd order interaction effect)

Add factors, can consider higher order interaction effects: how does the interaction of a third factor impact the interaction effect of a second factor on the level changes of a first factor (try saying that 10 times fast).

The so-called **sparcity-of-effect principle** suggests that only main effects and lower order interaction effects are likely to be important.

Factorial Design in Continuous Example

A classic 2x2 design for our Continuous Example

Factorial Design in Continuous Example

A classic 2x2 design for our Continuous Example

	Low- ϵ	**High-** ϵ
Discrete	L-D	H-D
Continuous	L-C	H-C

Where $\epsilon = f(\pi_M, c, \overline{T})$

Factorial Design in Continuous Example

A classic 2x2 design for our Continuous Example

	Low- ϵ	**High-** ϵ
Discrete	L-D	H-D
Continuous	L-C	H-C

Where $\epsilon = f(\pi_M, c, \overline{T})$

In addition to point predictions of model, can test following hypotheses using this data to get at main experimental questions:

- Continuous is different from Discrete: $\frac{t^{L-D}+t^{H-D}}{2} < \frac{t^{L-C}+t^{H-C}}{2}$
- ϵ influences wedge between Continuous and Discrete: $|t^{L-D} - t^{H-D}| > |t^{L-C} - t^{H-C}|$
- ϵ increases wedge between Continuous and Discrete: $t^{L-C} - t^{L-D} < t^{H-C} - t^{H-D}$

Which are main effects and which are interaction effects?

Fractional Factorial Designs

The sparcity-of-effect principle suggests an alternative to the Full Factorial Design that allows one to recover only a subset of the total set of possible interactions.

Fractional Factorial Designs

The sparcity-of-effect principle suggests an alternative to the Full Factorial Design that allows one to recover only a subset of the total set of possible interactions.

This is a big topic, but an example can illustrate. Suppose you are interested in 3 factors, each at 2 levels. A full fractional design requires $2^3 = 8$ cells. You can run a $2^{3-1} = 4$ cell design instead and if done properly can recover the main effects (but none of the interactions). To avoid confounds of main effects:

- Run every level in exactly two cells.
- Make sure no two cells have two factors sharing the same levels.

Fractional Factorial Designs

The sparcity-of-effect principle suggests an alternative to the Full Factorial Design that allows one to recover only a subset of the total set of possible interactions.

This is a big topic, but an example can illustrate. Suppose you are interested in 3 factors, each at 2 levels. A full fractional design requires $2^3 = 8$ cells. You can run a $2^{3-1} = 4$ cell design instead and if done properly can recover the main effects (but none of the interactions). To avoid confounds of main effects:

- Run every level in exactly two cells.
- Make sure no two cells have two factors sharing the same levels.

Doing this will confound main effects with interaction effects so it only works if interactions are thought to be unimportant.

Half Factorial Design in Real Options Example

Real Options example uses a half-factorial design.

Treatment τ	Step Size h	Uptick Prob p	Expiration Prob q	Option Prem w^*_τ
Low	0.0155	0.513	**0.007**	0.179
Medium A	0.0155	**0.524**	0.003	0.490
Medium B	**0.03**	**0.524**	**0.007**	0.499
High	**0.03**	0.513	0.003	0.804
Effect on Premium	0.317	0.003	-0.308	

Main Effects:

$$\bar{h} = \frac{w^*_{High} + w^*_{MedB}}{2} - \frac{w^*_{Low} + w^*_{MedA}}{2} = 0.317 \quad (3)$$

$$\bar{p} = \frac{w^*_{MedA} + w^*_{MedB}}{2} - \frac{w^*_{Low} + w^*_{High}}{2} = 0.003 \quad (4)$$

$$\bar{q} = \frac{w^*_{Low} + w^*_{MedB}}{2} - \frac{w^*_{High} + w^*_{MedA}}{2} = -0.308 \quad (5)$$

Between and Within Design

An important design question is how to assign treatments to subjects:

Between and Within Design

An important design question is how to assign treatments to subjects:

- Between subject design: Subjects are observed under one level of the factor only.
- Within subject design: The subjects are observed under multiple levels of the factor.

Between and Within Design

An important design question is how to assign treatments to subjects:

- Between subject design: Subjects are observed under one level of the factor only.
- Within subject design: The subjects are observed under multiple levels of the factor.

In many experiments, some of the factors/treatment variables are applied between subject while others are within subject.

Within Subject Design

A within subject design has a few advantages and disadvantages.

Pro:

- Can substantially reduce variability (creates tighter standard errors).
- Essentially, individual fixed effects difference out.
- Less formally, treatment effects found within-subject are considered especially powerful.

Within Subject Design

A within subject design has a few advantages and disadvantages.

Pro:

- Can substantially reduce variability (creates tighter standard errors).
- Essentially, individual fixed effects difference out.
- Less formally, treatment effects found within-subject are considered especially powerful.

Con:

- Behavior under one treatment may impact decisions under the other.
- Because treatments must be applied in sequence, possible to confound learning with treatment effects.
- (Can be partially eased by varying the order of treatment across subjects or following an ABABAB... design)

Blocked Designs

A within subject design is an example of a statistical principle called "blocking".

The idea is, if you can find "nuisance factors" (individual subjects or sessions bring many of these with them) that are sources of variability, you can run all of your treatments under each nuisance factor to increase precision (by reducing the degree to which variance in nuisance factors drives variance in measured treatment difference).

Blocked Designs

A within subject design is an example of a statistical principle called "blocking".

The idea is, if you can find "nuisance factors" (individual subjects or sessions bring many of these with them) that are sources of variability, you can run all of your treatments under each nuisance factor to increase precision (by reducing the degree to which variance in nuisance factors drives variance in measured treatment difference).

- Difference on observations within block and average to measure treatment effect (similar to diff-in-diff).

Blocked Designs

A within subject design is an example of a statistical principle called "blocking".

The idea is, if you can find "nuisance factors" (individual subjects or sessions bring many of these with them) that are sources of variability, you can run all of your treatments under each nuisance factor to increase precision (by reducing the degree to which variance in nuisance factors drives variance in measured treatment difference).

- Difference on observations within block and average to measure treatment effect (similar to diff-in-diff).
- Note $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$. We increase covariance between X and Y by having them measured on the same subject, thus reducing variance of the treatment comparison.

Blocked Designs

A within subject design is an example of a statistical principle called "blocking".

The idea is, if you can find "nuisance factors" (individual subjects or sessions bring many of these with them) that are sources of variability, you can run all of your treatments under each nuisance factor to increase precision (by reducing the degree to which variance in nuisance factors drives variance in measured treatment difference).

- Difference on observations within block and average to measure treatment effect (similar to diff-in-diff).
- Note $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$. We increase covariance between X and Y by having them measured on the same subject, thus reducing variance of the treatment comparison.

Some common blocking variables

Blocked Designs

A within subject design is an example of a statistical principle called "blocking".

The idea is, if you can find "nuisance factors" (individual subjects or sessions bring many of these with them) that are sources of variability, you can run all of your treatments under each nuisance factor to increase precision (by reducing the degree to which variance in nuisance factors drives variance in measured treatment difference).

- Difference on observations within block and average to measure treatment effect (similar to diff-in-diff).
- Note $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$. We increase covariance between X and Y by having them measured on the same subject, thus reducing variance of the treatment comparison.

Some common blocking variables

- Sessions/subjects
- Lab/experimenter
- Demography (i.e. gender, culture)
- Time/Order

Between and Within in Our Examples

In **Real Options example**:

Between and Within in Our Examples

In **Real Options example**:

- I need a lot of variation in stochastic realizations to identify behavioral rules.
- I really want to get a feel for long term learning.

Both suggest I should avoid a within-subject design.

Between and Within in Our Examples

In **Real Options example**:

- I need a lot of variation in stochastic realizations to identify behavioral rules.
- I really want to get a feel for long term learning.

Both suggest I should avoid a within-subject design.

In **Continuous example**:

Between and Within in Our Examples

In **Real Options example**:

- I need a lot of variation in stochastic realizations to identify behavioral rules.
- I really want to get a feel for long term learning.

Both suggest I should avoid a within-subject design.

In **Continuous example**:

- I want ample opportunity to get a feel for joint payoff maximizer (continuous prediction).
- ϵ doesn't effect this if I design carefully.

Suggests I *could* vary either of these within session but I would want to block carefully with time to avoid confounding time with treatment:

- If Continuous/Discrete is applied between-subject run in each: High-ϵ / Low-ϵ / High-ϵ / Low-ϵ / ...
- else a fully within-design: L-D / H-D /H-C / L-C/ L-D / H-D/...

Repetition Within Session

Will subjects will be repeating the experimental task? Most experiments feature repetition and for three possible reasons:

Repetition Within Session

Will subjects will be repeating the experimental task? Most experiments feature repetition and for three possible reasons:

- Improve salience: Subjects often get a feel for the game based on a bit of experimentation.
- Study the learning process: In many studies how subjects learn is the center of the investigation.
- Enable within-session treatment variation: You can't achieve within-subject variation without having subjects make more than one decision!

Three Ways to Repeat

There are three basic ways to do repetition.

Three Ways to Repeat

There are three basic ways to do repetition.

- **Stationary repetition:** Subjects do the exact same thing in the exact same environment repeatedly.

Three Ways to Repeat

There are three basic ways to do repetition.

- **Stationary repetition:** Subjects do the exact same thing in the exact same environment repeatedly.
- **Within-session treatment application:** As above but elements of the design are changed from period-to-period

Three Ways to Repeat

There are three basic ways to do repetition.

- **Stationary repetition:** Subjects do the exact same thing in the exact same environment repeatedly.
- **Within-session treatment application:** As above but elements of the design are changed from period-to-period
- **Time-blocking:** Treatments are varied within a block of periods and these blocks are repeated in order to avoid confounding experience with treatment.

Matching

Every experiment with interaction between subjects repeated over time must settle on a way to match the subjects within the session. Two most common:

Matching

Every experiment with interaction between subjects repeated over time must settle on a way to match the subjects within the session. Two most common:

- **Partner matching:** Experiments explicitly studying supergames and reputations match subjects with the same partner(s) repeatedly.
- **Strangers matching:** Experiments aimed at gathering information on one shot games often randomly and anonymously rematch subjects with one another from period to period to short circuit supergame/reputation effects.
    - "Zipper" designs are matching protocols that attempt to prevent subjects from being rematched with each other at all during the experiment.
    - Healy (2007) offers a useful caution: group reputations are possible even in such settings.

Matching

Other increasingly popular matching strategies:

- **Repeated-repeated:** Since Selten and Stoecker (1986, at least) and increasingly recently experimentalists have studied repeated supergames. Subjects are matched with new partners in each period but periods are divided into a number of "subperiods" during which subjects stay matched with the same partner.
- **Population matching:** In some experiments the entire session's worth of subjects must interact simultaneously in each period (i.e. in most market experiments, population games).
- **Silos:** Subjects within session are divided into subgroups and matched only with other members of the sub-group (to generate more fully independent data).

How to Match in Continuous Example

How to Match in Continuous Example

Definitely want to do repeated-repeated to observe learning about timing.

How to Match in Continuous Example

Definitely want to do repeated-repeated to observe learning about timing.

Might group subjects into two 6-subject matching silos.

- **Pro:** Get more fully independent data.
- **Con:** If I run 30 periods, each sub matched with each other subject 6 times. Should I worry about group reputations?

How to Match in Continuous Example

Definitely want to do repeated-repeated to observe learning about timing.

Might group subjects into two 6-subject matching silos.

- **Pro:** Get more fully independent data.
- **Con:** If I run 30 periods, each sub matched with each other subject 6 times. Should I worry about group reputations?

Obviously no matching decisions necessary for real options example

How Many Subjects?

A lot of things to consider when deciding this:

- Resources
- Time
- Level of independence across subjects (more on this later)
- Statistical power

How Many Subjects?

A lot of things to consider when deciding this:

- Resources
- Time
- Level of independence across subjects (more on this later)
- Statistical power

Statistical Power: The probability a statistical test rejects a false null hypothesis.

- If probability of a Type I error is β then power is $1 - \beta$
- A low powered experiment is ill suited to test a hypothesis.
- With very low power, it is very likely you will fail to find effects that are meaningful and real.

Components of Power Analysis

Power analysis is a way of ensuring you are giving the rejection of the null a "fair shot." Likelihood of rejecting a null depends on:

Components of Power Analysis

Power analysis is a way of ensuring you are giving the rejection of the null a "fair shot." Likelihood of rejecting a null depends on:

- Expected effect size (Δ)
- Variance (σ^2)
- n (number of subjects)
- α (probability of type I error you can tolerate, conventionally 0.05)

Components of Power Analysis

Power analysis is a way of ensuring you are giving the rejection of the null a "fair shot." Likelihood of rejecting a null depends on:

- Expected effect size (Δ)
- Variance (σ^2)
- n (number of subjects)
- α (probability of type I error you can tolerate, conventionally 0.05)

Under assumption of approximate normality, can calculate power $(1 - \beta)$ using standard methods (or using any statistical software package).

Two Ways to Apply

A priori power analysis: figure out how many observations you need.

- Get Δ from a theoretical prediction if possible (often not).
- Get σ from pilot data or previous studies.
- Convention: Set an n generating a power of 0.80

Two Ways to Apply

A priori power analysis: figure out how many observations you need.

- Get Δ from a theoretical prediction if possible (often not).
- Get σ from pilot data or previous studies.
- Convention: Set an n generating a power of 0.80

Post hoc power analysis: figure out how confident you should be in your results

- Get Δ from means differences in data.
- Get σ from your data.
- Calculate power (and credibility of your results).

Getting It Right

Fixing $1 - \beta$, you can reduce required N with careful design:

- Use a Within-Design and study across-treatment differences ($\downarrow \sigma$)
- Block on as many factors as possible ($\downarrow \sigma$)
- Include controls in statistical tests ($\downarrow \sigma$)
- Induce values carefully – minimize confusion, use steep payoffs, eliminate sources confounding dominance ($\downarrow \sigma$)
- Choose parameters that generate widely separating predictions ($\downarrow \Delta$)

These only scratch the surface.

Getting It Right

Fixing $1 - \beta$, you can reduce required N with careful design:

- Use a Within-Design and study across-treatment differences ($\downarrow \sigma$)
- Block on as many factors as possible ($\downarrow \sigma$)
- Include controls in statistical tests ($\downarrow \sigma$)
- Induce values carefully – minimize confusion, use steep payoffs, eliminate sources confounding dominance ($\downarrow \sigma$)
- Choose parameters that generate widely separating predictions ($\downarrow \Delta$)

These only scratch the surface.

Failures of independence complicate all of this significantly

- Multiple observations on the same subject are **not** independent sources of data.
- Subject interactions reduce and sometimes entirely destroy independence.
- Means you may need more data.

With Your Econometrics In Mind

Someone once told me experimental design is a substitute for (a lot of) econometrics.

With Your Econometrics In Mind

Someone once told me experimental design is a substitute for (a lot of) econometrics.

In order to function this way your design must

- exogenize your key hypothesis tests
- avoid confounds

With Your Econometrics In Mind

Someone once told me experimental design is a substitute for (a lot of) econometrics.

In order to function this way your design must

- exogenize your key hypothesis tests
- avoid confounds

This means your hypothesis tests should as much as possible be set up to concern the impacts of variation that is driven directly by treatments.

With Your Econometrics In Mind

Someone once told me experimental design is a substitute for (a lot of) econometrics.

In order to function this way your design must

- exogenize your key hypothesis tests
- avoid confounds

This means your hypothesis tests should as much as possible be set up to concern the impacts of variation that is driven directly by treatments.

This in turn means your design should be largely driven by the **specific** hypothesis tests you need to run to answer your questions. Thinking this way can have a huge impact on your design.

With Your Econometrics In Mind

Someone once told me experimental design is a substitute for (a lot of) econometrics.

In order to function this way your design must

- exogenize your key hypothesis tests
- avoid confounds

This means your hypothesis tests should as much as possible be set up to concern the impacts of variation that is driven directly by treatments.

This in turn means your design should be largely driven by the **specific** hypothesis tests you need to run to answer your questions. Thinking this way can have a huge impact on your design. Someone else once told me they do all of their econometrics before they run the experiment. This is great advice.

With Your Econometrics In Mind

Common to kick yourself shortly after data collection when you find exact questions are only answered using endogenous cuts of the data or are confounded in subtle ways. Someone else once told me they do all of their econometrics before they run the experiment. This is great advice.

With Your Econometrics In Mind

Common to kick yourself shortly after data collection when you find exact questions are only answered using endogenous cuts of the data or are confounded in subtle ways. Someone else once told me they do all of their econometrics before they run the experiment. This is great advice.

A really useful exercise (again) is to simulate your full dataset

1. Add noise to subject decisions and add random effects by subject.
2. Write code for your full set of econometrics and conduct a complete analysis.
3. Vary whether these robotic subjects' decisions are centered on your null or whether their behavior would reject the null.

With Your Econometrics In Mind

Common to kick yourself shortly after data collection when you find exact questions are only answered using endogenous cuts of the data or are confounded in subtle ways. Someone else once told me they do all of their econometrics before they run the experiment. This is great advice.

A really useful exercise (again) is to simulate your full dataset

1. Add noise to subject decisions and add random effects by subject.
2. Write code for your full set of econometrics and conduct a complete analysis.
3. Vary whether these robotic subjects' decisions are centered on your null or whether their behavior would reject the null.

Doing this does a few things:

1. It makes your future data concrete and will surprisingly often reveal problems with your design.
2. It will reveal subtle barriers to parameter recovery in structural exercises.
3. Overall, it forces you to think about your design in terms of your future data analysis.

Design With Your Paper In Mind

Think about what sort of papers might come out of your experiment as a function of the results.

Design With Your Paper In Mind

Think about what sort of papers might come out of your experiment as a function of the results.

The "safest" type of experiment (especially for a grad student) is one that will make an interesting paper regardless of the results.

Design With Your Paper In Mind

Think about what sort of papers might come out of your experiment as a function of the results.

The "safest" type of experiment (especially for a grad student) is one that will make an interesting paper regardless of the results.

Anticipate criticisms. What would a skeptical reader think of your argument? Build your design with a critical reader in mind.

- Are there objectionable elements in the way you've induced preferences?
- In the best case scenario, are there remaining ambiguities in the interpretation of data under your design?
- What follow-up treatments might you run in various contingencies to convincingly explain your data? Be as prepared as possible from the beginning as this may inform your initial design.

Bibliography

- Ballinger, T., M. Palumbo and N. Wilcox. (2002) "Precautionary savings and social learning across generations: An experiment." Economic Journal 113: 920-947.
- Healy, P. J., (2007) "Group Reputations, Stereotypes, and Cooperation in a Repeated Labor Market" American Economic Review, 97(5): 17511773.
- Selten, R. and R. Stoecker (1986) "End behavior in sequences of finite prisoner's dilemma supergames: a learning theory approach," Journal of Economic Behavior and Organization, 7: 47-70.
- Guth, W., R. Schmittberger and B. Schwarze (1982) "An experimental analysis of ultimatum bargaining," Journal of Economic Behavior and Organization, 3(4): 367-388.