Threats to validity

- For experiment results to be trustworthy, they need to be reliable, have high internal validity, and high external validity across time.
- Reliability is about reproducibility of our results and the extent to which a rerun of the exact same experiment would lead to the same result. It's about the stability of measurements and the replicability of the work.
- Internal validity is about the extent to which the treatment effect estimate of our experiment reflects the cause and effect relationship of interest instead of being influenced by other factors. This is all about the design of the experiment.
- External validity is about the extent to which our experiment results generalise both across different populations and across time. Results don't usually generalise across populations (a feature may work well in one country but not in another), which isn't a problem because we can (and usually should!) just run the experiment separately for different markets. Traditional experiments do, however, assume that experiments generalise over time, in that we think that the effect we estimate during the experiment period will be persistent – i.e. we think that the long-term effect of a change will be the same as the short-term effect.
- In my mind, some of the threats listed below, such as learning effects, could be listed as threats to either internal or external validity. I list them where I think they make most sense given the definitions, but one could argue about that and some authors make different choices – what matters is that you think about them when designing an experiment!

Threats to reliability

Threats to internal validity

Interference

- Basically, all violations to SUTVA
- Interference can happen due to
  - Network effects
  - Cannibalisation of resources in marketplaces
  - Shared resources (i.e. treatment slowing down site for everyone)

Interaction effects

- Users can be simultaneously part of multiple experiments, so that what we measure for reach of them is really the effect of the interaction of all of them. This means that, if only some features are implemented, the results after roll out could be different from those observed during the experiment period.
- However, with large sample sizes, this should not generally be a problem because effects of different experiments average out between treatment and control group.
- While the above may be true statistically, interaction effects can still lead to extremely poor user experiences (blue background interacted with blue font), which is why mature platforms aim to avoid them.

Non-representative users

Possible scenarios:

- Our marketing department launches an add campaign and attracts a lot of unusual users to the site temporarily.
- A competitor does the same and temporarily takes a ways users from our site.
- Heavy-user bias: heavy users are more likely to be bucketed in an experiment, biasing the results relative to the overall effect of a feature. Depending on the context, this can be an issue.
- Solution: run experiments for longer (thought this comes with opportunity costs, and will increase cookie churn)

Survivor bias

- This is really just a version of the above: if you select only users that have used the product for some time, your sample is not representative of all users. The classic demonstration of survivor bias is Abraham Wald's insight in WWII that you want to put extra armour where returning plans got hit the least, since it's presumable the planes that got hit there that didn't make it back.

Novelty and learning effects

- Challenge: behaviour might change abruptly and temporarily in response to a new feature (novelty or "burn in" effect) or it might take a while for behaviour fully adapt to a new feature (learning effects). In both cases, the results from a relatively short experiment will not provide a representative picture of the long-run effects of a feature.
- Examples: Increasing number of adds shows on Google led to increase in add revenue initially but then decrease of clicks in the long term because it increased add blindness @hohnhold2015focusing
- Solutions:
  - Measure long-term effects (by running experiments for longer)
  - Have a "holdout" group of users that isn't exposed to any changes for a pre-set period of time (a month, a quarter), to measure long-term effects
  - Estimate dynamic treatment effects to see the evolution of the treatment effect

Sample ration mismatch

Treats to external validity

Budget effects in Ads

- On an adds platform, a treatment might perform very well during an experiment in that it makes marketers launch more adds. But once scaled up may do less well because the increased traffic might exhaust marketer's budgets, leading them to reduce adds launched.

Feedback loops from personalisation

- Treatments might behave differently during experimentation and once they are scaled up if the performance of a feature is a function of the size of the audience it is exposed to (an example could be a recommendation algorithm, which performs better and better as it is being used more).

Day-of-week effects

- See below

Seasonality

- Seasonality comes in many forms: day of week effects, week of year effects, season effects, holiday effects, etc.
- The challenge is that, potentially, user behaviour might differ on certain days or over certain time periods either because we get different users or because users change their behaviour.
- Whether it is really a problem depends on the context. One aspect that is often forgotten here is that seasonality, first and foremost, is about a shift in levels – activity on LinkedIn might go down during the summer months. What we usually want to measure, however, is the difference between treatment and control units. Hence, if you don't have reason to believe that the effect of the treatment is different during a particular season (e.g. because you think it's additive), then seasonality might not be a problem for you.
- Having said that, it's actually quite likely that with either different users or different behaviour by the same users, users might react differently to featore on different days. So it really is a thread to external validity, and we thus should usually care about it.
- Solution: design your experiment so as to take seasonality into account. E.g. run your experiment for at least one week to account for day of week effects (that's generally a good idea), don't run crucial experiments during the holiday season or on major holidays or discard data from such periods, etc.
- What to take into account depends on your context. So understand the relevant seasonality for you (if you're a travel app, consider seasonality of travel demand, if you're an e-commerce site, consider seasonality of shopping behaviour)

Differences in time-to-action between users

- Some users may engage with a new features immediately, others might take a while and then react differently to it.
- When running experiments for a very short time, we might thus get a biased picture of the overall effect of the feature.

Resources

- Forbes article on when not to trust your A/B tests
- Dennis Meisner discussing threats to external validity