Behavioral field experiments

Angelino Viceisza [t: @aviceisza]

Spelman & Duke

Göttingen
November 24, 26, 27, 2015

Outline

Lecture 1: General (T, 11/24)
Introduction
Definitions
Methodology
Puritan issues
Lecture 2: LFEs (R, 11/26)
LFEs: Basics
LFEs: IV
LFEs: EV
LFEs: miscellaneous
Lecture 3: RCTs (F, 11/27)
RCTs: Basics
RCTs: IV
RCTs: EV
Lecture 4: Ways forward (F, 11/27)
Ways forward
Publication strategies & stories

Lecture 1: General (T, 11/24)

Syllabus

Review syllabus & assessment guidelines.

Exercise

1. What is an experiment?
2. What is the main issue you should consider when designing an experiment?

Exercise (cont'd)

What is the main issue you should consider when designing an experiment?

Q: What is the question you want the experiment to answer?

Notice that this is no different from any research project.

Example

A designs an experiment around a baseline two-person trust game and a modified three-person/third-party trust game.

Q: Is this a well-designed experiment?

Subfields & methodology

1. Decision theory, (traditional) game theory, behavioral game theory
2. Experimental economics, behavioral economics
3. Structural econometrics, reduced-form econometrics

Taxonomy & terminology

Harrison and List (2004)

1. Conventional lab experiments.
2. Artefactual field experiments (AFEs).
3. Framed field experiments (FFEs).
4. Natural field experiments (NFEs).
5. Natural experiments.
6. Quasi-experimental/synth. control: matching, RD, instrumental variable, DiD.

Viceisza (2015): Lablike field experiments (LFEs).

Development à la Duflo: Randomized controlled trials (RCTs) aka NFEs or FFEs & Lab-in-the-field.

Experimental à la Gneezy: Extralab experiments for LFEs aka AFEs, FFEs.

Elicitation

Given your research question, is your experiment eliciting what you think it is/should?

Example: Charness and Viceisza (CV; *RoBE forthcoming*, 2015), `https://goo.gl/7gfC3d`.

CV (2015), 1

Small-scale/pilot study ($N = 91$)

1. Senegalese farmers.
2. Exposed to Holt-Laury and Gneezy-Potters (between-subjects).
3. All subjects exposed to WTR scale à la Dohmen et al.
4. Punchline:
   - 4.1 HL high inconsistency, but some predictive ability.
   - 4.2 WTR some predictive ability, but women are more likely than men to report highest WTR ($= 10$).
   - 4.3 GP no apparent failure, but limited predictive ability.

CV (2015), 2

One can imagine that such issues extend to other forms of elicitation such as time preferences and beliefs.

Caution: Be careful what elicitation instrument one uses and know what subjects are taking away from it. I.e. do not mis-specify the DGP!

Potential for:

1. Extensive piloting.
2. Objective/quantitative tests of understanding and decisionmaking.
3. Qualitative understanding & debriefing, but beware of "endogenous" explanations (ex-post rationalization).

Demand/Hawthorne effects

Two examples:

1. LFE: Cilliers et al. (WP, 2014), `http://goo.gl/0JwMme`.
2. RCT/survey: Zwane et al. (*PNAS*, 2010), `http://www.pnas.org/content/108/5/1821.full.pdf`.

Cilliers et al. 1

1. Experimentally vary white foreigner presence in dictator games across 60 villages in Sierra Leone.
2. Presence of a silent white foreigner increases player contributions by 19 %.
3. Players may want to impress the white foreigner by being more generous. Consistent with this, game contributions are no longer predicted by real-world public good contributions when the white foreigner is present.
4. The white foreigner's presence may make those more familiar with aid perceive the games as a form of means testing. Consistent with this, in the presence of the white foreigner, players in more aid-exposed villages give less and are more likely to believe that the games are testing them for aid suitability.
5. This is what Viceisza (Guide 2012, *JES* 2015) calls the "mzungu effect".

Cilliers et al. 2

Take-away: Hold foreigner presence across your treatments if at all.

Crudely:

1. Assume bias$= \beta$ such that behavior in treatment A is $A + \beta$ and in treatment B is $B + \beta$.
2. So, treatment effect $\tau = (A + \beta) - (B + \beta) = A - B$ under certain conditions.
3. This is not the case if β is different in A than in B.

Zwane et al. 1

1. In the three health experiments, they find that being surveyed increases use of water treatment products and take-up of medical insurance.
2. Frequent surveys on reported diarrhea also led to biased estimates of the impact of improved source water quality.
3. In two microlending studies, they do NOT find an effect of being surveyed on borrowing behavior.
4. They conclude that the results suggest that limited attention could play an important but context-dependent role in consumer choice.

Zwane et al. 2

Take-away: Researchers should reconsider whether, how, and how much to survey their subjects.

Related to "priming" and behavioral cues.

Other related issues

1. Experimenter/enumerator-subject gender-interaction effects (sociology).
2. Same for race- or ethnicity-interaction effects (same).
3. Non-white (other) foreigner presence effects.
4. Language, look, etc. effects.

Key role for experimental design to mitigate such effects:
If β is constant across treatments, it drops off.

If β is nonconstant, then one should elicit/control for it!

Key Q: How do you know if it is constant across treatments?
Role for e.g. piloting, theory, previous findings.

Power 1

Spybrook et al. (2011).

Optimal Design software, `http://hlmsoft.net/od/`.

Documentation, `http://hlmsoft.net/od/od-manual-20111016-v300.pdf`.

Power 2

Useful for ex-ante and ex-post power calculations.

1. Experimental/empirical design.
2. Grant applications (e.g. DOE/FITW grant).

Ex ante: priors for (1) variance, (2) effect sizes, (3) intra-cluster correlations.

Can do sensitivity analysis and pull from previous literature (comparable to calibration exercise).

Viceisza (2012): Also mentions references such as Grosch and Muñoz (World Bank LSMS, 1996); see `http://goo.gl/ymNuhT`.

Deception

What is deception?

Difference between lies of omission and lies of commission.

Is deception right or wrong?

Are there ramifications for publication?

Anecdotes RE Hill et al. (*JDE*, 2012) and Torero and Viceisza (2015).

IRB 1

What is IRB?

What is the history of IRB?

Is IRB only a US thing and if so, should it be?

How does one deal with IRB internationally?

IRB 2

Why seek IRB approval even if not in the US?

Is the researcher the best judge of potential risks associated with the research?

Journals and IRB.

IRB anecdotes...

Pre-analysis plans 1

What is a pre-analysis plan (PAP)?

Main reference: Coffman and Niederle (2015), `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.29.3.81`.

A pre-analysis plan is a credibly fixed plan of how a researcher will collect and analyze data, which is submitted before a project begins. So, in its extreme, you tell me before the project begins the EXACT regressions you plan to run (no more, no less).

Pre-analysis plans have recently becomes more popular across the social sciences; e.g. Humphreys et al. (2013), Monogan (2013), Miguel et al. (2014), & Berge et al. (2015).

Purpose: Transparency?

Pre-analysis plans 2

Pos. Pred. Val. (PPV)$= \frac{[1-\beta^k(1-u)^k]\pi}{[1-\beta^k(1-u)^k]\pi+(1-\pi)[1-(1-\alpha)^k(1-u)^k]}$

where:

1. $\alpha = 0.05$ (statistical significance)
2. $\beta = 0.2$ (type II error; $1-\beta$ is power of the study)
3. $\pi \in \{0.3, 0.5, 0.7, 0.9\}$ (proportion of studies testing the true hypothesis)
4. $u \in \{0.01, 0.1, 0.25\}$ (study bias–the probability with which a study that would have been reported false without any bias is instead reported positive)
5. $k \in \{1, 10, 25\}$ (the number of studies that were (or could be) investigated.)

Pre-analysis plans 3

Benefit?

Table 1 calculates PPV and ΔPPV as study bias u is reduced for any given k. Findings:

1. A PAP that reduces u is most effective when $k = 1$ and/or π is lower.
2. If $k > 1$, a PAP is most helpful when it reduces u to almost zero.
3. For $k > 1$, a PAP that reduces u has very little effect unless it is almost eliminating it.
4. Conclusion: PAPs mainly make sense when $k = 1$ and π is low, e.g. large-scale and costly field experiments/studies that are unlikely to be replicated. Hence the title of the paper!

Pre-analysis plans 4

Figure: PPV from Coffman and Niederle (2015)

Table 1

How Reducing Within-Study Bias Affects Probability that a Published Positive Result Is True (PPV), by Number of Substitute Studies and Expected Probability That a Hypothesis Is True

Number of substitute studies:		*1 study*		*10 studies*		*25 studies*	
Expected probability of true hypothesis	***Bias***	*PPV*	*ΔPPV (from row above)*	*PPV*	*ΔPPV (from row above)*	*PPV*	*ΔPPV (from row above)*
0.30	**0.25**	0.56	–	0.31	–	0.30	–
	0.10	0.71	0.15	0.35	0.04	0.30	0.00
	0.01	0.86	0.14	0.52	0.17	0.37	0.07
0.50	**0.25**	0.75	–	0.51	–	0.50	–
	0.10	0.85	0.10	0.56	0.05	0.50	0.00
	0.01	0.93	0.08	0.71	0.16	0.58	0.08
0.70	**0.25**	0.87	–	0.71	–	0.70	–
	0.10	0.93	0.06	0.75	0.04	0.70	0.00
	0.01	0.97	0.04	0.85	0.11	0.76	0.06
0.90	**0.25**	0.96	–	0.90	–	0.90	–
	0.10	0.98	0.02	0.92	0.02	0.90	0.00
	0.01	0.99	0.01	0.96	0.04	0.93	0.03

Note: A significance level of 0.05 and power of 0.8 is used throughout; "PPV" refers to the "positive predictive value" as in Ioannidis (2005), which is the probability of a result being true given a positive result.

Pre-analysis plans 5

Other considerations

1. Costs & benefits of PAPs.
2. Costs & benefits of hypothesis registries.
3. Costs & benefits of replication studies.
4. Suggested *Journal of Replication Studies*.

Lecture 2: LFEs (R, 11/26)

LFEs: References

Two main references

1. Viceisza (2012), IFPRI guide, `https://goo.gl/1h1Jkt` (complete download form).
2. Viceisza (2015), *JES* article, `https://goo.gl/FIWYP3`.

LFE purposes

Viceisza (2015)

1. Test theory/heuristics.
2. Elicit parameters/characteristics.
3. Unpack black box/combine with other empirical methods.
4. Methodological explorations.
5. To teach concepts/alter behavior in other contexts.

Exercise: Qs

What is internal validity (IV)?

What is external validity (EV)?

Which is more important?

Exercise: Response

Crudely, IV is about whether or not your study does what it sets out to.

More specifically, IV is about whether or not an experiment that seeks to identify a causal treatment effect actually achieves such goal.

So, IV is intimately tied to proper experimental design (avoiding confounds).

We will discuss external validity (EV) later, but in principle, one must first worry about IV. Crudely, one could say that IV is necessary, but not always sufficient for a well-designed experiment.

IV considerations 1

Well-defined research question.

Clean/simple treatments to test such research question.

Experimental design issues (not exhaustive):

1. Between (balancedness) and/or within-subjects (order/learning) effects?
2. Random assignment to treatment and if so, at what level? Clusters or individual? This has implications for ex-ante power calculations as well as ex-post analysis (clustering of standard errors).
3. Mitigate spillovers (dividers) or exploit this? (location in the room)

Notice that 2 has historically not been a major concern in lab experiments. Why?

IV considerations 2

Well-defined experiment protocol:

1. To frame or not to frame?
2. To incentivize or not to incentivize?
3. Classroom or individual?
4. Paper and pencil or computerized?
5. Live translation or train experimenter – mzungu effects?
6. How to gauge subject understanding?

IV considerations 3

Other considerations:

1. Pilot, pilot, pilot..
2. Do not underestimate intuition cultivated by in-country discussions.
3. Collect and exploit additional survey data as well as administrative data.
4. Do not underestimate the power of ex-post debriefing.
5. E.g. Torero and Viceisza (*JEBO*, 2015) – built into the design, `http://goo.gl/karlAk`.

Exercise: Qs

How do we check for internal validity?

What to do if one lacks internal validity?

Exercise: Response 1

How do we check for internal validity?

Check for baseline equivalence using pre-characteristics.

Check for balancedness ex post.

Typical: Between-subjects – check whether characteristics are different across treatments on average. But, one can do full distribution tests as well.

Typical: Within-subjects – check whether there are any order/learning effects across treatments.

Exercise: Response 2

What to do if one lacks internal validity?

Control for confounding issues.

Simple: include certain covariates in regression and/or test for order effects and include time/order dummies.

More complicated: treat experimental data as non-experimental $\rightarrow$ instrumental variable, matching, weighting.

E.g. Hill and Viceisza (*Exp. Econ.*, 2012); `http://goo.gl/76Uj4S`.

Exercise: Response 3

The biggest concern is to what extent lack of IV gives rise to unbalanced unobservables that cannot be controlled for!

This can also be an issue if there is NO evidence of lack of IV. But, people are less likely to raise it in such case.

EV

Recall the question we asked earlier: "What is external validity (EV)"?

EV

EV aka generalizability asks to what extent findings from one context generalize to another.

See Al-Ubaydli and List (*NBER WP* 19757, 2012) for a model; `http://www.nber.org/papers/w17957`. I recap some of this in Viceisza (*JES*, 2015). Also see *NBER WP 19666*.

Also see `http://www.nber.org/papers/w20877` on "control" in NFE relative to LFE/lab.

Exercise: Qs

How do you assure EV?

How do you test for EV?

Exercise: Response 1

In a sense, the ONLY way to test for generalizability is to run a study in more than one context.

Case for replication?

Many people think of EV on the spectrum from the lab to the field, but that is not necessarily the way to think about it.

Again, it depends on the purpose of one's experiment/study more generally.

Exercise: Response 2

Typically, it is tied to representativeness of a sample. So, be careful when sampling and make sure you have some characteristics of the full population.

But, note that generalizability has to do with the treatment effect – not the characteristics per se!

So, perhaps use other sources of data to correlate with the main outcome and treatment effect? E.g. actual commercialization with contributions in a coordination game OR actual production with effort in a real-effort task. And so on...

In some sense, generalizability is also related to the concept of parallelism discussed in seminal experimental literature.

Camerer-List debate 1

Is it a debate?

Let us assume it is. What is it really about?

How does this debate relate to Coffman and Niederle?

Camerer-List debate 2

Reference 1: Levitt and List (*JEP*, 2007), `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.21.2.153`.

Reference 2: Camerer (Handbook, 2015), `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1977749`.

Reference 3: Al-Ubaydli and List (*NBER WP 19666*, 2013), `http://www.nber.org/papers/w19666`.

Important note

The issue of "external validity" is not one that ONLY plagues experimental work.

One can ask of any research project whether its findings "generalize". E.g. theoretical work hinges on certain assumptions; other empirical work rests on certain assumptions and/or data limitations.

As some have argued, the reason why other work potentially makes less of a fuss about EV is because this issue is usually trumped by that of IV (identification of causal effects).

Budgetary aspects 1

Typically, subject payments (potentially, high stakes).

Labor: main experimenter, translator, assistants, enumerators for complementary data collection.

Expenses for setting up or renting a field lab (classrooms, equipment, etc.)

Public good contributions – e.g. payments to farmer groups for facilitation.

Transportation costs for experiment team & sometimes, for the subjects.

E.g. in Torero and Viceisza (2015) as well as Sänger et al. (*Ag. Econ.*, 2013) – bused farmers who were far away to/from the experiment site (Vietnam).

Budgetary aspects 2

Depending on the context, one may be able to contract all this out – e.g. Busara lab in Kenya.

Potential concern: less control as to what is done.

Proposal 1

Recall that one of your proposals must focus on an LFE.

In writing your three-pager, consider the above issues.

The proposal must strike a balance between addressing these main issues and being succinct.

Notice that you are also supposed to discuss your empirical strategy and the types of tables/analyses you anticipate coming from the project.

Proposal 2: Exercise (time permitting)

Spend X] minutes thinking about your LFE proposal in the context of what we just discussed.

Let $X = 90 - lecture \geq 0$ mins hopefully!

If not, see me during OHs.

Lecture 3: RCTs (F, 11/27)

Exercise: Qs

What is a randomized controlled trial (RCT)?

Why has this approach gained extensive popularity in (development) economics recently?

Exercise: Response

An RCT *randomly* assigns units of a population or subpopulation (sample) to (experimental) conditions.

The term is borrowed from the medical professions where historically there have been placebos and double-blind protocols.

Reasons for fame:

1. Economics cares (A LOT) about causal effects.
2. Some argue that randomized experiments are the "gold standard" for obtaining causal effects.
3. By-pass selection bias.
4. Good marketing, particularly to donors.

Caveat: terminology

Note that in development and several other economics subfields, RCT is taken to be synonymous with an NFE.

That is, a study in which units are randomly assigned to control and treatment conditions (without typically knowing so) AND outcomes are measured based on behavior/data collected in a naturally-occurring context. E.g. random assignment to "metacognition" at Spelman, `http://goo.gl/5JITP5`.

But, an LFE in which units are randomly assigned to treatment conditions is ALSO an RCT. I.e. nothing says that outcomes MUST be measured in a naturally-occurring environment. E.g. random assignment to intention games in Senegal (Aflagah et al. 2015).

Take heed, depending on who you are talking to.

RCT reference

Main reference on RCTs as defined in development: Duflo et al. (2007), `http://economics.mit.edu/files/806`.

Some of the issues discussed for LFEs "generalize" to RCTs (no pun intended).

RCT IV 1

Earlier it was said that RCTs are sometimes considered the "gold standard" for identifying causal effects.

Some dispute this – see later discussions of Deaton (2010).

Even if one accepts the "gold standard" statement, it is only true if the research design is internally valid! I.e. if a design does not achieve IV, an RCT could potentially lead to the same complications as non-experimental work.

So, let us discuss some design principles and potential threats to IV in RCTs.

RCT IV: Principle 1

Find a reliable implementing partner.

Most RCTs are conducted in conjunction with an existing agency (NGO, government, firm, school system, etc.).

This reduces recruitment costs, increases credibility, and ensures non-intrusive access to subjects (people not necessarily knowing that they are in an experiment).

- Can you see links with Al-Ubaydli and List (2015) discussion of NFEs potentially rendering more "control"?

Reliable partner: Understands and respects IV and why randomization achieves IV.

RCT IV: Principle 1 – Exercise: Q

How can an implementing partner pose a threat to IV?

RCT IV: Principle 1 – Threat 1

Incentives to depart from randomization, e.g. friends or elites not in treatment group.

RCT IV: Principle 1 – Dealing with Threat 1

Retaining control of randomization (clearly).

Stressing importance of protocol. If this is messed up, it reduces lessons learned for their own purposes.

Separate "church and state" (avoid collusion). Make sure that tasks are performed by people with different incentives.

Track participants through the whole process – unique id, not just name – match on different characteristics. E.g. bring official ids (state-issued, firm-issued).

RCT IV: Principle 1 – Dealing with Threat 1

Take pictures during information sessions (if applicable).

Conduct randomized audits ex post to check who showed up/was exposed to treatment.

Most importantly: You want this to act as an ex ante incentive to behave properly – so, let people know, but do not disclose full details such that there is low probability of cheating!

Potentially more important than avoiding manipulation is knowing if and when it occurred such that one can control for it!

RCT IV: Principle 1 – Dealing with Threat 1

The worst is thinking all is OK when it is not.

Some manipulations may manifest themselves in imbalances of certain characteristics, but what if they do not?

What to do if you suspect problems? Treat data as if non-experimental. Correct for potential selection (classic methods as applicable).

Punchline

It is important to find a trustworthy and reliable implementing partner.

Examine their incentives for manipulation.

Bandiera et al. (2011), `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.25.3.63`.

RCT IV: Principle 2

Think carefully about (1) population, (2) sample, (3) sample size, and 4) power.

RCT IV: Principle 2 – Exercise: Q

How can sample size and power pose a threat to IV?

RCT IV: Principle 2 – Threat 2

Too few observations within each treatment condition may lead to a study that is underpowered and unable to detect effects.

While this is not a lack of IV per se, one wants to be confident that a null-effect is a true effect!

In that sense, an underpowered study is a badly designed study.

Recall Optimal Design software discussed for LFE.

RCT IV: Principle 2 – Threat 2

To assess power, one must consider not just the sample size, but also the outcome variable(s) under consideration – malleability, distribution (variation).

Some outcomes may be easier to manipulate than others (e.g. perceptions versus actual behavior).

This still begs the question of short, medium, and long run impacts.

RCT IV: Principle 2 – Threat 2

What do you do if a study is underpowered?

One approach is to collect additional data, but that is costly and not clean.

You can potentially use small-sample estimation strategies.

Also see literature on bootstrapping.

But, preferably, avoid such situations.

It is better to cut the number of treatments and increase $N/treatment$!

RCT IV: Principle 3

Think carefully about issues raised in Principle 2 PLUS level/unit of randomization.

E.g. level of randomization may mitigate spillovers.

RCT IV: Principle 3 – Exercise: Qs

How can the level of randomization increase IV?

RCT IV: Principle 3 – Threat 3

Example:

1. You want to study the impact of a health information campaign on beliefs and cleanliness practices.
2. You randomize 2000 individuals in 50 villages across an information treatment and a control.
3. Is there a concern?

RCT IV: Principle 3 – Threat 3

Example:

1. Given you have individuals who are treated and untreated in the same village, this is problematic.
2. Spillovers could cause treatment and control individuals to behave virtually the same.
3. If you find an effect, great, since spillovers work against you, BUT what if you do NOT find an effect?
4. You could also try to elicit social networks and control for spillovers. In fact, this may be interesting nonetheless.
5. Also, see Duflo et al. (2007), pp. 56-58.

RCT IV: Principle 3 – Threat 3

How to deal with this?

1. Well, what if you randomize at the village level?
2. I.e. $N = 25$ treatment villages and $N = 25$ control villages.
3. You can still perform individual-level analysis, but control for village fixed effects/cluster standard errors at the village level.
4. Spillovers are less likely.
5. You could also do $N = 10$ where randomization is within village to study spillovers and then remaining $N = 40$ over treatment and control.
6. Clearly, randomization at a higher level comes at a cost.

RCT IV: Principle 4

Think carefully about whether or not the actual treatment is the same as the intention to treat!

E.g. participants decline to participate (cannot force).

E.g. attrition – participants agree to participate, but do not show up or drop out over time.

RCT IV: Principle 4 – Exercise: Qs

How can the above impact IV?

RCT IV: Principle 4 – Threat 4

If participants systematically decline to participate or drop out and this is correlated with treatment status, this is problematic.

It could be like selection bias.

What to do? Intent-to-treat $\rightarrow$ instrument actual treatment with randomized treatment (intent-to-treat). See Duflo et al. (2007), pp. 50-56.

RCT IV: Principle 4 – Threat 4

What to do with attrition?

1. Report attrition statistics and the extent to which it appears balanced across treatments and controls.
2. Even if balanced, it may be problematic. Assess to what extent attrition is correlated with other observables.
3. Are those who drop out significantly different from those who do not or the population? Compare on observables at baseline for full sample.
4. See Duflo et al. (2007), pp. 58-61 for parametric and non-parametric references such as Manski-Lee bounds.

RCT IV: Principle 4 – Threat 4

Significant attrition will reduce power of the study.

This is another reason for thinking carefully about N during the experimental design phase.

This is of particular importance if you have a panel study in mind. I.e. tracking subjects over multiple years. Especially, if subject pool is likely to migrate.

Recall importance of tracking principles discussed previously. They come into play here too!

RCT EV: Exercise: Qs

Why may EV be a concern in RCTs?

RCT EV: Exercise: Answers

For the same reasons that it was a concern in LFEs.

People will ask whether the findings of an RCT generalize to the naturally-occurring environment.

Particularly when the RCT is intended to inform policy.

Also see Duflo et al.'s (2007) discussion RE ITT versus ATE in the context of policy recommendation.

RCT EV: Role of theory

Models can guide the extent to which an effect may be expected to generalize.

This can be combined with data to do out-of-sample predictions or simulations.

Recall that it is about the extent to which a treatment effect generalizes and therefore, we must know something about mechanisms and conditions.

RCT EV: *JEL* symposium

Key article: Deaton (2010), `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jel.48.2.424`.

A must read!

In what follows, selected quotes/aspects raised.

RCT EV: *JEL* symposium

"...instrumental variables have moved from being solutions to a well-defined problem of inference to being devices that induce quasi-randomization..."

See his examples on pp. 427-30.

RCT EV: *JEL* symposium

"...I shall argue that, under ideal circumstances, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes..."

RCT EV: *JEL* symposium

"...I argue that evidence from randomized controlled trials can have no special priority. Randomization is not a gold standard because "there is no gold standard" Cartwright (2007a.) Randomized controlled trials cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence..."

"...More positively, I shall argue that the analysis of projects needs to be refocused toward the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work..."

RCT EV: Exercise: Qs

Deaton (2010) draws a distinction between "external" variables and "exogenous" variables.

Are these the same? If not, how are they different? Which is stronger?

Which more closely resembles our "current" use of the term "exogeneity"?

RCT EV: Exercise: Answers

Exogeneity is a stronger requirement than externality.

Exogeneity requires that an instrument meet the exclusion restriction, be orthogonal to the error term.

There is no empirical test for it, as it is an identifying assumption that must be made prior to the analysis.

RCT EV: *JEL* symposium

Section 4:

1. Assume one implements an ideal RCT.
2. It tells you something about the average treatment effect, but not distributional effects per se; see pp. 438-39.
3. It is also incorrect to estimate $Y_i = \beta_0 + \beta_1 * T_i + \Sigma_j \theta_j * X_{ij} + \Sigma_j \phi_j * X_{ij} * T_i + u_i$. Technically, one needs to run (new) RCTs on those subgroups.
4. As he argues, it is tempting to run such regressions to test mechanisms (via heterogeneous impacts), but that does not fit the strict RCT standard!
5. See any links with pre-analysis plans?

RCT EV: *JEL* symposium

Section 4:

1. Most RCTs are NOT ideal, as discussed before.
2. It is often incorrect to just compare means or run OLS on $Y_i = \beta_0 + \beta_1 * T_i + u_i$ because variances of treatments and controls are impacted by intervention.
3. It is also incorrect to estimate $Y_i = \beta_0 + \beta_1 * T_i + \Sigma_j \theta_j * X_{ij} + u_i$.
4. Several aspects of concern; in particular, data mining.
5. See any links with pre-analysis plans?

Lecture 4: Ways forward (F, 11/27)

Tips 1

Note that while the term 'experiment' is somewhat popular nowadays, a paper will not get published just because it is based on an experiment!

Several considerations impact publication:

1. Design, IV, EV, novelty, policy relevance.
2. Your network – who you know; are you part of a mafia?
3. Editor's preferences.
4. More when discussing publication strategies.

Tips 2

Ask yourself if an experiment is the best way to test your question?

If so, what type of experiment or experiment**S**?

LFE, NFE/RCT, Lab, Natural experiment in isolation or combined with each other or other methods such as structural based on survey/admin data?

Always think through the role for theory (or at least a conceptual framework).

In development/ag-econ, it is not uncommon for reviewers to disclaim your model as unrealistic/restrictive. So, sometimes, let reviewers tell you what they want in a model.

Tips 3

We return to the core: What is the question you would like to ask/answer?

Assume an experiment is part of the way to answer this question.

What is popular nowadays?

Tips 4

Combined experiments: Lab or LFE combined with RCT.

See Viceisza (*JES*, 2015) – conclusion.

Examples abound in the literature – some of my own; Aflagah et al. (2015a, 2015b); Bernard et al. (2015a, 2015b).

See Camerer (2015), `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1977749`.

Better yet: Combined experiments with other methods (structural)?

Aflagah et al. (2015a)

Link to presentation: `https://goo.gl/TBG7xh`.

LFEs: coordination & intentions.

RCTs: control & intentions treatments.

Use findings from LFE to argue empirical strategy/mechanisms in RCT.

Aflagah et al. (2015b)

Link to presentation: `https://goo.gl/TBG7xh`.

LFEs: coordination & intentions.

RCTs: control & intentions treatments.

Exploit individual- and group-level variation in exposure to LFEs to identify impacts on behavior in RCT. Claim: Lab experiments can impact naturally-occurring behavior.

Do we see the link with Camerer-List debate?

Tips 5

If just an LFE, think carefully about:

1. What is the question I would like to ask/answer?
2. Can I answer it with just the data from the LFE?
3. What do I gain from conducting an LFE relative to a lab experiment?
   - Note: The answer should NOT just rest on the subject pool!
4. If the question is policy-relevant, be concerned about generalizability.
5. Can you combine these data with pre- and post-survey and administrative data?
6. Can the LFE complement other empirical methods?

Tips 6

Combined empirical methods, i.e. experiments & non-experiments.

Example 1: Mahajan and Tarozzi (2011), `https://are.berkeley.edu/~aprajit/DDC.pdf`.

Approach: Combine an LFE on time preferences with a structural model. Specifically, use LFE to identify parameters of naiveté.

Example 2: Avis, Ferraz, and Finan (2015).

Approach: Natural experiment of government audits combined with structural model/simulation to identify mechanisms.

Tips 7

Other considerations:

1. Multiple experiments over time.
2. Long-run impacts, e.g. Schaner (2015) – `http://www.dartmouth.edu/~sschaner/main_files/Schaner_LongRun.pdf`.
3. Behavioral dimensions – Mullainathan-Shafir-type stuff.
4. Neuro- and bio-economic data – e.g. Giné et al. (2012) – `http://econweb.umd.edu/~goldberg/docs/gine_goldberg_yang_AER2012.pdf`.

Tips 8

One significant area where more work is needed in developing country contexts: methodology.

E.g. risk, time, demand, framing, long-run, etc.

But, beware: Reviewers (both for grant proposals and publications) are very critical of methodological papers.

Also, agencies may be unwilling to fund methodology, as they prefer policy relevance. So, build methodology into policy.

As a general rule, always diversify your portfolio of treatments – unfortunately, for better or worse, there is publication bias towards nonzero/non-null effects!

Tips 9

Again: Given your focus, many of you may be drawn to conduct LFEs in developing country contexts.

Ask yourself why a particular experiment in that context?

For example, is a conventional lab experiment better suited?

Is an RCT better suited? Think with "contributions" in mind first; then, with your wallet!

We really are in a world of combined methods; both within and outside economics...

Exercise: Qs

What do you consider a good publication?

Specifically, what do you consider the top journal(s) in your field?

Exercise: Answers 1

Australia has a ranking across disciplines: `http://www.abdc.edu.au/pages/abdc-journal-quality-list-2013.html`.

Tinbergen also has a ranking within econ/finance: `http://www.tinbergen.nl/discussionpaper/?paper=2283`.

Exercise: Answers 2

Economics: AER, Econometrica, JPE (Chicago), QJE (Harvard Econ), ReStat (Harvard KSG), ReStud (EU), EJ (UK/RES).

Experimental: GEB, Experimental, JEBO [JEPsych, JESA, RoBE]

Development: AEJ: Applied, JDE, $\sim$ JPubE, $\sim$ JoLE [EDCC]

Ag-Econ: AJAE, above [Ag. Econ, ERAE, FP]

Exercise: Answers 3

You should ask yourself: Where do I want to publish and why?

Given the answer to this question, you should have a submission strategy in place!

What follows are necessary, but not sufficient conditions for publication in top journals (general interest or field, with second-tier as fallback).

My two worst (lack of) publication stories...

1. 11-month review followed by a rejection.
2. 10-month resubmission review followed by an acceptance.

My two best publication stories...

1. Relatively quick rejection with 1 review claiming deception followed by a 3-month R&R with 1 out of 2 reviews claiming the best design.
2. Editors disagreeing with reviewers & accepting paper subject to revisions, which improved the paper.

Top-5 experiences

Desk rejections and then what?

Passed on to me: "The cigarette break" anecdote.

One thing is certain: If you do NOT submit, you will NOT get in!

So, at the end of your career, you do not want to say: "I didn't get in, cause I never submitted there!"

References

Most of what follows is borrowed from presentations by Tim Cason, Yan Chen, and Daniel Houser at the ESA International Meetings.

See this link, `https://www.economicscience.org/resourceDisplay.html?catId=1`.

Note that Marie Claire Villeval also has some useful tips in that same presentation. Due to time limitations, I have not included them here. Time permitting, we will review some of them.

Dan Houser: Advice

1. A good paper is only as long as it needs to be.
2. Papers must be written in native English – there is no shame in asking (or paying) for a paper to be edited by a native speaker.
3. Hypotheses must be compellingly motivated and, ideally, tested using simple procedures.
4. Conclusions must be supported by the data; a strong prior with inconclusive data is not generally sufficient (or why run the experiment?)

Yan Chen: Advice 1

1. Present your paper at many venues and use feedback.
2. Ask yourself whether an idea is AER-worthy.
3. After a paper is polished (i.e., copy edited), you should circulate it among experts (potential referees). Reduce the expert's search cost: "Your paper is cited on page 17".
4. How to get invited? Invite yourself, e.g., when you circulate your paper. Reduce your host's cost (combining with conferences). Organize seminars at your department.

Yan Chen: Advice 2

1. *Management Science*: DE $\rightarrow$ AE (anonymous) $\rightarrow$ referees $\rightarrow$ AE $\rightarrow$ DE.
2. When you suggest AEs and referees, do not waste slots. Suggest relevant referees, e.g., on your reference list. Suggested referees should not be too senior.
3. *Games and Economic Behavior*: You cannot suggest AEs, but might influence who handles your paper (through citation).
4. Other journals: should you suggest referees? Yes!

Tim Cason: Advice 1

What to submit?

1. Length. Is shorter always better? Usually.
2. Include appropriate appendices to ensure a thorough review (see old Econometrica guidelines) such as exact experiment instructions. And, translate them correctly!
3. Supplemental materials – be judicious; do not show everything you tried in the analysis that was not interesting enough to include in the paper.
4. Inserted by Viceisza: A dissertation chapter, essay, or job market paper is typically NOT a paper ready for submission, unless you really wrote it in that style!

Tim Cason: Advice 2

Dealing with resubmissions and rejections:

1. Drop everything so that you can do a thorough revision quickly (especially when your tenure clock is ticking!)
2. Response letters should be detailed, explaining how each point is addressed.
3. Rejections can be good. Sometimes they provide the best feedback on your paper.
4. If a referee did not get an important aspect of your paper, it is your fault not the referee's.
5. Inserted by Viceisza: Address important comments even in a rejection – you are likely to get that referee again at another journal!

Additional Q&A

Any other questions or comments?