The Jungle of Science in CRO and Experimentation

Rigour, evidence, and how we should use science in our work

First, a question…

Who thinks we should aim for the highest rigour in our experiments?

Well…if you put your hand up, I disagree with you

Can I give you an example from scientific research we *might* be able to agree on?

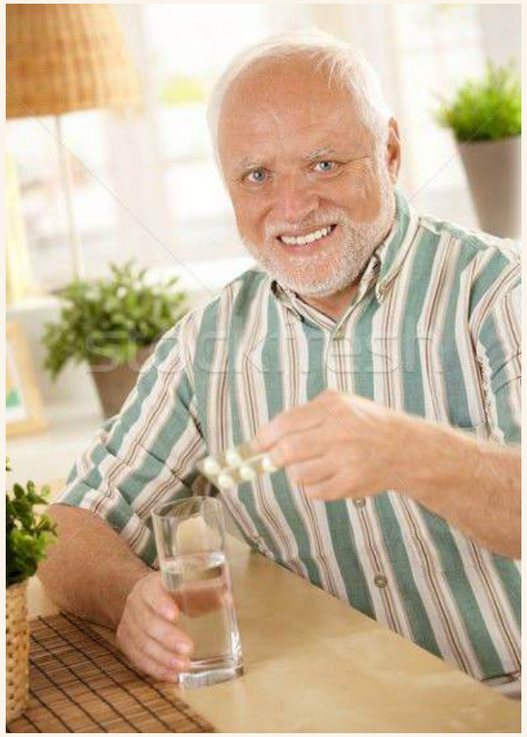
Scientific Trials

Explanatory Trials

Can this work?

Pragmatic Trials

Does this work?

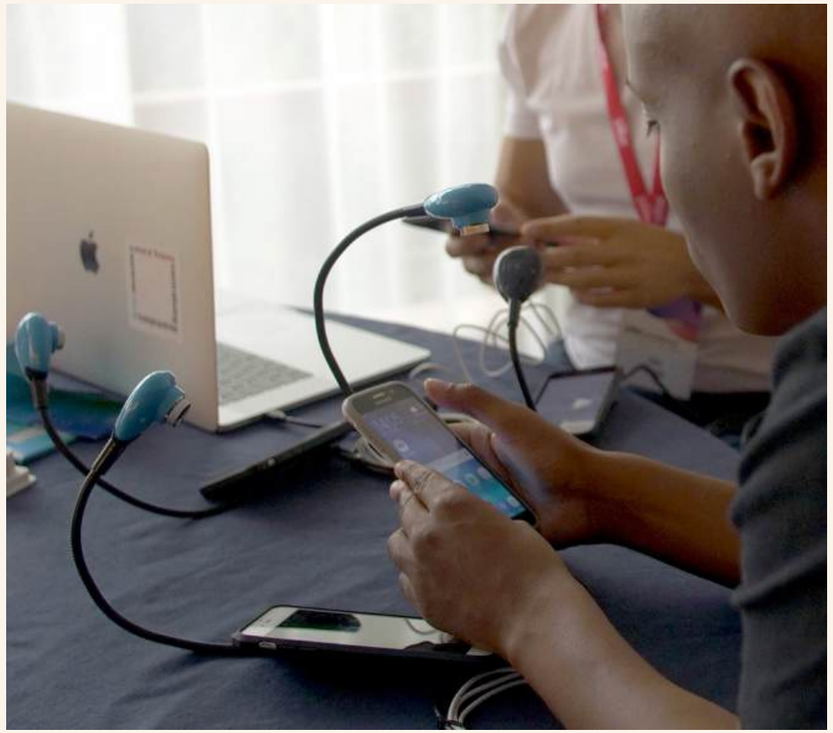


Medicine: Ideal conditions or real world context

Experimentation Methods

Explanatory Method

Can this work?



Pragmatic Method

Does this work?

A/B Testing: Real-world context, not ideal conditions.

We know users experience our tests in a real world context

But what context do we **build them** in?

CRO's Environment Is Not a Laboratory

Scientific Ideal

- Perfect randomisation
- Fixed, stable population
- Isolated variables
- Perfect instrumentation
- Stability of research
- Desired runtime

CRO Reality

- Cookie loss, device switching, bots
- Traffic cycles, campaigns, competitors
- Multiple variables, product releases
- Noisy data, missing tracking, analytics flaws
- Reactive stakeholders
- Limited time windows, deadlines

CRO Borrows from Science… and Then Gets Lost

What We Borrow

Control groups, randomisation, significance

Why It Helps

Reduces bias, structures decisions

Where It Breaks

Mimicking lab rigour in uncontrolled contexts

- CRO in real businesses: deadlines, politics, messy organisations and data
- Lab discipline ignores these realities
- Undervalues human organisational work

CRO Borrows from Science… and Then Gets Lost

What We Borrow

Control groups, randomisation, significance

Why It Helps

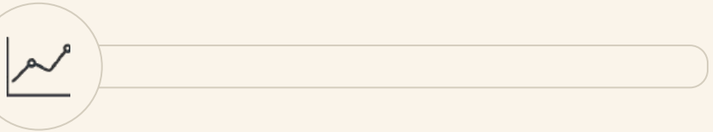
Reduces bias, structures decisions

Where It Breaks

Mimicking lab rigour in uncontrolled contexts

- CRO in real businesses: deadlines, politics, messy organisations and data
- Lab discipline ignores these realities
- Undervalues human organisational work

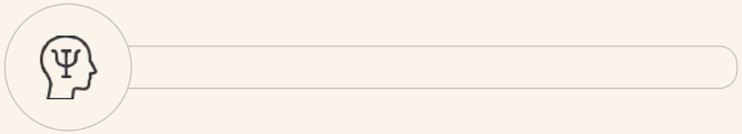
Rigour In Different Disciplines



CRO

Primary risk: Financial loss, sub-optimal performance.

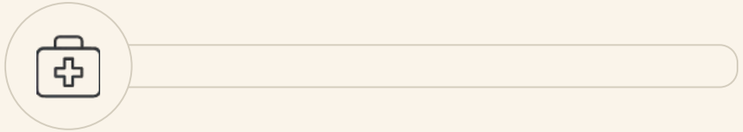
Dynamic rigour appropriate.



Psychology

Ethical considerations, human impact.

Significant rigour needed.



Medicine

Human lives at stake.

Highest rigour needed.

Even within science, rigour varies by purpose.

However, rigour still varies within disciplines

High Stakes Psychology

Childhood trauma research: high ethical concern for the participants.

High Stakes in CRO

Government website tools: can impact vulnerable people.

Low Stakes Psychology

Academic pricing studies: limited ethical impact for the participants.

Low Stakes in CRO

Button colour testing: generally low impact on users.

Core principle: Rigour must always match the stakes in any discipline

What Are We Trying To Do?

CRO **Isn't** Trying To

- Discover universal laws
- Publish academically
- Achieve scientific purity

CRO **Is** Trying To

- Support better decisions
- Build practical understanding
- Learn what works in *this* context
- Move forward without causing harm

What do we learn along the way?

Experimenters accumulate insights for a "practical meta-analysis"

Audience Behaviour

How audiences respond

Pattern Recognition

Consistent UX patterns, friction

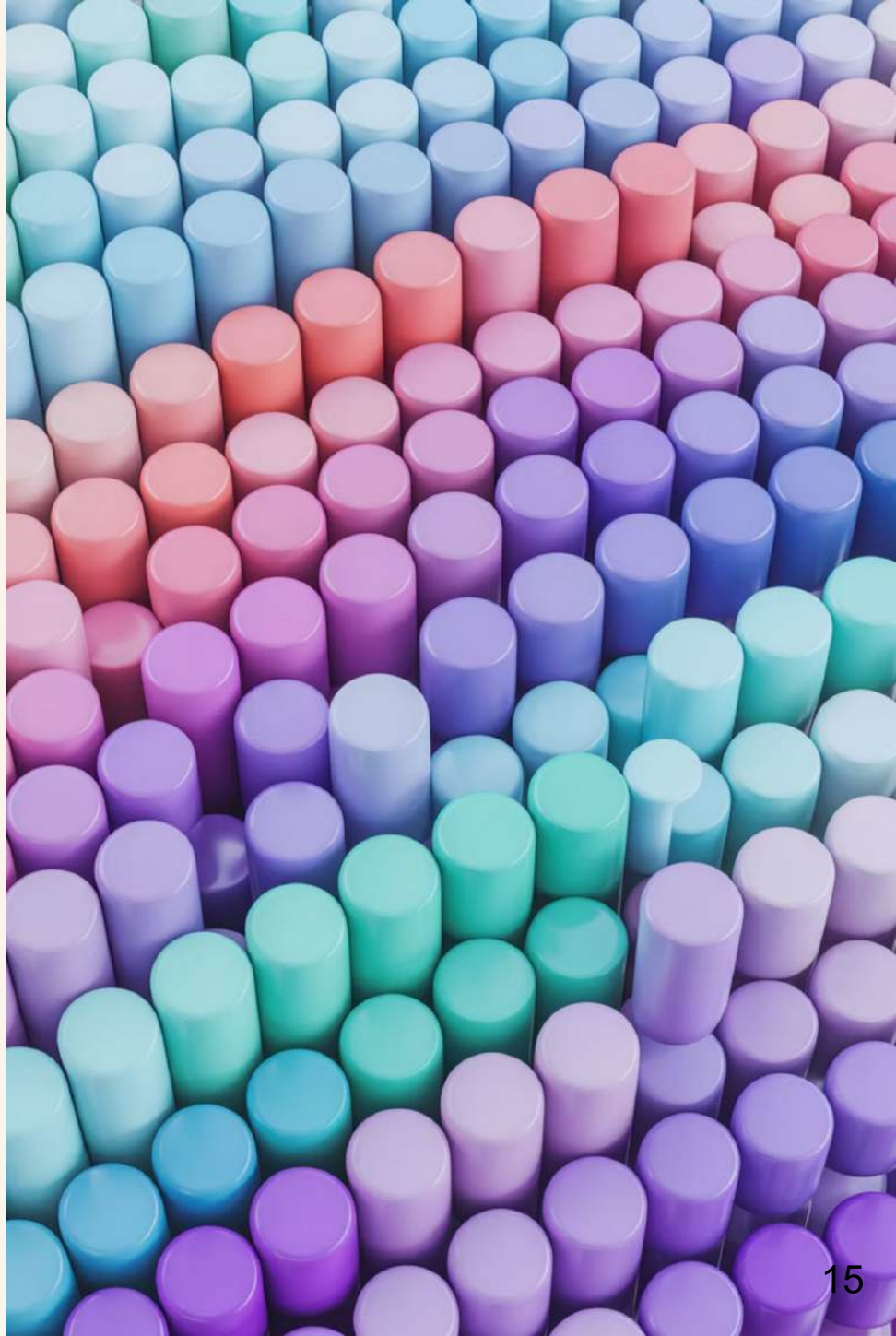
Contextual Learning

Business and product

Organisational Reality

Feasibility, cost, political barriers

This context is crucial for deciding **what to test**



But…could the wrong approach to rigour mean that you accumulate less of this understanding?

The Two Ways Teams Get Rigour Wrong

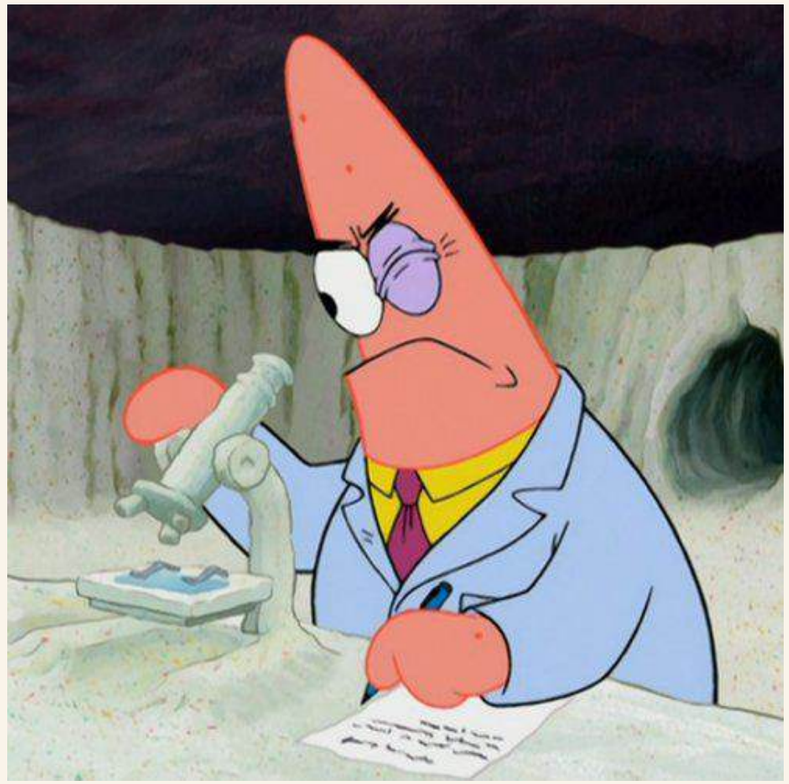
Trap 1: Oversimplification

- Just run loads of quick tests and see what happens
- Bad statistics
- Lack of well thought out process
- Volume over quality
- No genuine learning

The Two Ways Teams Get Rigour Wrong

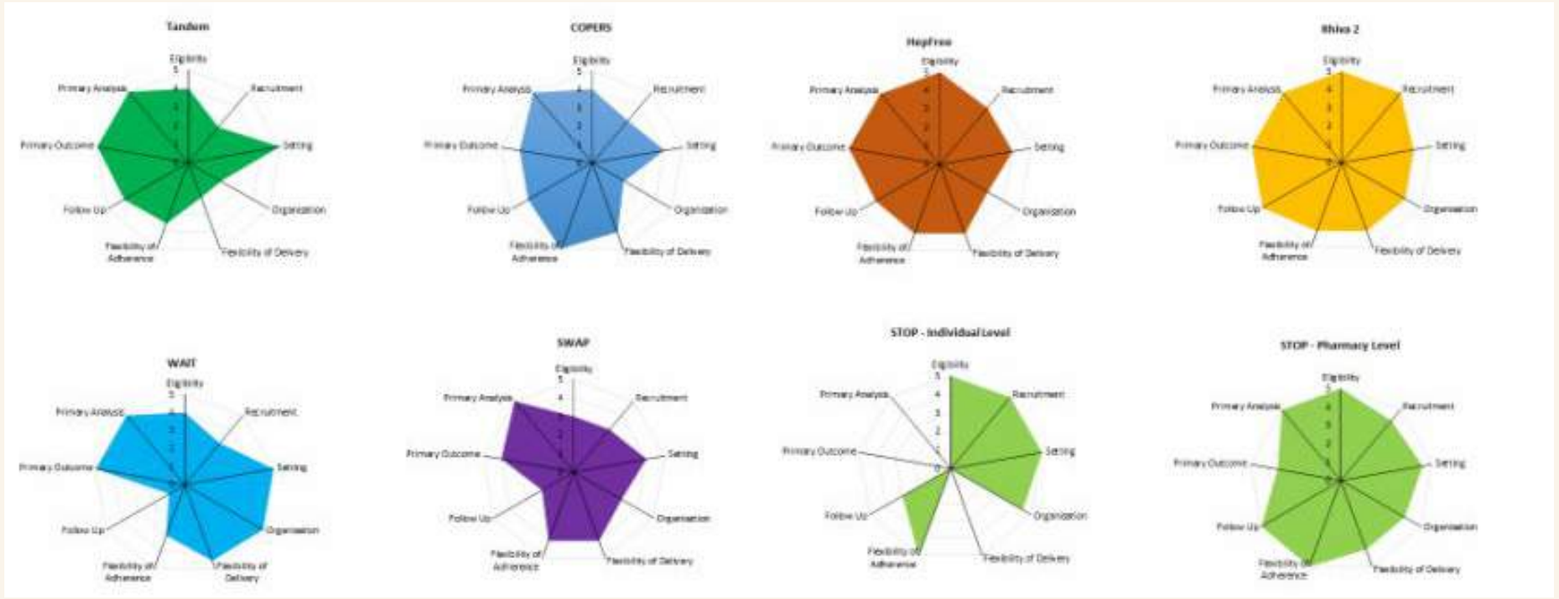
Trap 2: Overcomplication

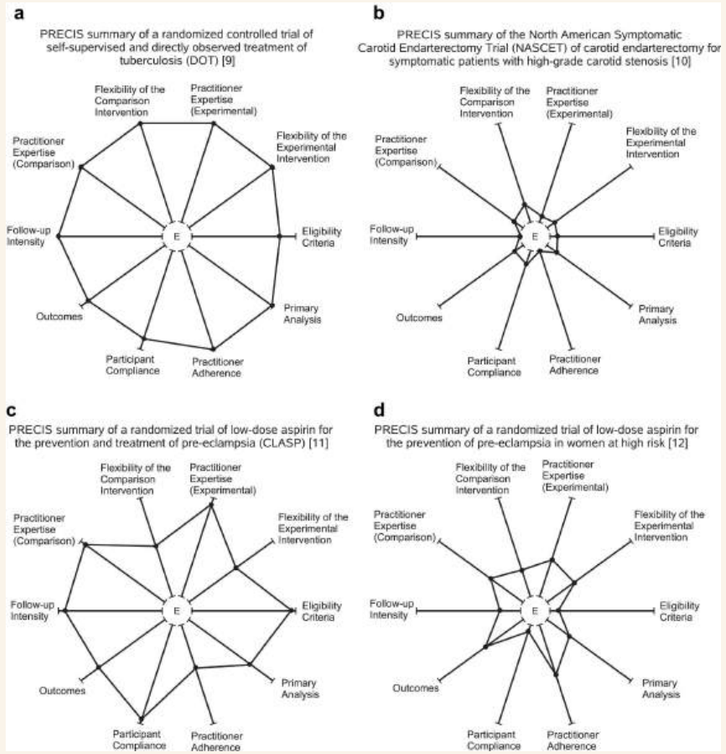
- Too many people in the decision chain
- Delays in starting and calling tests
- Fear of being wrong
- Shipping slows down
- Decision paralysis!



So How Do They Get It Right?

It's not this…





PRECIS-2: Look it up if you have ***nothing*** better to do

Or this…

Before I go into how I think we should approach rigour, a caveat

Basic Hygiene, Not Purity

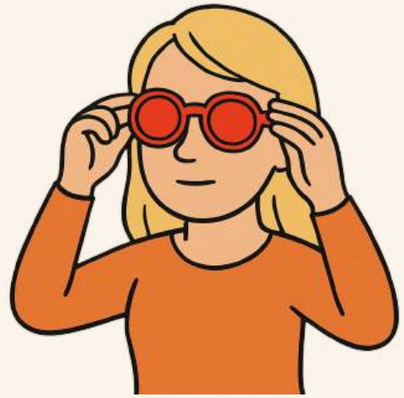
Minimum Viable Process

- Quality assurance
- Accurate tracking & instrumentation
- Basic statistical literacy
- Guardrails against harm
- Writing hypotheses
- Data quality checks

Uphold hygiene so your rigour can be proportional and pragmatic

What should you consider when varying rigour?

5 Lenses of Rigour



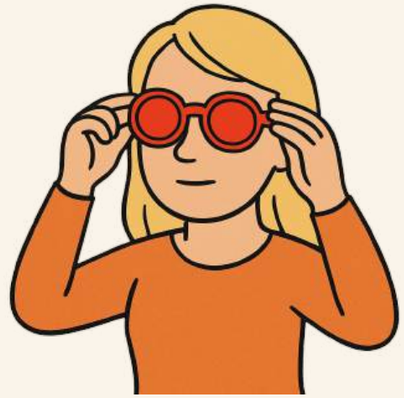
Stakes

- Rigour should rise with the consequences of being wrong
- If the decision is high-stakes, you need stronger evidence and more confidence
- If the stakes are low, don’t over-engineer it as this slows the accumulation of learnings

What is the consequence if this test misleads us and we make a decision based on the wrong signal?

5 Lenses of Rigour

Stakes



*“In practice, **each A/B test should be planned by taking into account the probabilistic risks and rewards involved.***

*Such an approach results in **a balance between type I and type II errors** which maximizes the return on investment from testing.*

If a threshold of 95% seems too high for some tests, it may well be so. For other, high-stakes scenarios, I’d argue that even 99% might be way too low a bar.”

- **Georgi Georgiev**, Top Misconceptions About Scientific Rigor in A/B Testing

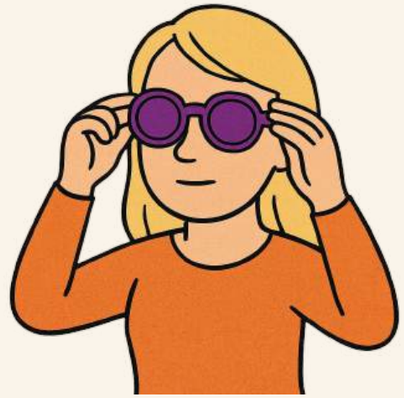
5 Lenses of Rigour

Learning

- The goal in CRO isn’t scientific purity
- The goal is learning that moves us forward
- The best CRO programmes work because they compound insight

What do I need to learn here, and what outcome gives me that learning most effectively?

5 Lenses of Rigour

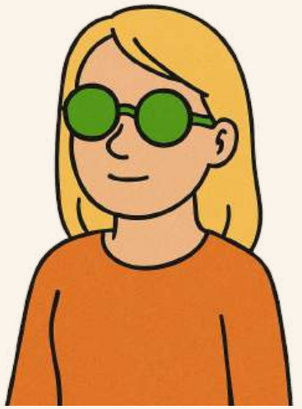


Detectability

- This lens is about what the test can realistically show given traffic, effect size, and context
- If the stakes are low, a weaker signal can still give useful directional learning
- If the stakes are high, low signal means we need a different design, metric, or question

Given the test setup, what can we realistically learn and is that enough for the decision we need to make?

5 Lenses of Rigour



Capability

- This is about whether the team can safely and consistently use the process
- If the capability isn’t there, advanced techniques reduce rigour by creating confusion, mistrust, and inconsistent interpretation
- True rigour is choosing methods the team can run well, not the ones that look scientifically impressive

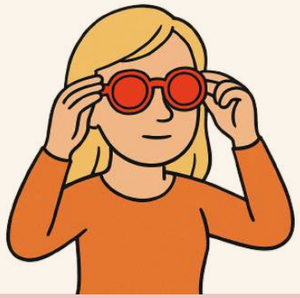
What level of rigour can this team execute well and consistently?

5 Lenses of Rigour

Confidence

- The barrier to action isn't always the data itself
- It’s whether people trust the process enough to make a decision from it
- This is about the everyday political realities of using tests to challenge ego, assumptions and cherished notions

Is this enough rigour for people to believe the result, make a decision, and not argue about it endlessly?

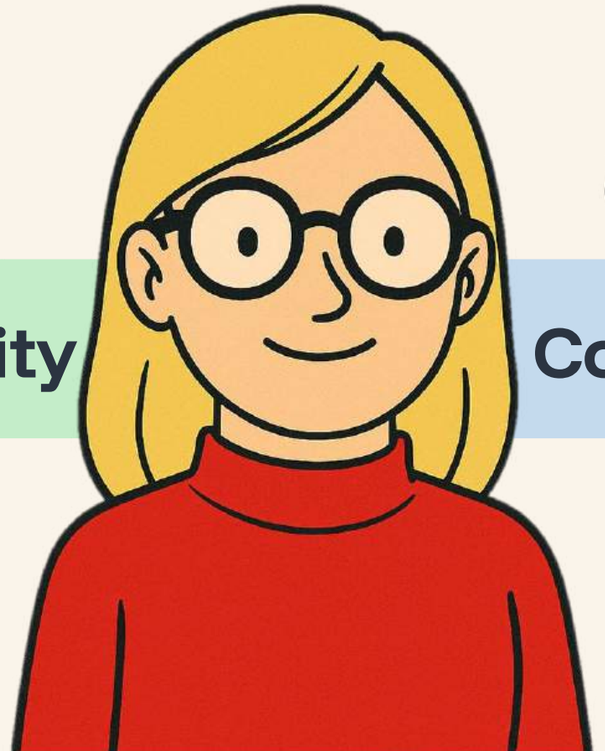


Stakes

Learning

Detectability

Capability



Confidence

Human Factors Are Part of Every Test

- [ ] Panic stops after peeking
- [ ] Winners never shipped
- [ ] Conflicting priorities
- [ ] Time pressure
- [ ] Fear of failure

In summary

CRO is not a lab experiment

We should aim to:

- Make better decisions under uncertainty
- Build practical understanding over time
- Move forward without causing harm

It’s a jungle out there!

- Use enough rigour to make confident decisions, not so much that you slow learning
- Science gives us brilliant methods but appreciate your context
- If your tests are helping make better product decisions, more often...
- You're doing a great job!

Thank you!

LinkedIn: bit.ly/Marcellasullivan