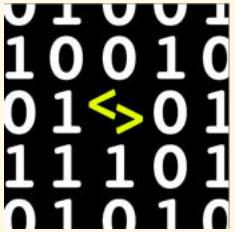
Encouragement designs for A/B testing



bit summit

Ben Elbers

About me

Data Science at Spotify since 2022, working in Product Insights.

Check out my website (elbersb.com) or reach out on LinkedIn (linkedin.com/in/elbersb)

This talk is based on a blog post:
Check out engineering.atspotify.com

Encouragement designs are an alternative design to the standard RCT design that is used in A/B testing.

It can* solve the problem that in a standard A/B test, some users don't have access to the feature we're introducing. This can be a problem if we…

- have lots of marketing,
- the feature has a viral/sharing component,
- or if the user expects the feature to be present (e.g., Spotify Wrapped)

* if some important assumptions are satisfied 🙃

In an encouragement design, we make the feature available to everyone, but only encourage users in the treatment group to use it.

Under certain assumptions, we can still learn something about the value of the feature.

One definition up-front:

If we launch a new feature, we have four different user groups:

- **Compliers**: Users who use the feature if they are encouraged to use it, but don't if they are not.
- **Always-takers**: Users who *always* use the feature regardless of whether they are encouraged or not.
- **Never-takers**: Users who *never* use the feature regardless of whether they are encouraged or not.
- **Defiers**: Users who always do the opposite: When they are assigned to the treatment they don't get treated; when they are assigned to control, they get treated.

A typology of A/B test designs

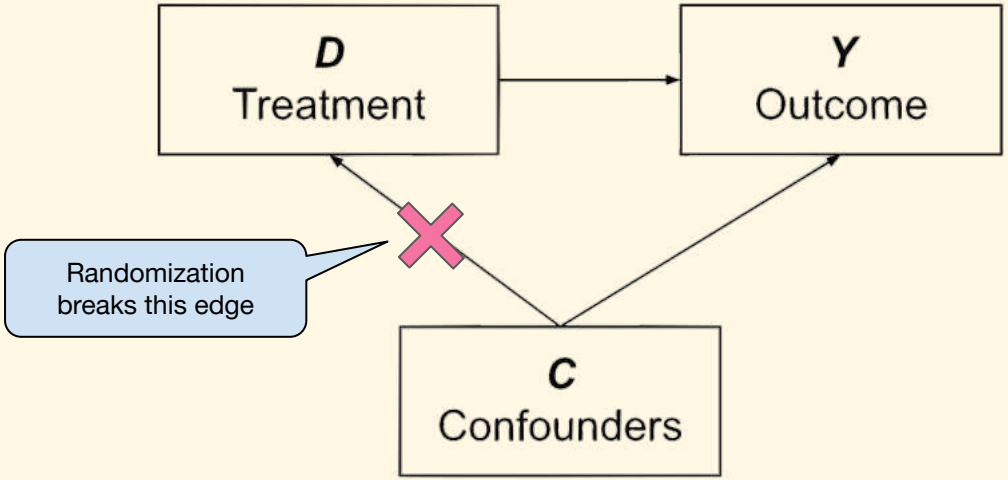
Type A
The gold standard

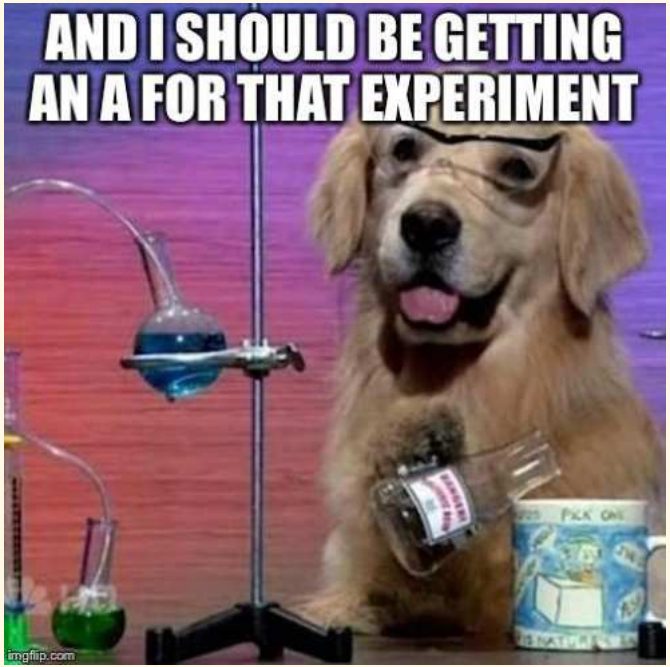
- We assign users randomly to treatment and control.
- Users are **fully complying** with the assignment, i.e. those in the treatment cell are treated, and those in the control cell are not treated.
- When we compare the difference in means between treatment and control, we identify the causal effect of the treatment on the outcome (**ATE**).

By definition, in a Type A test, we only have "**compliers**"

Random assignment	**Treatment status**
Treatment	Treated
Control	Untreated

Type A
The gold standard





Type B
Partial compliance

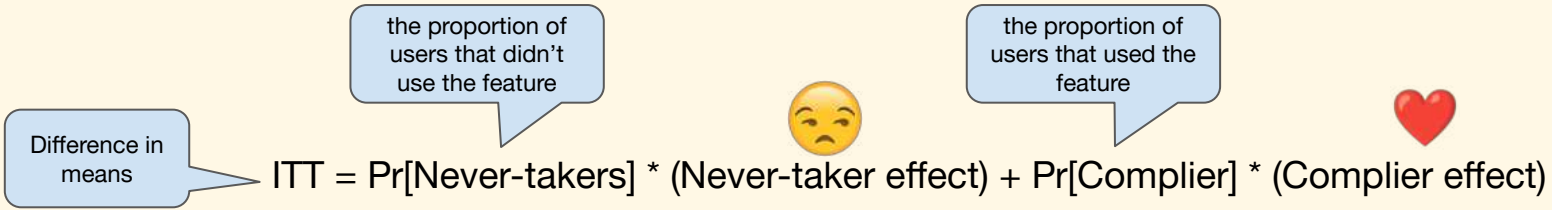
- Some A/B tests have partial compliance – for instance when we introduce a new feature.
- The difference in means now ***identifies the causal effect of being assigned to the treatment group*** – not the causal effect of using the feature.
- This is known as the **Intent-to-Treat effect** (ITT).

Because the control group doesn't have access to the feature, we can't have defiers and always-takers

Random assignment	**Treatment status**
Treatment	Treated (compliers)
	Untreated (never-takers)
Control	Untreated (compliers and never-takers)

Type B
Partial compliance

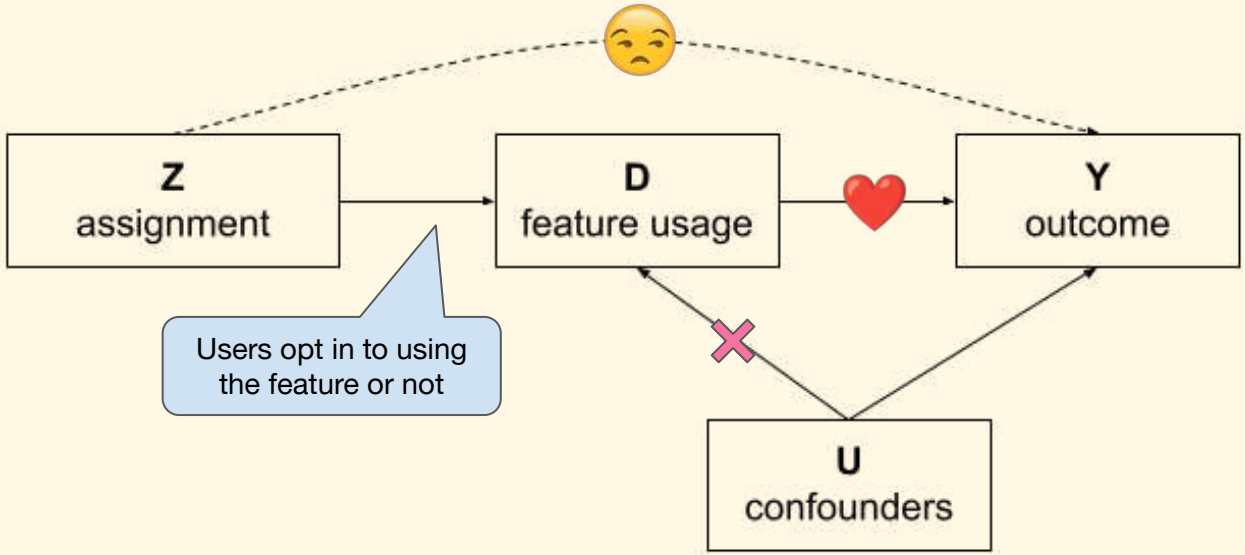
Write the ITT as **a weighted average**:



The complier effect could be large and positive, but if the never-taker effect is large and negative, the ITT will be 0.

Type B A/B tests mix "feature impact" and "feature presentation impact."

Type B
Partial compliance



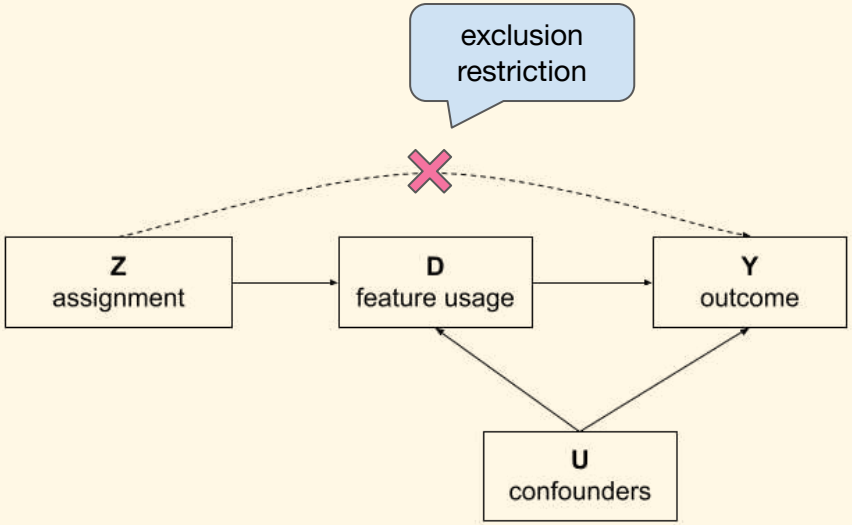
Type B
Partial compliance

If we assume that there is no direct effect Z → Y, then **the only reason for a change in the outcome has to be the feature usage**, D.

This is called the **exclusion restriction**.

We're saying that the never-takers are not bothered by the new feature.

Is this realistic?



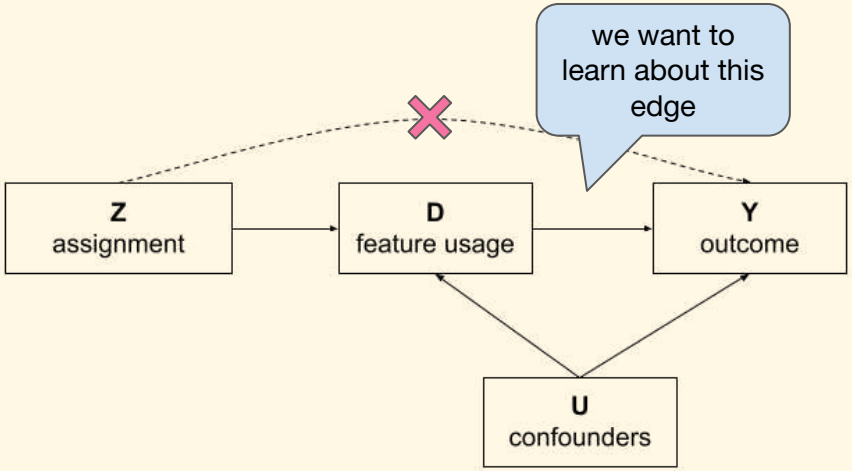
Type B
Partial compliance

With the exclusion restriction, estimate the causal effect of the feature itself *for the compliers* by dividing the ITT by the proportion of compliers.

Example: The ITT showed a retention uplift of 1pp, and 10% of users used the feature, then:

LATE = ITT / Pr[Complier] = 1pp/10% = 10pp

This is an **instrumental variables estimator** (Z is an instrument for D).

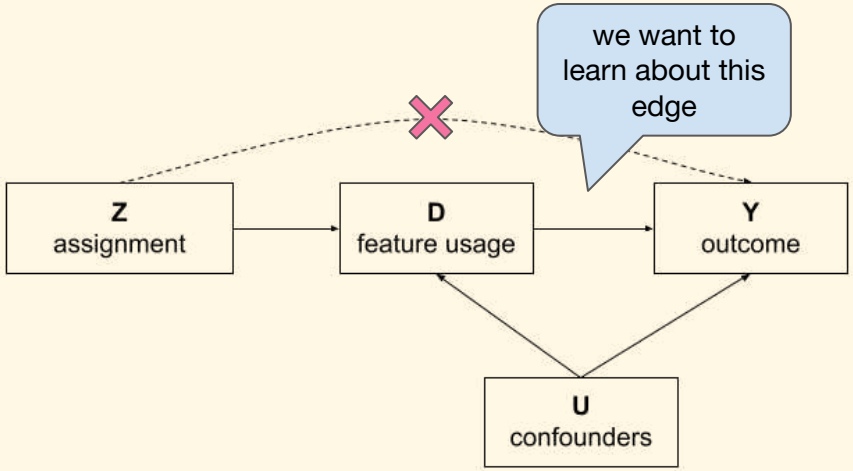


Type B
Partial compliance

The formula gives us the **LATE** – *local* average treatment effect –, not the **ATE**:

LATE = ITT / Pr[Complier]

We only learn about the causal effect for the compliers, not the entire population.



Wait, wasn’t this supposed to be about encouragement designs?

Type C
Encouragement designs

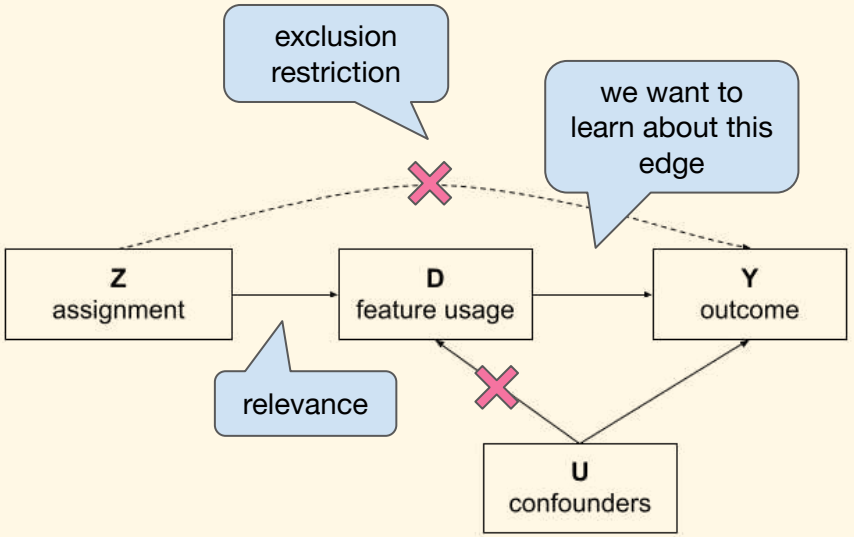
- We launch the feature for everyone
- We randomize whether a user is *encouraged* to use the feature (e.g., tooltip, message)
- The ITT measures the **causal effect on the outcome of being encouraged** – which is not what we are usually interested in.

Random assignment	**Treatment status**
Treatment (Encouraged)	Treated (always-takers and compliers)
	Untreated (never-takers and defiers)
Control (Non-encouraged)	Treated (always-takers and defiers)
	Untreated (never-takers and compliers)

Type C
Encouragement designs

To learn something from an encouragement design, we have to make three assumptions:

- The encouragement works: encouraged users use the feature more than non-encouraged users (relevance)
- No path Z → Y except through D (exclusion restriction)
- No defiers

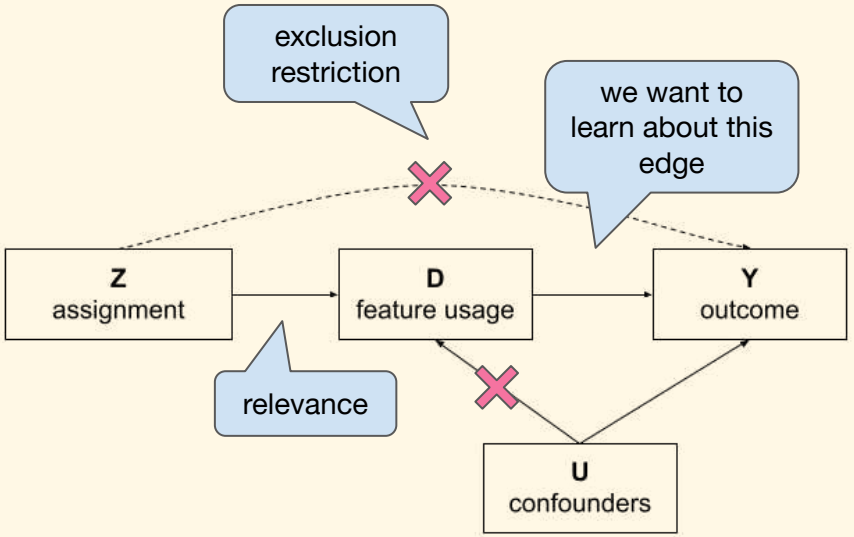


Type C
Encouragement designs

The logic is then the same as before for Type B: We estimate the proportion of compliers, and “blow up” the ITT:

LATE = ITT / Pr[Complier]

How’s a complier defined when we have non-compliance in both the treatment and control groups?



Type C
Encouragement designs

Estimate the proportion of always-takers in the population as
Pr[Treated | Control], and then subtract that proportion from the first cell:

Pr[Compliers] = Pr[Treated | Treatment] - Pr[Treated | Control]

The final formula for the LATE is then:

LATE = ITT / Pr[Compliers]
= (E[Y | Treatment] - E[Y | Control]) /
(Pr[Treated | Treatment] - Pr[Treated | Control])

Type C
Encouragement designs

The IV estimator takes an indirect way of estimating the treatment effect (think two regressions instead of one). This means that typically, standard errors are a lot larger compared to a standard A/B test.

If the encouragement doesn't work well, the IV estimator is very unstable. The better the encouragement works, the smaller the standard errors.

More on standard errors:
elbersb.com/public/posts/2023-10-07-iv-standard-error/

Type C
Encouragement designs

The most important part of the an encouragement design is the **exclusion restriction**.

The encouragement needs to be effective enough so that users that make use of the feature, but not too overbearing so that the exclusion restriction is violated.

This is difficult.

Compared to a regular A/B test with non-compliance, using an encouragement design solves some problems, but creates new ones:

Pros ("Better LATE Than Nothing")

We can give the feature to everyone –
no problems with marketing, word of mouth
(relaxes the SUTVA somewhat)

Easier to justify to stakeholders

If the feature works, potentially higher impact from launch

Cons

Unverifiable assumptions:
exclusion restriction, no defiers

Inference about a latent population — who's a complier is not observable

Higher standard errors – requires good encouragement and large sample sizes

Whether the encouragement design is useful depends on an evaluation of the tradeoffs and domain knowledge.

Thank you!