Intro to

HETEROGENEITY & MODERATION

Elizabeth Tipton

TREATMENT EFFECTS

- So far, we've focused a lot on the **average treatment effect**. Sometimes this is referred to as "the" effect or a "summary" effect — neither of which is quite right.
- It is possible — and maybe even likely — that treatment effects will vary. Why?
  - Often a "treatment effect" is a comparison of an intervention to **'business as usual'.** We know 'business as usual' differs a lot in schools in the US.
  - Interventions are often **implemented differently** in different places — sometimes adapted, sometimes combined with other curricula, and so on. Again, this would suggest variation in effects are likely.
  - Student **prior knowledge** differs — which suggests that effects might be larger for those who aren't doing as well (on the test) as those that are already proficient.

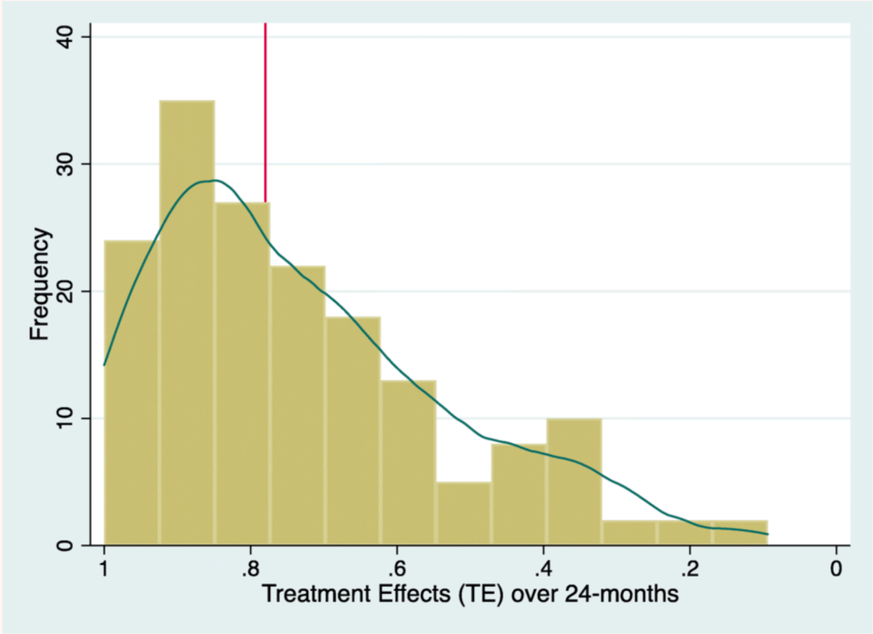
POTENTIAL OUTCOMES

- We can be more formal about this. For a very simple RCT, let $i = 1,...,n$ individuals be included in a study. Each of these individuals has two potential outcomes:
  - $Y_i(0)$: what their outcome would be if they continue with business as usual
  - $Y_i(1)$: what their outcome would be if they instead take part in the intervention
- This means that for each individual there is a unit specific treatment effect:

$$\delta_i = Y_i(1) - Y_i(0)$$

SUMMARIES

- This means that there is a distribution of treatment effects. This distribution has:
  - Mean = $E(\delta_i) = \delta = \mu_1 - \mu_0$
  - Variance = $V(\delta_i) = \tau^2 = \sigma_0^2 + \sigma_1^2 - 2\rho_{01}\sigma_0\sigma_1 \approx 2\sigma^2(1 - \rho_{01})$
- Notice that $\rho_{01} = Corr(Y_i(0), Y_i(1))$ is the correlation between the potential outcomes.



FUNDAMENTAL PROBLEM

- Unfortunately, we can't observe both $Y_i(0)$ and $Y_i(1)$ for each unit, and thus we can't observe δ_i.

Unit	Covariates	Treatment	Potential outcomes		Observed
i	X(i)	Z(i)	Y(1)	Y(0)	Yobs
1	5	1	**100**	90	100
2	15	1	**110**	80	110
3	10	0	100	**90**	90
...	...	...	...	...	...
n	20	0	120	**100**	100

WHAT WE DO OBSERVE

- However, because of randomization, we can get an unbiased estimate of our average treatment effect: $\hat{\delta} = \bar{Y}_1 - \bar{Y}_0$. This is what we've been doing so far!
- But the variance is more tricky — we don't know what ρ_{01} is. **Of course, we can try different values, but any exploration of heterogeneity thus requires additional assumptions.**
- Put another way: randomization allows us to estimate the average causal effect of an intervention without assumptions. The analysis is simple and straightforward to explain. But treatment effect heterogeneity requires assumptions and models — it is complex.

WHAT CAN WE DO?

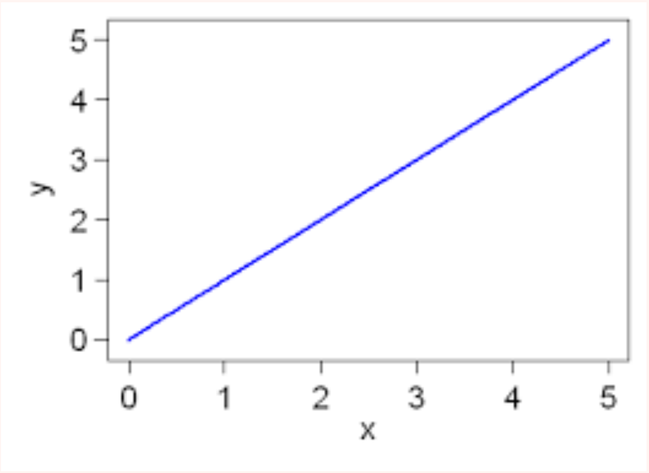
- In a simple RCT we might have the simple model:

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

- Or we might have the moderator model:

$$Y_i = \beta_0^A + \beta_1^A T_i + \beta_2^A X_i^c + \beta_3^A T_i X_i^c + \epsilon_i^A$$

INTERACTION



- Because we centered X_i^c, we have $\beta_0 = \beta_0^A$ and $\beta_1 = \beta_1^A$. However, now we have two different relationships:
  - For those in the comparison: $E(Y_i \mid C) = \beta_0^A + \beta_2^A X_i^c$
  - For those in the treatment: $E(Y_i \mid T) = (\beta_0^A + \beta_1^A) + (\beta_2^A + \beta_3^A) X_i^c$
  - **And thus our treatment effects:** $E(Y_i \mid T) - E(Y_i \mid C) = \beta_1^A + \beta_3^A X_i^c$

BUT LET'S LOOK MORE CLOSELY

- Returning to the unit specific treatment effects, we now have:

$$\delta_i \approx [\beta_0^A + \beta_1^A X_i^c] + \eta_i$$

- But we can't see this η_i because we can't observe both potential outcomes.
  - Thus when we observe an interaction, it is **part** of the treatment effect variation, but **not all** of it.
  - It is important to keep this in mind. We are trying to understand and explain variation in an outcome that we cannot observe — we are very much feeling around in the dark.

MODERATORS

TYPES

- It is helpful to conceive of different types of moderators.
- **Cronbach proposed 4 types:**
  - **Units:** e.g., different types of students, prior knowledge, subgroups, etc
  - **Treatments:** e.g., different versions of a treatment, different comparison conditions
  - **Outcomes:** e.g., different measures
  - **Settings:** e.g., different school types

PRIOR TO TREATMENT

- Like a covariate, a moderator needs to be observed before the intervention is implemented.
- Otherwise it is a mediator!
- Be careful here:
  - e.g., make sure the pre-test is measured before (or close to) the beginning
  - e.g., make sure any classification of students into subgroups is before

THINK THINK THINK

- If I say “moderator” the first things you will likely think of are:
  - Race / Ethnicity
  - SES
  - Gender
- **Is there a reason to expect that the treatment effect is different for these groups? This requires considering the mechanism of the intervention.**

MECHANISM

- Go back to your logic model. Where in this model do you see that the intervention effect might differ?
  - e.g., perhaps you suppose that a ‘problem X’ —> need for intervention. This suggests that different degrees of ‘problem X’ might impact the effect.
  - e.g., perhaps there are supports and resources required for it to be implemented well. This suggests that measuring the presence of these supports and resources may affect the impact.

Table 1 | Examples of common sources of treatment-effect heterogeneity in behavioural intervention research

Source of heterogeneity	Definition	Examples
Experimental procedure	Details of an intervention's implementation that might seem trivial can have a substantial impact on its effectiveness.	An intervention in which tax preparer H&R Block automatically pre-populated the Free Application for Federal Student Aid form for parents of college-eligible students using data already collected for tax returns increased college enrolment by eight percentage points[22]. A subsequent intervention in which participants were merely informed that tax data could be used to pre-populate the form and directed to a website that could help them do this had no detectable effect[27].
Research population	Members of some cultural or demographic groups or people with particular psychological characteristics (for example, high need for cognition or reward sensitivity) are more responsive to an intervention than others.	Many effects foundational to the nudge movement[53] (for example, conformity, heuristics and biases) were found to be substantially stronger in subpopulations that closely resemble the college-student samples in which they were originally documented (that is, younger, more educated and wealthier) than in the population at large. This finding is based on meta-analysis of replications conducted in nationally representative samples[62].
Objective or structural affordances of the context	Objective features of the context can afford more or less opportunity for the psychological effect of an intervention to lead to the targeted behaviour.	A growth-mindset intervention, which teaches participants that intelligence can grow with effort, was designed to prevent ninth graders from failing core courses. Pre-registered analyses revealed that it was effective in low- and middle-achieving schools, but had no effect on course failures in high-achieving schools. This is probably because high-achieving schools have such ample resources to prevent failures that the intervention was superfluous for that purpose[61].
Psychological affordances of the context	Subjectively experienced features of the context can afford more or less opportunity for the intervention to have the intended psychological effect.	An intervention that frames voting as a way to claim (or re-affirm) a desirable identity ('voter') increases turnout in major elections[23]. The same treatment has no effect in uncompetitive congressional primaries where the identity 'voter' does not feel important or meaningful[48,59,118].
	Even if an intervention has the intended psychological effect immediately, subjectively experienced features of the context can either support or undermine that psychological state.	A growth-mindset intervention, which teaches participants that intelligence can grow, has a larger effect in classrooms with norms that are supportive of a growth mindset. Its effect in classrooms with norms that do not support growth mindset is weaker[61] (this result comes from pre-registered analyses).

Bryan, Tipton, Yeager (2021)

WHATEVER YOU DO,
THINK CAREFULLY ABOUT THIS

WHY? MODELS? BEWARE

WHY DO WE CARE?

- This seems hard. Why should we care?
  - If an intervention works differently for different students, teachers, schools, communities and so on, then this means the ATE is simply not enough information to summarize the intervention's efficacy.
  - If an intervention effect varies information about this could be helpful for:
    - Understanding 'for whom and under what conditions' the intervention works — something decision makers care about.
    - Understanding the mechanism of the intervention — something scientists care about.

1. SUBGROUP EFFECTS

- The simplest question we could ask is: What is the ATE for different subgroups that are important to decision-makers?
  - e.g., providing separate ATE estimates for those with ‘low’ ‘average’ ‘high’ reading ability.
  - e.g., providing separate ATE estimates for different demographic subgroups
  - e.g., providing separate ATE estimates for different school-types — urbanicity, region, school structure, grade-levels and so on.

SUBGROUPS

- Subgroup effects can be thought of as ‘descriptive’ — our goal is simply to provide ATE estimates for different slices of the population.
- Some things to keep in mind:
  - We need to know what this population is so that we can describe the population and subgroup appropriately (e.g., the ATE isn’t for all “rural” schools, its for all rural schools in this population).
  - Splicing the data into subgroups reduces sensitivity — the overall ATE might have adequate power, but the subgroups likely do not. Thus, be careful with hypothesis testing.

2. DIFFERENTIAL EFFECTS

- Alternatively, we might be interested in understanding if the effect for one subgroup is different from another.
  - e.g., Is the ATE in rural schools different form the ATE in urban schools?
- We have to be careful here for a few reasons:
  - Interpretation issues
  - Confounding and causality
  - Power

A. INTERPRETATION

- Suppose we investigate this model:

$$Y_i = \beta_0^A + \beta_1^A T_i + \beta_2^A X_i^c + \beta_3^A T_i X_i^c + \epsilon_i^A$$

- And we find that β_3^A is non-zero (putting aside power).
- Interpretation:
  - If X_i^c is continuous we have: “The ATE is β_1^A and for each 1-unit change in X_i^c, the expected treatment effect changes by β_3^A units.”
  - If X_i^c is a centered dummy variable, now we have “The ATE is β_1^A and for the effects for those in Group 2 are β_3^A units larger [smaller] than those in Group 1.”

B. CONFOUNDING

- When we move to comparisons — interactions — it is easy to slip into causal language.
- However, while the ATE is a causal effect (due to randomization), interaction effects are **not** causal. They are observational.
- **We have to be worried about confounders.** For example, we find that an intervention reduces student suspensions for Black students more than non-Black students.
  - **What does the intervention have to do with race/ethnicity?**
    - Examining the data, we might find that *suspension rates (pre-test) are higher* for Black students than others. Thus we can reduce suspensions among Black students because they actually get suspended — whereas we cannot for other groups because they are less likely to get suspended.
  - This difference is subtle — but it points to how we interpret and attribute these differential effects.

C. STATISTICAL POWER

- Let's imagine we estimate ATEs for two subgroups (1, 2) and then we compare them:

$$\hat{\beta}_3 = \hat{\delta}_2 - \hat{\delta}_1$$

- Now let's look at standard errors:

$$SE(\hat{\beta}_3) = \sqrt{SE^2(\hat{\delta}_2) + SE^2(\hat{\delta}_1)}$$

- Notice that this standard error is more than twice as large!

- (This is not always the case — we will discuss situations later in which power is actually better for interactions than the ATE)

BEWARE TYPE II

- A good principle to remember is “Absence of evidence is not evidence of absence.”
  - You cannot prove the null hypothesis to be true. You can only prove it false.
- Put another way, if you planned your study design with a focus on the ATE — and the study is adequately powered for the ATE — it is possible that power if substantially lower for interactions / moderators / comparisons of subgroups.
- Thus, if you do not find the interaction effect significant, it could be because the treatment effects *do not* differ OR because the test *is very* underpowered.
- **Altogether this means you can prove that effects vary — but proving that the effect is constant is not possible.**

3. WHAT ABOUT OTHER VARIABLES?

- So far, I've focused on a single moderator with a single interaction. We might think of this as the case in which we have a "confirmatory" test.
- But what if we simply want to explore the data to see if we can build a model that predicts treatment effects? For example, maybe we collected p variables in the data and we want to know which subset of these p variables best explains variation in treatment effects.
  - If we approach this using hypothesis testing, we're going to need to worry about inflated Type I errors. That is, with multiple testing we're likely to end up with something significant just by chance.
  - Taking a step back, we can see that if there are 4 variables, we have many possible models: ABCD, ABC, ABD, ACD, BCD, AB, AC, AD, BC, BD, A, B, C, D. This suggests that this is a model selection problem — and a predictive model selection problem at that.

PREDICTIVE MODELS

- Once we move into the predictive model world, everything becomes more complicated.
- For example:
  - Instead of hypothesis testing, measures of model fit matter more.
  - These measures of model fit include penalties for the inclusion of too many variables (e.g., think adjusted-R^2, AIC, BIC).
  - Fitting this many models manually is tricky. Here is where algorithms / computational tools can help.
  - Some of these best methods out there are Bayesian Causal Forests — which involve a combination of Bayesian models, random forests, tuning parameters, and so on.
- **Overall, this is to say: Proceed carefully. This work can be rewarding — but it requires strong methods, sound reasoning, and good computational skills.**

CLUSTER DESIGNS

2-LEVEL CLUSTER RANDOMIZED

- **Let's return to the CRT.** Recall, we have student i in school j, and schools are randomized.

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij}^c + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_0 + \gamma_1 T_j + \gamma_3 W_j + \gamma_5 W_j T_j + \eta_{0j}$$

$$\beta_{1j} = \gamma_2 + \gamma_4 T_j$$

- In a single model we have:

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_{ij}^c + \gamma_3 W_j + \gamma_4 X_{ij}^c T_j + \gamma_5 W_j T_j + \eta_{0j} + \epsilon_{ij}$$

IN DETAIL

- Let's look more carefully:

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_{ij}^c + \gamma_3 W_j + \gamma_4 X_{ij}^c T_j + \gamma_5 W_j T_j + \eta_{0j} + \epsilon_{ij}$$

- In this model:
  - β_4 **is a cross-level interaction.** It describes how the treatment effect differs across student characteristics.
  - β_5 **is a site-level interaction.** It describes how the treatment effect differs across different types of schools.

EXAMPLES

- **Cross level moderators:**
  - Are treatment effects different for students in 3rd and 4th grade? (Perhaps the intervention is better in one grade than another)
  - Are treatment effects different for students from historically excluded groups than for others?
- **Cluster level moderators:**
  - Are treatment effects different for schools with lower pre-test scores?
  - Are treatment effects different for schools that had been using program X versus those using Z in the prior year?

CLUSTER LEVEL MODERATOR

- Let's start with a simpler model, with only a cluster level moderator.
- Let W_j indicate if a school is small (=1) or not (=0).
- Assume for now that we include $M = m_t + m_c$ schools in the study and that $M/2$ schools are small or not (equal allocation to the subgroups).
- Then we have:

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_3 W_j + \gamma_5 W_j T_j + \eta_{0j} + \epsilon_{ij}$$

CONT'D

- We can estimate γ_5 using: $\hat{\gamma}_5 = [\bar{Y}_{t2} - \bar{Y}_{c2}] - [\bar{Y}_{t1} - \bar{Y}_{c1}]$

- Thus we have:

$$SE(\hat{\gamma}_3) = 4\sqrt{\frac{n\bar{R}_w^2\sigma_2^2 + \sigma_1^2}{nm}}$$

MDESD

- If we move to the standardized effect $\delta = \frac{\gamma_3}{\sqrt{\sigma_2^2 + \sigma_1^2}}$ with ICC with $\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$ then,

$$SE(\hat{\delta}_5) = 4\sqrt{\frac{n\bar{R}_w^2\rho + (1-\rho)}{nm}}$$

- Thus, the **Minimum Detectable Effect Size Difference (MDESD)** is:

$$\delta_{MM} = 4M_{df}\sqrt{\frac{n\bar{R}_w^2\rho + (1-\rho)}{nm}}$$

STUDENT LEVEL MODERATOR

- Now, let's focus on a dummy variable D_{ij} which indicates if students are in 3rd grade (versus 4th). Again, assume this is balanced. Our model is:

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 D_{ij} + \gamma_4 D_{ij} T_j + \eta_{0j} + \epsilon_{ij}$$

- Notice here that I did not center the dummy variable. Thus:
  - γ_1 is the ATE for 4th grade classrooms
  - γ_4 is the difference in ATEs between 3rd vs 4th grade classrooms

CONT'D

- Now we have: $\hat{\gamma}_3 = [\bar{Y}_{t3} - \bar{Y}_{c3}] - [\bar{Y}_{t4} - \bar{Y}_{c4}]$

- And the standard error:

$$SE(\hat{\gamma}_3) = 4\sqrt{\frac{\bar{R}_D^2 \sigma_1^2}{nm}}$$

MDESD

- If we move to the standardized effect $\delta = \frac{\gamma_3}{\sqrt{\sigma_2^2 + \sigma_1^2}}$ with ICC with $\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$ then,

$$SE(\hat{\delta}_3) = 4\sqrt{\frac{\bar{R}_D^2(1-\rho)}{nm}}$$

- Thus, the **Minimum Detectable Effect Size Difference (MDESD)** is:

$$\delta_{MM} = 4M_{df}\sqrt{\frac{\bar{R}_D^2(1-\rho)}{nm}}$$

COMPARING THESE

- **Recall that in this model, our MDES is:**

$$\delta_M \approx \sqrt{2} M_{df} \sqrt{\frac{n\rho + (1-\rho)}{mn}}$$

- **If we use a cluster-moderator we have MDESD:**

$$\delta_{MM} \approx 4 M_{df} \sqrt{\frac{n\bar{R}_w^2\rho + (1-\rho)}{nm}}$$

- **If we use a student-level moderator we have MDESD:**

$$\delta_{MM} \approx 4 M_{df} \sqrt{\frac{\bar{R}_D^2(1-\rho)}{nm}}$$

TAKE-AWAYS

- **For a given design:**
  - Cluster-level moderators are less sensitive than the ATE.
  - Student-level moderators are more sensitive than the ATE.
- **Caveat:**
  - We have focused on dummy variables that are balanced as moderators. In real-life, these are likely not balanced, *thus reducing sensitivity*.