Better preanalysis plans through design declaration & diagnosis

Graeme Blair, UCLA

Joint work with Jasper Cooper, Alexander Coppock, and Macartan Humphreys
The Journal of Politics is phasing in a requirement for all experimental research to be preregistered. I believe that this is just the second journal across the social-behavioral sciences to require preregistration.

**Pre-registration**: authors who want to submit manuscripts containing original experimental work, including laboratory, field, and survey experiments are required to submit proof of study/design pre-registration with one of the available research registries (e.g., EGAP, RCT, Open Science). Pre-registration of other types of research design is very much encouraged. The submission of unregistered laboratory, field, and survey experiments will not be accepted. This policy will be phased in: For manuscripts submitted in 2021, authors need to justify in a letter to the editor why the study was not or could not be preregistered.

Dear Colleagues: before submitting to @The_JOP please have a look at the new guidelines for contributors, including pre-registration, replication, ethical considerations etc.: journals.uchicago.edu/journals/jop/i... @VETroeger
What should go in a plan?

1. Research design declaration

   Model
   Inquiry
   Data strategy
   Answer strategy

2. Research design diagnosis
What is a preanalysis plan?

What are your hypotheses?
How will you test them?

Timestamped publicly
Why preanalysis plans?

Clarify what you thought *before*, *(in the middle)*, and *after*

What tests are confirmatory vs. exploratory
Writing plans changes plans
1) Data collection. Have any data been collected for this study already?

- Yes, I have already collected the data.
- No, I have not yet collected any data.
- It's complicated. I have already collected some data but explain in Question 3 why readers may consider this a valid pre-registration nevertheless.

(Note: "Yes" is not an accepted answer.)

What goes in a plan?

As Predicted: 9 items

McKenzie (2012): 10 items

AEA registry: ~ 30 items

EGAP registry: ~ 30 items

Journal of Development Economics: 44 items

Ganimian (2018): 60 items

2) What is the dependent variable or hypothesis being tested in this study?

Example: A month-long academic summer program for disadvantaged kids will reduce the drop in academic performance that occurs during the

3) Dependent variable. Describe the key dependent variable(s) specifying how they will be measured.

Example: Simple average GPA across all courses during the first semester after the intervention.

4) Conditions. How many conditions do participants have to choose from?

Example 1: Two conditions: Offering summer program: yes vs no.

Example 2: 12 conditions in a mixed design lab study. Participants will be assigned to one of four conditions: math training, verbal training, memory task, or control (4 between-subject conditions). Each participant will complete a math test, a verbal test, and a memory test (3 within-subject conditions).
We need

1. Language for research designs
2. Algorithm for choosing one
Graphical models

Covariate
Used for blocking

X

Unknown heterogeneity

U

Random assignment

Z

Outcome

Y
Structural models

\[ Y = 0.1 \times Z + X + U \]

\[ X \sim \text{binom}(1, 0.3) \]

\[ U \sim \mu(0, 1) \]
Design declaration

Model

Inquiry

Data strategy

Answer strategy
Model

Theory of the system under study

— What causes what — and how
— How variables are distributed
— Correlations between variables
— Sequence of events
— Theory if we are right, and if we are wrong
Inquiry

Features of the model you want to study

— Units
— Conditions
— Outcomes

— Descriptive, causal, predictive
Data strategy

Procedures to gather information about the world

- Sampling
- Random assignment
- Measurement
Answer strategy
How you summarize data from the data strategy

— Data cleaning
— Data transformation
— Estimation
— Visualization
— Interpretation
Answer strategy
How you summarize data from the data strategy

— Data cleaning
— Data transformation
— Estimation
— Visualization
— Interpretation
— Document selection
— Coding procedures
— Narrative
Declaring a design in practice

R package DeclareDesign
design <-

# Model
declare_model(
  N = 100,
  X = rbinom(N, 1, 0.3),
  U = rnorm(N),
  potential_outcomes(Y ~ 0.1 * Z + X + U)
) +

# Inquiry
declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0)) +

# Data strategy
declare_assignment(blocks = X, block_prob = c(0.1, 0.5)) +
declare_measurement(Y = reveal_outcomes(Y ~ Z)) +

# Answer strategy
declare_estimator(Y ~ Z, model = lm, inquiry = "ATE")
Draw simulated data

draw_data(design)

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>U</th>
<th>Y_Z_0</th>
<th>Y_Z_1</th>
<th>Z</th>
<th>Z_prob</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>1</td>
<td>0.66</td>
<td>1.66</td>
<td>1.76</td>
<td>0</td>
<td>0.5</td>
<td>1.66</td>
</tr>
<tr>
<td>002</td>
<td>1</td>
<td>-1.69</td>
<td>-0.69</td>
<td>-0.59</td>
<td>1</td>
<td>0.5</td>
<td>-0.59</td>
</tr>
<tr>
<td>003</td>
<td>0</td>
<td>-1.03</td>
<td>-1.03</td>
<td>-0.93</td>
<td>0</td>
<td>0.9</td>
<td>-1.03</td>
</tr>
<tr>
<td>004</td>
<td>1</td>
<td>-0.62</td>
<td>0.38</td>
<td>0.48</td>
<td>0</td>
<td>0.5</td>
<td>0.38</td>
</tr>
<tr>
<td>005</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.13</td>
<td>0</td>
<td>0.9</td>
<td>0.03</td>
</tr>
<tr>
<td>006</td>
<td>1</td>
<td>0.34</td>
<td>1.34</td>
<td>1.44</td>
<td>0</td>
<td>0.5</td>
<td>1.34</td>
</tr>
</tbody>
</table>
## Draw mock estimates

draw_estimates(design)

<table>
<thead>
<tr>
<th>est</th>
<th>std.err</th>
<th>statistic</th>
<th>p.value</th>
<th>conf.lo</th>
<th>conf.hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.64</td>
<td>0.29</td>
<td>2.22</td>
<td>0.03</td>
<td>0.07</td>
<td>1.22</td>
</tr>
</tbody>
</table>
Draw mock estimand

draw_estimands(design)

<table>
<thead>
<tr>
<th>estimand_label</th>
<th>estimand</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Descriptive inquiries

# survey targeting average policy preferences
declare_inquiry(mean_preferences = mean(Y))

# list experiment studying binary trait
declare_inquiry(proportion = mean(Y_star))
Observational research designs

# regression discontinuity
cutoff <- 0.5
control <- function(X) {
  as.vector(poly(X, 4, raw = TRUE) %*% c(.7, -.8, .5, 1))
} 
treatment <- function(X) {
  as.vector(poly(X, 4, raw = TRUE) %*% c(0, -1.5, .5, .8)) + .15
}

declare_model(
  N = 1000,
  U = rnorm(N, 0, 0.1),
  X = runif(N, 0, 1) + U - cutoff,
  potential_outcomes(Y ~ Z * treatment(Z) + (1 - Z) * control(X) + U),
  Z = 1 * (X > 0)
) +
declare_measurement(Y = reveal_outcomes(Y ~ Z))
Algorithm for selecting designs

Declare

Diagnose

Redesign
Algorithm for selecting designs
Diagnosing a design

What are the properties of a research design?

1. Through analytical expressions

2. Through simulation
Is my design powered?

\[
\text{power} \approx \Phi \left( \frac{|\mu_t - \mu_c| \sqrt{N}}{2\sigma} - \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right)
\]
Is my design powered?

\[
\text{power} \approx \Phi \left( \frac{|\mu_t - \mu_c| \sqrt{N}}{2\sigma} - \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right)
\]

- **Model**: normally-distributed outcome; \( \sigma_t = \sigma_c \)
- **Data strategy**: simple random assignment
- **Answer strategy**: equal-variance t-test with N-2 degrees of freedom
Is my design biased?

— Blocking with varying assignment probabilities
— Random assignment of clusters of different sizes
— Differential attrition
— Logit with fixed effects
— Posttreatment bias
How many people should I interview?
How many men and women?
How often should I interview them?
Should I assign 2 or 3 treatment arms?
Is it important in this case to use blocking?
How many items should I include in my index?
More survey items or more respondents?
Robust or cluster-robust standard errors?
Should I control for emotions in my regression?
Is it okay to drop people who didn't respond?
To where can I generalize these results?
Diagnosing a design through simulation
Simulated results from the study

```
run_design(design)
```

<table>
<thead>
<tr>
<th>est</th>
<th>std.err</th>
<th>statistic</th>
<th>p.value</th>
<th>conf.lo</th>
<th>conf.hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.64</td>
<td>0.29</td>
<td>2.22</td>
<td>0.03</td>
<td>0.07</td>
<td>1.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>estimand_label</th>
<th>estimand</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Estimand

Estimate

\( x = 0 \)

Confidence interval

\( \Delta \)

Estimand
Error
<table>
<thead>
<tr>
<th>sim_ID</th>
<th>estimate</th>
<th>std.error</th>
<th>conf.low</th>
<th>conf.high</th>
<th>p.value</th>
<th>estimand</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.08</td>
<td>0.04</td>
<td>0.00</td>
<td>0.16</td>
<td>0.04</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.05</td>
<td>0.11</td>
<td>0.45</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>0.04</td>
<td>0.04</td>
<td>0.20</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>0.04</td>
<td>0.02</td>
<td>0.18</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>0.07</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.15</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.04</td>
<td>-0.03</td>
<td>0.14</td>
<td>0.21</td>
<td>0.09</td>
</tr>
<tr>
<td>7</td>
<td>0.09</td>
<td>0.04</td>
<td>0.01</td>
<td>0.17</td>
<td>0.03</td>
<td>0.11</td>
</tr>
<tr>
<td>8</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.05</td>
<td>0.12</td>
<td>0.44</td>
<td>0.12</td>
</tr>
<tr>
<td>9</td>
<td>0.16</td>
<td>0.04</td>
<td>0.09</td>
<td>0.24</td>
<td>0.00</td>
<td>0.09</td>
</tr>
<tr>
<td>10</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.04</td>
<td>0.13</td>
<td>0.33</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Diagnosands

What are your objectives and does your design meet them?

Ethics: min(subjects_harmed)

Cost: mean(cost), max(cost)

Bias: mean(estimate - estimand)

Power: mean(p.value <= 0.05)

Probability of getting sign wrong: mean(sign(estimate) != sign(estimand))
## Diagnose design

```python
diagnose_design(design)
```

<table>
<thead>
<tr>
<th>Bias</th>
<th>RMSE</th>
<th>Power</th>
<th>Pr(sign wrong)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.56</td>
<td>0.65</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Algorithm for selecting designs

Declare

Diagnose

Redesign
Back to preanalysis plans

Declare your design in MIDA

Present a diagnosis

Register it for a timestamp
Research lifecycle

Brainstorming
Planning
Realization
Integration
Planning

Ethics
Partners
Funding
Piloting
Criticism
Preanalysis plan

Declare
Diagnose
Redesign
We're caught between higher research standards and lack of ideas for how to assess and communicate about designs.
Take aways

Minimum:
Write a plan to change your plan

Medium: Register your plan

Maximum:
Declare in code, diagnose
Thank you

More at declaredesign.org

Big team effort: Clara Bicalho, Jasper Cooper, Neal Fultz, Sisi Huang, Markus Konrad, Lily Medina, Aaron Rudkin, Luke Sonnet, and John Ternovski

Supported by EGAP and the Arnold Foundation