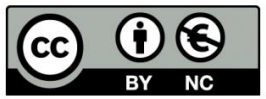
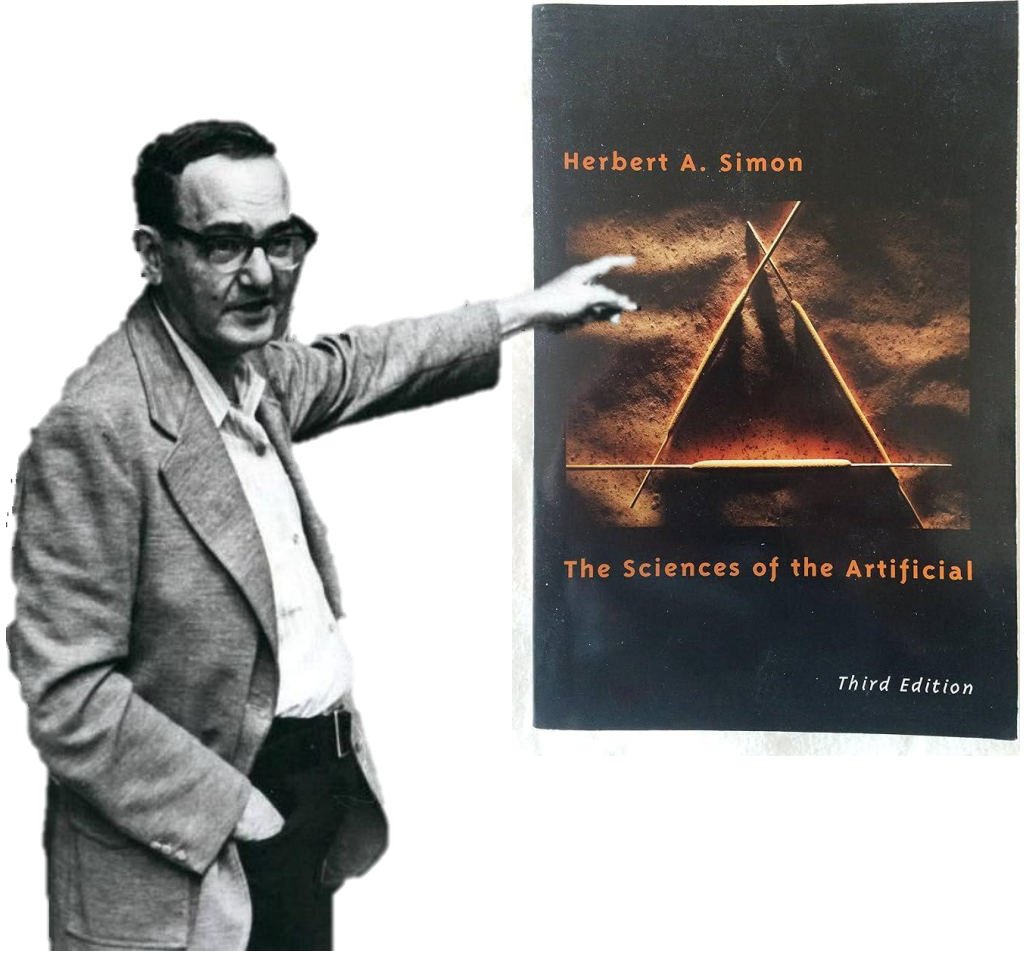
Design of Experiments
Empirical IT Research with humans

Prof. Federico Cabitza
A.Y. 2025-2026





‘design sciences’ approach

Scientific Method	Objective
1. Observations	Gather empirical data
2. Literature Review	Understand the existing theories
3. Hypothesis Formulation	Develop a testable proposition
4. Experiment Design	Plan how to test the hypothesis
5. Data Collection	Perform experiments to collect data
6. Analysis	Interpret the data
7. Hypothesis Testing	Statistical analysis to reject/confirm hypothesis
8. Interpretation	Final interpretation of the results

‘design sciences’ approach

a science of design
and design as a science

Scientific Method	Objective	Design Science	Objective
1. Observations	Gather empirical data	**1. Problem Identification**	Understand the user needs and constraints
2. Literature Review	Understand the existing theories	**2. Literature & Benchmark Review**	Study existing solutions and design principles
3. Hypothesis Formulation	Develop a testable proposition	**3. Objective Setting**	Formulate design goals and hypotheses regarding the artifact's value
4. Experiment Design	Plan how to test the hypothesis	**4. Design & Development**	Create artifact prototypes using informed methods
5. Data Collection	Perform experiments to collect data	**5. Artifact Evaluation**	Employ user testing, analytics, and other methodologies to evaluate the artifact's performance
6. Analysis	Interpret the data	**6. Analysis**	Synthesize user feedback and performance metrics
7. Hypothesis Testing	Statistical analysis to reject/confirm hypothesis	**7. Hypothesis Testing**	Determine if the design goals and value propositions are met, often via statistical methods
8. Interpretation	Final interpretation of the results	**8. Delivery**	Summarize design efficacy and areas for improvement

‘design sciences’ approach

a science of design
and design as a science

Scientific Method	Objective	Design Science	Objective
1. Observations	Gather empirical data	**1. Problem Identification**	Understand the user needs and constraints
2. Literature Review	Understand the existing theories	**2. Literature & Benchmark Review**	Study existing solutions and design principles
3. Hypothesis Formulation	Develop a testable proposition	**3. Objective Setting**	Formulate design goals and hypotheses regarding the artifact's value
4. Experiment Design	Plan how to test the hypothesis	**4. Design & Development**	Create artifact prototypes using informed methods
5. Data Collection	Perform experiments to collect data	**5. Artifact Evaluation**	Employ user testing, analytics, and other methodologies to evaluate the artifact's performance
6. Analysis	Interpret the data	**6. Analysis**	Synthesize user feedback and performance metrics
7. Hypothesis Testing	Statistical analysis to reject/confirm hypothesis	**7. Hypothesis Testing**	Determine if the design goals and value propositions are met, often via statistical methods
8. Interpretation	Final interpretation of the results	**8. Delivery**	Summarize design efficacy and areas for improvement

Scientific Method	Objective	Design Science	Objective
1. Observations	Gather empirical data	**1. Problem Identification**	Understand the user needs and constraints
2. Literature Review	Understand the existing theories	**2. Literature & Benchmark Review**	Study existing solutions and design principles
3. Hypothesis Formulation	Develop a testable proposition	**3. Objective Setting**	Formulate design goals and hypotheses regarding the artifact's value
4. Experiment Design	Plan how to test the hypothesis	**4. Design & Development**	Create artifact prototypes using informed methods
5. Data Collection	Perform experiments to collect data	**5. Artifact Evaluation**	Employ user testing, analytics, and other methodologies to evaluate the artifact's performance
6. Analysis	Interpret the data	**6. Analysis**	Synthesize user feedback and performance metrics
7. Hypothesis Testing	Statistical analysis to reject/confirm hypothesis	**7. Hypothesis Testing**	Determine if the design goals and value propositions are met, often via statistical methods
8. Interpretation	Final interpretation of the results	**8. Delivery**	Summarize design efficacy and areas for improvement

HYPOTHESIS ⟷ **PROTOTYPE**

1.Def: A hypothesis in the scientific method is a testable statement that posits a relationship between variables based on empirical evidence.
2.Testability: Hypotheses are subject to falsification through experiments, which collect data to either confirm or reject the hypothesis.
3.Iteration: If the hypothesis is rejected, the scientist revises the hypothesis or develops a new one, returning to experimental testing. The cycle continues until sufficient evidence is gathered to support or reject the hypothesis.

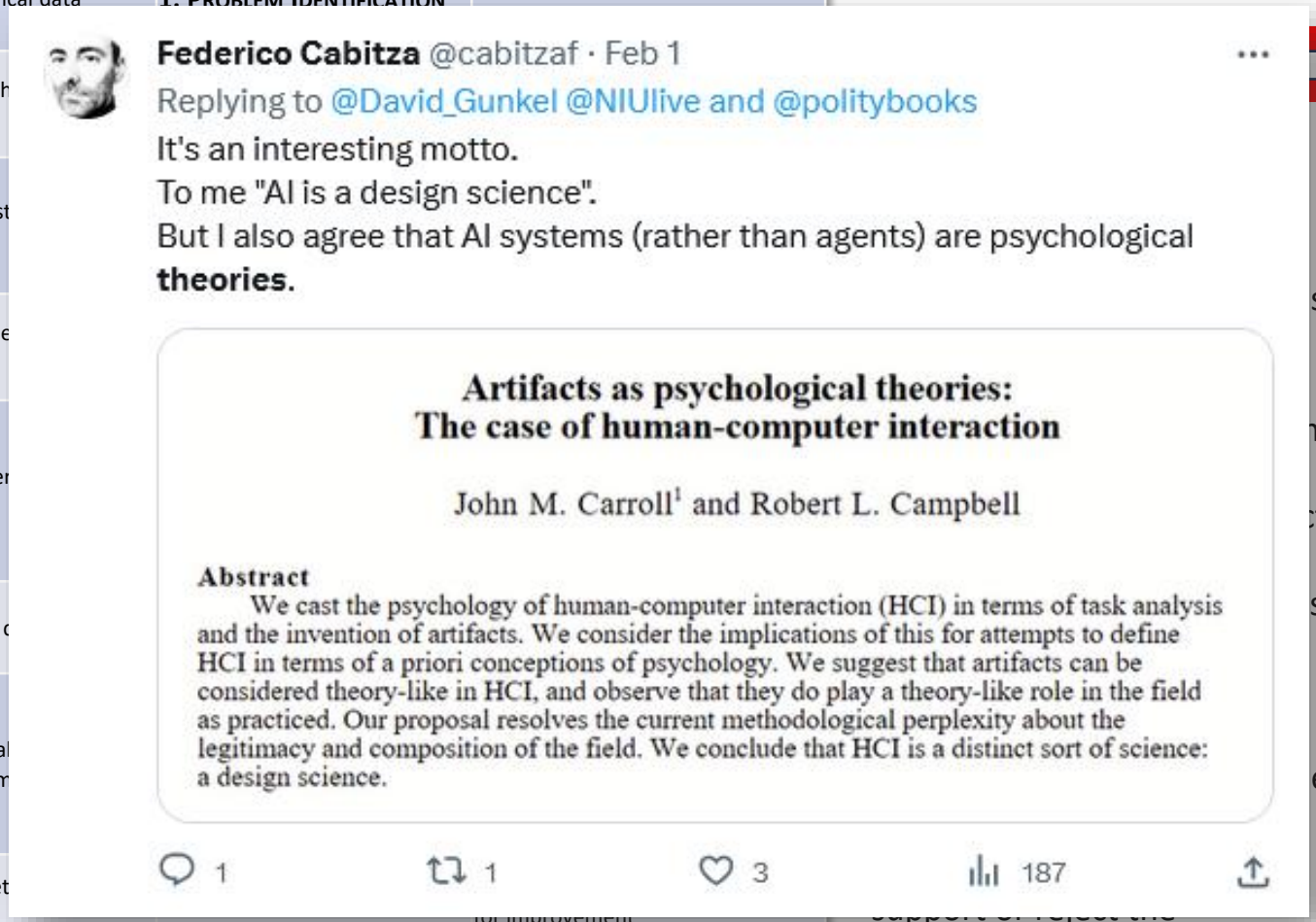
Scientific Method	Objective	Design Science	Objective
1. Observations	Gather empirical data	**1. Problem Identification**	Understand the user needs and constraints
2. Literature Review	Understand the existing theories	**2. Literature & Benchmark Review**	Study existing solutions and design principles
3. Hypothesis Formulation	Develop a testable proposition	**3. Objective Setting**	Formulate design goals and hypotheses regarding the artifact's value
4. Experiment Design	Plan how to test the hypothesis	**4. Design & Development**	Create artifact prototypes using informed methods
5. Data Collection	Perform experiments to collect data	**5. Artifact Evaluation**	Employ user testing, analytics, and other methodologies to evaluate the artifact's performance
6. Analysis	Interpret the data	**6. Analysis**	Synthesize user feedback and performance metrics
7. Hypothesis Testing	Statistical analysis to reject/confirm hypothesis	**7. Hypothesis Testing**	Determine if the design goals and value propositions are met, often via statistical methods
8. Interpretation	Final interpretation of the results	**8. Delivery**	Summarize design efficacy and areas for improvement

HYPOTHESIS ⟷ PROTOTYPE

1.Def: A hypothesis in the scientific method is a testable statement that posits a relationship between variables based on empirical evidence.
2.Testability: Hypotheses are subject to falsification through experiments, which collect data to either confirm or reject the hypothesis.
3.Iteration: If the hypothesis is rejected, the scientist revises the hypothesis or develops a new one, returning to experimental testing. The cycle continues until sufficient evidence is gathered to support or reject the hypothesis.

1.Def: each prototype acts as an "embodied hypothesis." It is a tangible or virtual manifestation of a solution that aims to fulfill specific user needs and requirements.
2.Testability: The effectiveness of the prototype—i.e., its capacity to meet users' needs and improve their experience—is assessed through various forms of evaluations like user testing, A/B testing, and analytics.
3.Iteration: if the prototype does not meet the design, it is either revised or replaced with another prototype (another "hypothesis"). The design cycle continues with new evaluations.

Scientific Method	Objective	Design Science	Objective
1. Observations	Gather empirical data	1. Problem Identification	Understand the user needs and
2. Literature Review	Understand th theories		
3. Hypothesis Formulation	Develop a test proposition		
4. Experiment Design	Plan how to te hypothesis		
5. Data Collection	Perform exper collect data		
6. Analysis	Interpret the d		
7. Hypothesis Testing	Statistical anal reject/confirm		
8. Interpretation	Final interpret results		



hypothesis.

PROTOTYPE

1.Def: In the context of design science, each prototype acts as an "embodied hypothesis." It is a tangible or virtual manifestation of a solution that aims to fulfill specific user needs and requirements.

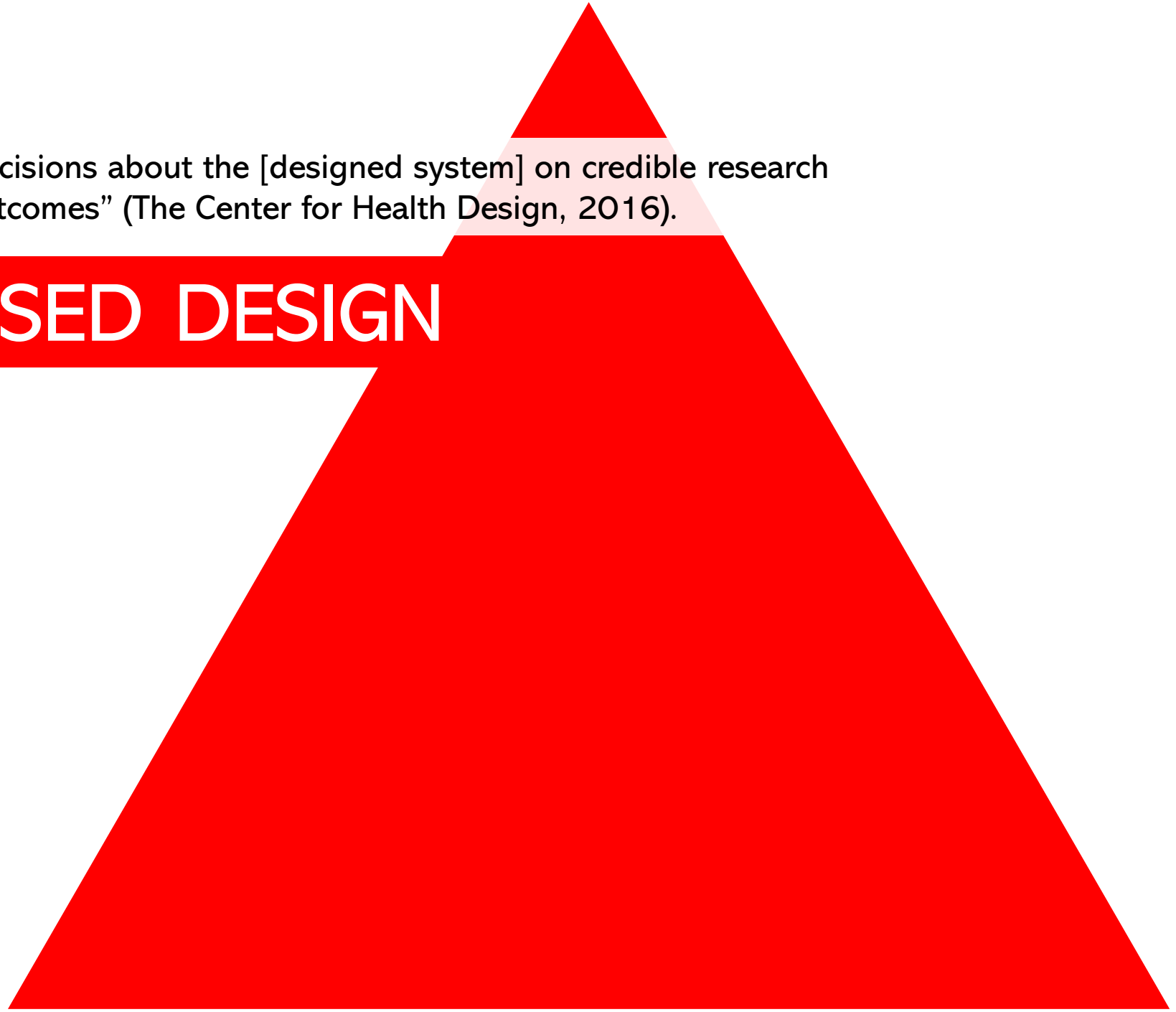
2.Testability: The effectiveness of the prototype—i.e., its capacity to meet users' needs and improve their experience—is assessed through various forms of evaluations like user testing, A/B testing, and analytics.

3.Iteration: if the prototype does not meet the design, it is either revised or replaced with another prototype (another "hypothesis"). The design cycle continues with new evaluations.

EVIDENCE-BASED DESIGN

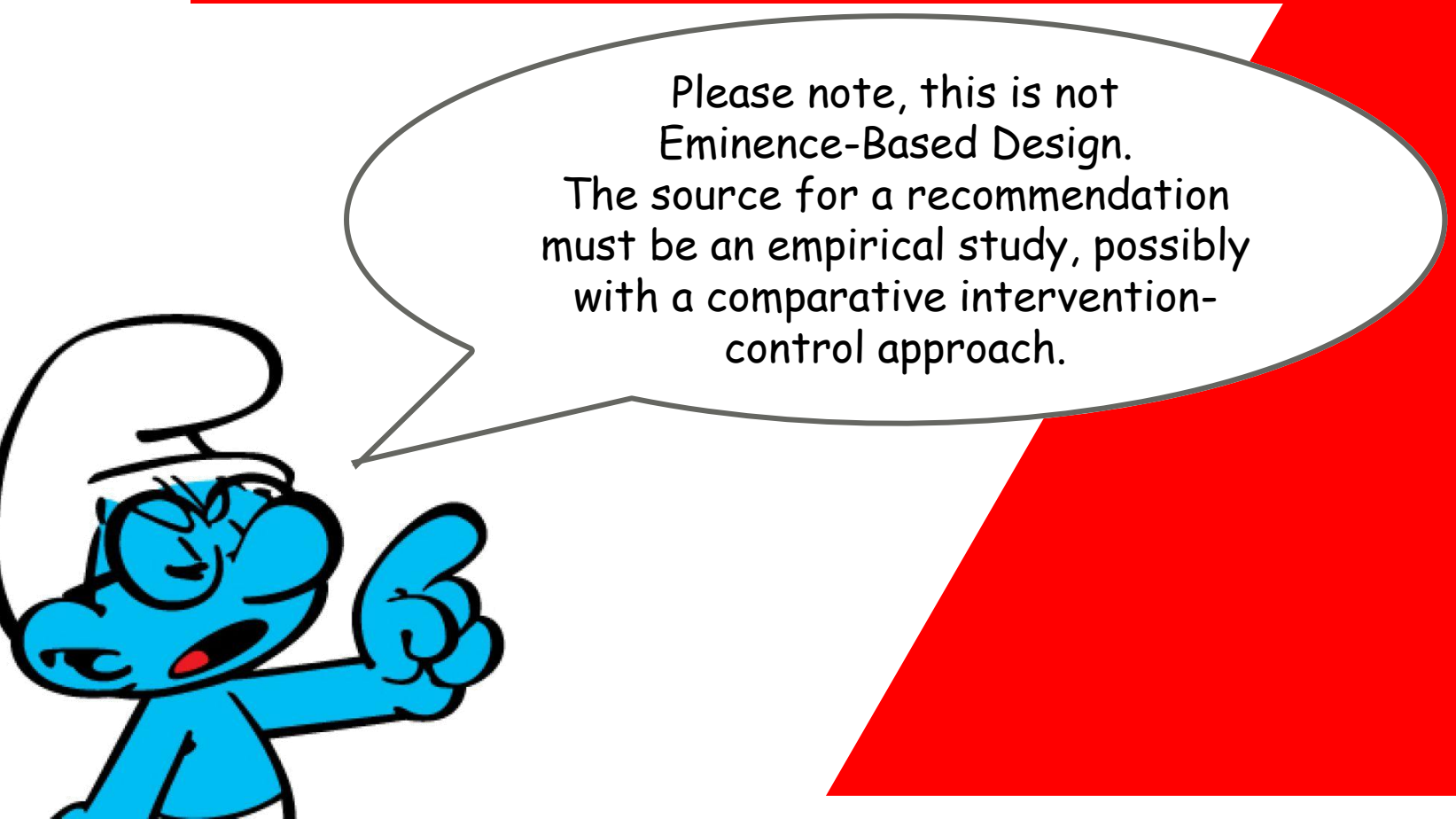
- EBD: “the process of basing decisions about the [designed system] on credible research to achieve the best possible outcomes” (The Center for Health Design, 2016).

EVIDENCE-BASED DESIGN



- EBD: “the process of basing decisions about the [designed system] on credible research to achieve the best possible outcomes” (The Center for Health Design, 2016).

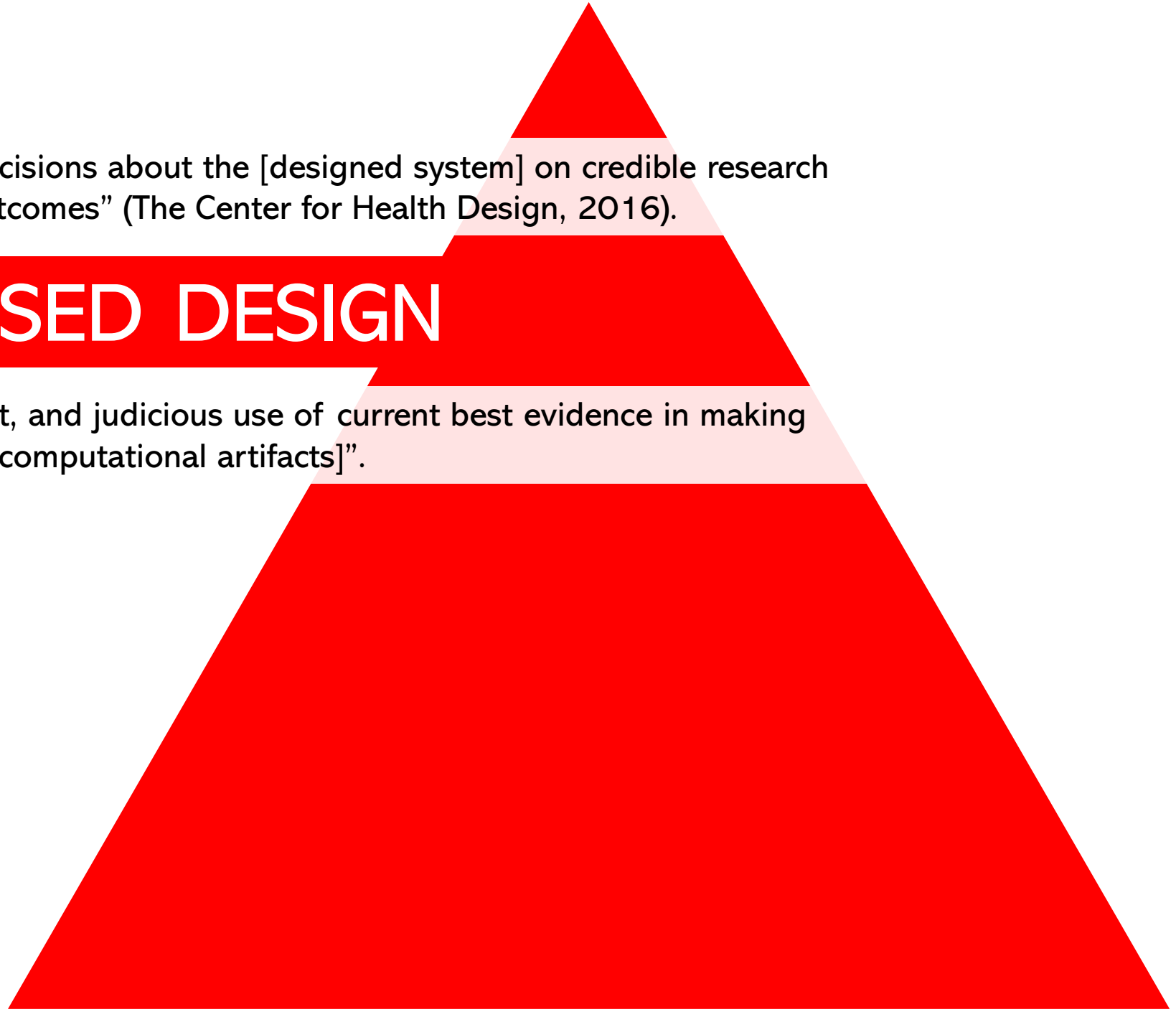
EVIDENCE-BASED DESIGN



EVIDENCE-BASED DESIGN

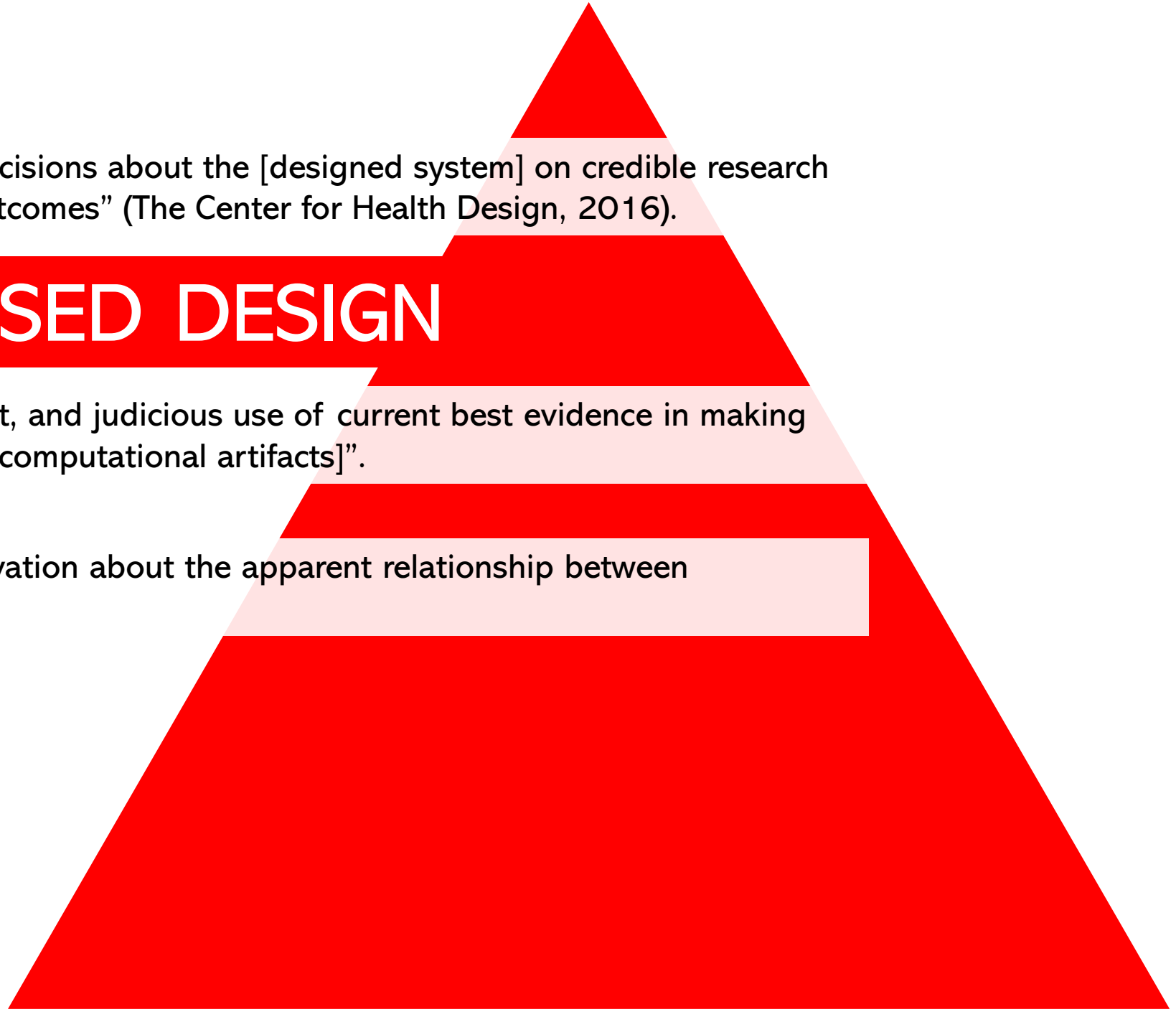
- EBD: “the process of basing decisions about the [designed system] on credible research to achieve the best possible outcomes” (The Center for Health Design, 2016).

- EBD: “the conscientious, explicit, and judicious use of current best evidence in making decisions about the [design of computational artifacts]”.



EVIDENCE-BASED DESIGN

- EBD: “the process of basing decisions about the [designed system] on credible research to achieve the best possible outcomes” (The Center for Health Design, 2016).
- EBD: “the conscientious, explicit, and judicious use of current best evidence in making decisions about the [design of computational artifacts]”.
- Evidence: “any empirical observation about the apparent relationship between events”/phenomena.



- EBD: “the process of basing decisions about the [designed system] on credible research to achieve the best possible outcomes” (The Center for Health Design, 2016).

EVIDENCE-BASED DESIGN

- EBD: “the conscientious, explicit, and judicious use of current best evidence in making decisions about the [design of computational artifacts]”.

- Evidence: “any empirical observation about the apparent relationship between events”/phenomena.

- EBD advocates a balanced integration of the skills and experience of the design practitioner, the client’s needs, and critically assessed evidence of various types.

- These include evidence grounded in rigorous scientific methodology as well as a continuum of levels of evidence including personal experience and intuition.

(strongest)

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

EVIDENCE-BASED DESIGN

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
Level 3: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving prospective real-world cases considered by real practitioners in real-world settings.
Level 4: Single experimental study (randomized, controlled) with retrospective real-world cases considered by real practitioners in simulated/laboratory settings.
Level 5: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving retrospective real-world cases considered by real practitioners in simulated/laboratory settings
Level 6: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by real practitioners in simulated/ laboratory settings.
Level 7: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by human participants but not real practitioners in laboratory settings
Level 8: Supervised machine learning train/test studies with external validation (multiple datasets in longitudinal or cross-section/multi-site settings)
Level 9 Supervised machine learning train/test studies with internal validation
Level 10: Consensus opinions of authoritative bodies (e.g., nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 11 (weakest): Opinions of recognized experts and case studies

(weakest)

(strongest)

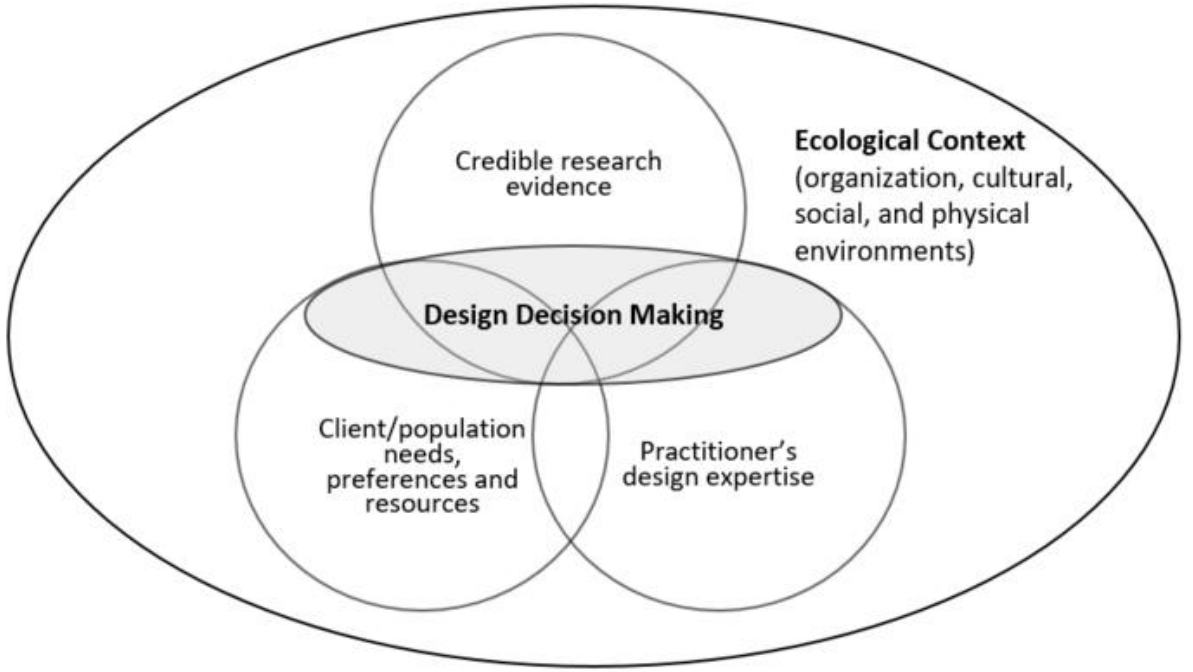
LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

EVIDENCE-BASED DESIGN

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
Level 3: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving prospective real-world cases considered by real practitioners in real-world settings.
Level 4: Single experimental study (randomized, controlled) with retrospective real-world cases considered by real practitioners in simulated/laboratory settings.
Level 5: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving retrospective real-world cases considered by real practitioners in simulated/laboratory settings
Level 6: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by real practitioners in simulated/ laboratory settings.
Level 7: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by human participants but not real practitioners in laboratory settings
Level 8: Supervised machine learning train/test studies with external validation (multiple datasets in longitudinal or cross-section/multi-site settings)
Level 9 Supervised machine learning train/test studies with internal validation
Level 10: Consensus opinions of authoritative bodies (e.g., nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 11 (weakest): Opinions of recognized experts and case studies

(weakest)

EVIDENCE-BASED DESIGN



Peavey, E., & Vander Wyst, K. B. (2017). Evidence-based design and research-informed design: What's the difference? Conceptual definitions and comparative analysis. *HERD: Health Environments Research & Design Journal, 10*(5), 143-156

EVIDENCE-BASED DESIGN

About **what** do we have to find or collect evidence (to base our design on)?

EVIDENCE-BASED DESIGN

About **what** do we have to find or collect evidence (to base our design on)?

About the design solutions that are associated (or cause) the best effect.

EVIDENCE-BASED DESIGN

About **what** do we have to find or collect evidence (to base our design on)?

About the design solutions that are associated (or cause) the best effect.
About the design solutions that make the technological intervention adopting them have the greater desired effect on the context of use.

EVIDENCE-BASED DESIGN

About **what** do we have to find or collect evidence (to base our design on)?

About the design solutions that are associated (or cause) the best effect.
About the design solutions that make the technological intervention adopting them have the greater desired effect on the context of use.

(AI/system) intervention: an intervention (which relies on an AI system/component/technology) to serve its purpose (cf. CONSORT-AI)

https://www.nature.com/articles/s41591-020-1034-x

EVIDENCE-BASED DESIGN

About **what** do we have to find or collect evidence (to base our design on)?

About the design solutions that are associated (or cause) the best effect.
About the design solutions that make the technological intervention adopting them have the greater desired effect on the context of use.

(AI/system) intervention: an intervention (which relies on an AI system/component/technology) to serve its purpose (cf. CONSORT-AI)

Human-AI/system interaction: The process of how humans interact with the (AI) intervention, for this to *function* as *intended*.

How to Prepare a Research Protocol

A clear, review-ready template for PhD research in HCI / Human-Centred AI

Based on: “Template for a HCI Research Protocol”

What a research protocol is (and why you need it)

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

Why?

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

Why?

Editors and reviewers are increasingly requesting authorization from the ethics committees, Ethics Boards, Institutional Review Boards (IRB) or Research Ethics Boards (REBs) of a research institution in order to publish a paper (often before it has been deemed suitable for publication in all other respects).

However, **all** ethics committees **always** request a **research protocol**, based on models that are all very similar.

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

What a research protocol is (and why you need it)

A protocol is a **precise blueprint** of your study: rationale → design → ethics → execution → analysis.
It is written **before** data collection so others can **understand**, **evaluate**, and **reproduce** your work.

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

What a research protocol is (and why you need it)

A protocol is a precise blueprint of your study: rationale → design → ethics → execution → analysis.
It is written before data collection so others can understand, evaluate, and reproduce your work.

Purpose

- Lock in decisions early (design, measures, analysis).
- Prevent “researcher degrees of freedom”.
- Make your study executable by someone else.

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

What a research protocol is (and why you need it)

A protocol is a precise blueprint of your study: rationale → design → ethics → execution → analysis.
It is written before data collection so others can understand, evaluate, and reproduce your work.

Purpose

- Lock in decisions early (design, measures, analysis).
- Prevent “researcher degrees of freedom”.
- Make your study executable by someone else.

Audience

- Supervisors & thesis committee.
- Ethics / IRB review.
- Future readers (papers, OSF, replication).

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

What a research protocol is (and why you need it)

A protocol is a precise blueprint of your study: rationale → design → ethics → execution → analysis.
It is written before data collection so others can understand, evaluate, and reproduce your work.

Purpose

- Lock in decisions early (design, measures, analysis).
- Prevent “researcher degrees of freedom”.
- Make your study executable by someone else.

Audience

- Supervisors & thesis committee.
- Ethics / IRB review.
- Future readers (papers, OSF, replication).

What reviewers check

- A justified question and measurable aims.
- Ethical, feasible procedure and sampling.
- Analysis plan aligned with variables & RQs.

Tip: if you cannot run the study from the protocol alone, it is not specific enough.

Protocol anatomy (the “map”)

Use this structure to keep your protocol both complete and easy to review.

Front matter

Rationale

Method

Back matter

Protocol anatomy (the “map”)

Use this structure to keep your protocol both complete and easy to review.

Keep sections short, but make methods and analysis operational (ready to execute).

Front matter

- Title page
- Abstract (≤250w)
- Keywords

Rationale

- Introduction
- Prior work
- Justification
- Aims / RQs / Hs

Method

- Design
- Ethics
- Participants
- Materials
- Procedure
- Analysis plan

Back matter

- Anticipated results
- Limitations
- Implications
- References

Protocol anatomy (the “map”)

Use this structure to keep your protocol both complete and easy to review.

Keep sections short, but make methods and analysis operational (ready to execute).

Front matter

- Title page
- Abstract (≤250w)
- Keywords

Rationale

- Introduction
- Prior work
- Justification
- Aims / RQs / Hs

Method

- Design
- Ethics
- Participants
- Materials
- Procedure
- Analysis plan

Back matter

- Anticipated results
- Limitations
- Implications
- References

Abstract & keywords: the “front door”

Abstract (max 250 words)

- Background: what problem and why now?
- Aim: what you will test/understand.
- Method: design, sample, main measures.
- Contribution: why it matters (theory / practice / society).

Keywords (3–6)

- Use specific terms (methods + domain + construct).
- Prefer standard vocabulary used in HCI venues.
- Avoid overly broad words (e.g., “technology”).

Write the abstract last, but design it to be readable in 30 seconds.

Abstract & keywords: the “front door”

Abstract (max 250 words)

- Background: what problem and why now?
- Aim: what you will test/understand.
- Method: design, sample, main measures.
- Contribution: why it matters (theory / practice / society).

Keywords (3–6)

- Use specific terms (methods + domain + construct).
- Prefer standard vocabulary used in HCI venues.
- Avoid overly broad words (e.g., “technology”).

Mini-template

```
This study investigates [phenomenon] in
[context]. We aim to [aim], using a [design]
with N≈[size] [population]. Participants
will [task], and we will measure [DV1, DV2]
and analyse data with [model/test]. Findings
will contribute to [theory/practice/society]
by [contribution].
```

Write the abstract last, but design it to be readable in 30 seconds. DV stands for Dependent Variable

Abstract & keywords: the “front door”

Abstract (max 250 words)

- Background: what problem and why now?
- Aim: what you will test/understand.
- Method: design, sample, main measures.
- Contribution: why it matters (theory / practice / society).

Keywords (3–6)

- Use specific terms (methods + domain + construct).
- Prefer standard vocabulary used in HCI venues.
- Avoid overly broad words (e.g., “technology”).

Mini-template

```
This study investigates [phenomenon] in
[context]. We aim to [aim], using a [design]
with N≈[size] [population]. Participants
will [task], and we will measure [DV1, DV2]
and analyse data with [model/test]. Findings
will contribute to [theory/practice/society]
by [contribution].
```

IV – Independent Variable

The independent variable is the factor you manipulate, vary, or compare across conditions. It represents the presumed cause or explanatory factor in your study. Examples: interface type, presence vs absence of an AI explanation, level of automation, training condition.

DV – Dependent Variable

The dependent variable is what you measure as an outcome. It is expected to change as a function of the independent variable. It represents the presumed effect. Examples: task accuracy, completion time, error rate, trust score, workload, perceived usability.

Introduction: build the rationale

Your introduction should move from context → evidence → gap → contribution.

Theoretical background / prior work

- Key concepts and relevant theories.
- Empirical evidence that motivates your study.
- Define constructs you will operationalise later.

Justification (the gap)

- What is missing/uncertain in the literature?
- Why is your question novel or timely?
- Why does it matter (societal + scientific value)?

Aim for a tight narrative: every paragraph should earn its place.

Introduction: build the rationale

Your introduction should move from context → evidence → gap → contribution.

Theoretical background / prior work

- Key concepts and relevant theories.
- Empirical evidence that motivates your study.
- Define constructs you will operationalise later.

Justification (the gap)

- What is missing/uncertain in the literature?
- Why is your question novel or timely?
- Why does it matter (societal + scientific value)?

Quick self-check (common failure modes)

- Do your citations support the exact claim you make?
- Is the gap a real contradiction/unknown, not “nobody studied X”?
- Are the contribution(s) specific (what will change if you are right)?

Aim for a tight narrative: every paragraph should earn its place.

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

If you cannot point to the exact analysis that answers an RQ, rewrite the RQ.

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as ev

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

A good empirical HCI research question is

1. interaction-centered,
2. specific, and
3. empirically tractable:

it clearly

- defines users, tasks, systems, and context;
- targets observable interaction phenomena rather than vague outcomes;
- is grounded in prior theory without presupposing results;
- and is scoped so it can be meaningfully answered with appropriate empirical methods within a single study.

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as ev

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

Components:

- **User population** – who is interacting (e.g., novices vs experts, clinicians, students, lay users).
- **Interactive system or feature** – what artifact, interface, or AI capability is involved.
- **Task or activity** – what users are trying to accomplish through interaction.
- **Context of use** – under which conditions (e.g., time pressure, uncertainty, collaboration, real vs simulated setting).
- **Outcome or phenomenon of interest** – what is being observed or explained (e.g., performance, errors, reliance, sensemaking, workload, experience).
- **Relationship or mechanism** – how or why the interaction produces certain effects (comparison, influence, mediation, trade-off).

Make the logic explicit: what you want to know, and what would count as ev

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

How does [S] used by [U] during [T] in [C] affect [O], and through what [M]?”

CUSTOM

C – Context of use (setting, constraints, conditions)
U – Users (target population)
S – System / feature (artifact, interaction technique, AI behavior)
T – Task (what users are doing)
O – Outcomes (what you will observe/measure: performance, errors, workload, reliance, experience, etc.)
M – Mechanism / relationship (how/why; comparison, expected influence, mediators/moderators)

…a good research question is “custom” to a specific socio-technical setting, not abstractly universal.

Make the logic explicit: what you want to know, and what would count as ev

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

How does [S] used by [U] during [T] in [C] affect [O], and through what [M]?”

Examples:

How does the spatial arrangement of controls in a mobile navigation interface affect task completion time and error rates for first-time users performing turn-by-turn navigation in unfamiliar urban environments?

How does the explicit visualization of predictive uncertainty in an AI-based clinical decision support system influence clinicians’ reliance, error detection, and decision time when diagnosing cases under time pressure?

How does the integration of a generative AI assistant into multidisciplinary team meetings affect turn-taking, information sharing, and perceived epistemic authority among team members during collaborative planning tasks?

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

In empirical HCI research, **research hypotheses** are formal, testable statements that **operationalize** a research question by making explicit the expected relationship between interaction variables.

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

Their role is not to replace the research question, but to *instantiate it* in a form that can be directly examined through empirical data.

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

NB: The mapping from research question to hypotheses often follows a **one-to-many pattern***: a single research question can give rise to multiple hypotheses, each* **targeting a specific outcome or mechanism** *implied by the question..*

Make the logic explicit: what you want to know, and what would count as evidence

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

A good RH is:

1. **Theoretically grounded** – directly derivable from the research question and motivated by prior literature or design rationale, not post-hoc intuition.
2. **Precise and falsifiable** – formulated as a testable claim, specifying a comparison, direction of effect, or difference between conditions.
3. **Operationally explicit** – clearly indicates how key variables will be instantiated and observed in the empirical study.

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

How does the visualization of model uncertainty in an AI decision support system affect clinicians' interaction and decision making under time pressure?

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

How does the visualization of model uncertainty in an AI decision support system affect clinicians' interaction and decision making under time pressure?

Derived hypotheses

- **H1**: Clinicians using an uncertainty-aware visualization will exhibit 10% lower inappropriate reliance on AI recommendations than clinicians using a confidence-free interface.
- **H2**: Uncertainty visualization will increase 10% decision time under time pressure, relative to the baseline interface.
- **H3**: Uncertainty visualization will improve clinicians' ability to detect incorrect AI recommendations of 10%.

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Why direction (polarization) matters.

a one-sided test is **more powerful** than a two-sided test for effects in the prespecified direction, because the entire Type I error rate α is allocated to a single tail (so the critical value is less extreme in that direction).

This advantage comes with an important trade-off: a one-sided test has essentially no power to detect effects in the opposite direction, which it will treat as non-significant even if they are substantial.

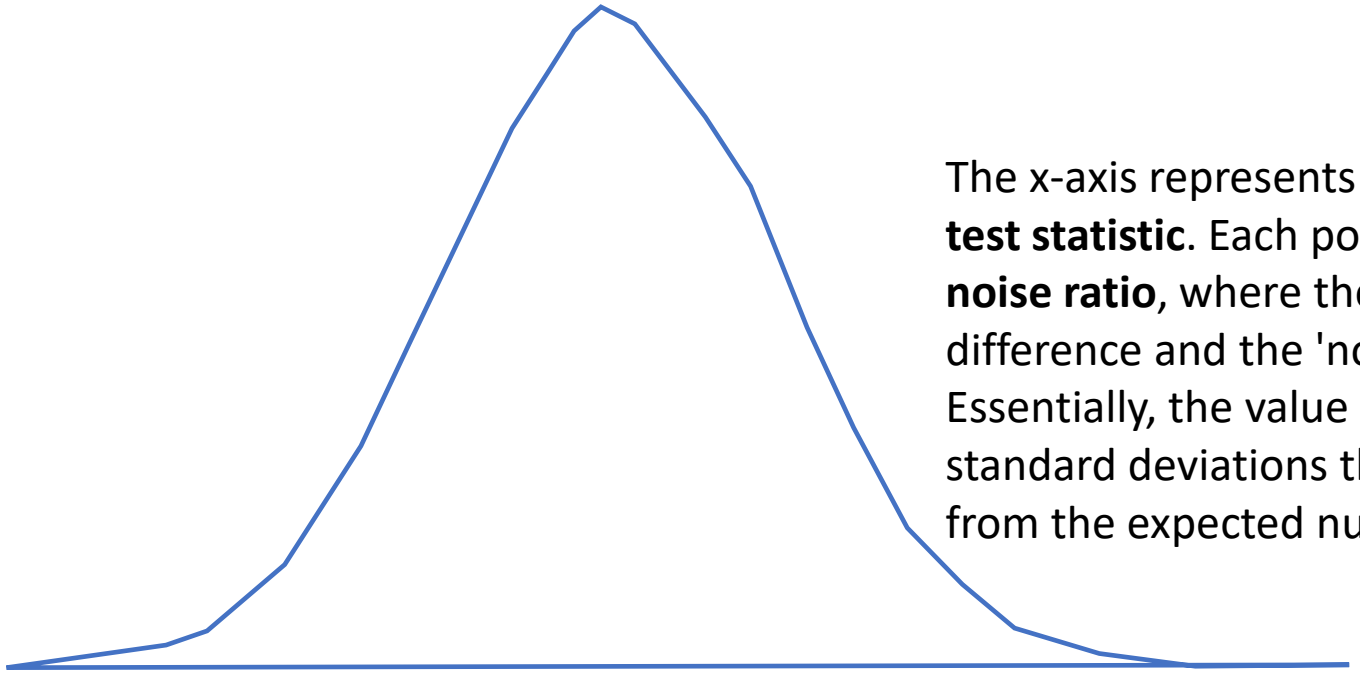
So, a one-sided test is appropriate only if "evidence against H0" is intended to be assessed exclusively in the hypothesized direction, and one is willing to regard an effect in the opposite direction as non-significant even if it is large.

How does the visualization of model uncertainty in an AI decision support system affect clinicians' interaction and decision making under time pressure?

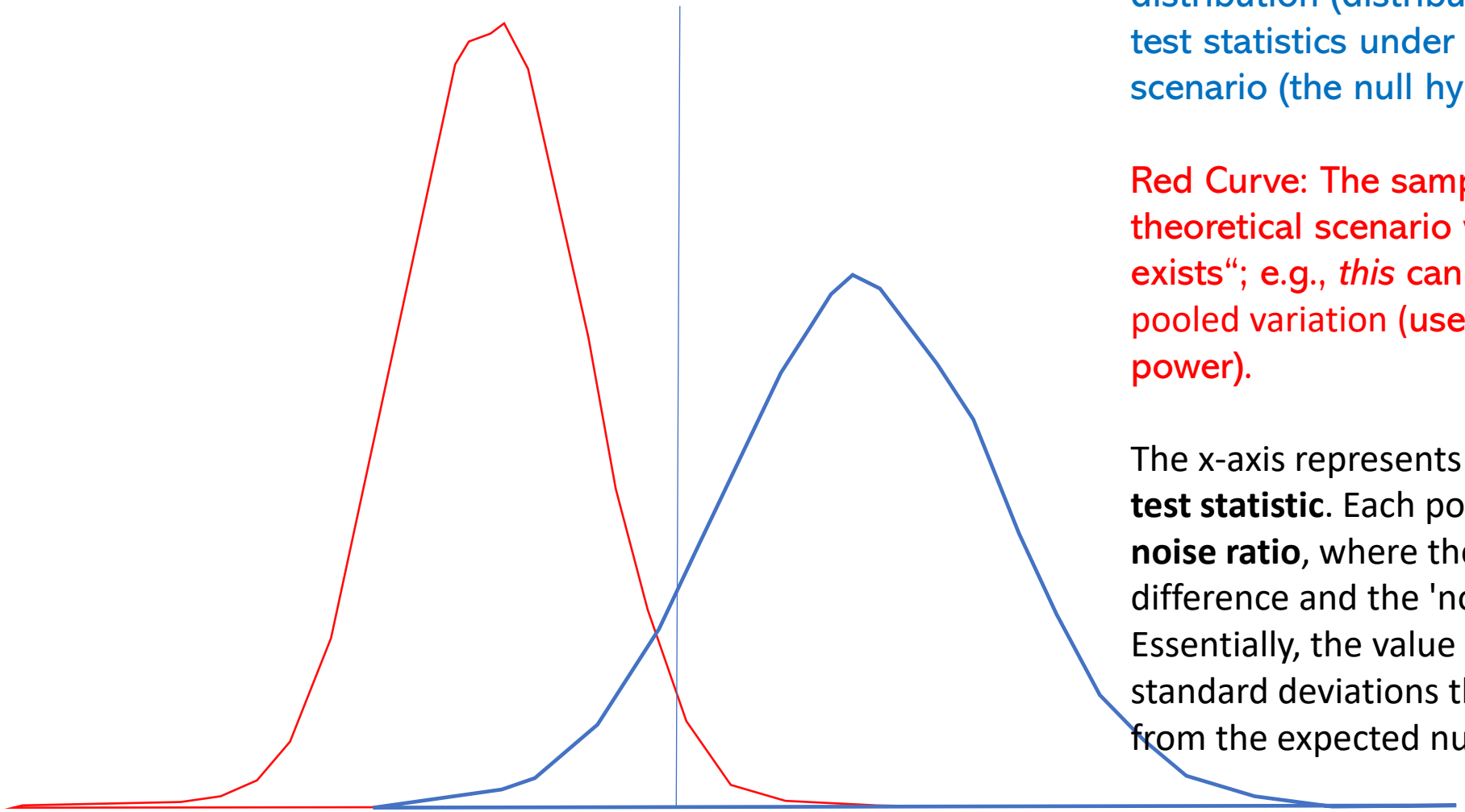
Derived hypotheses

- **H1**: Clinicians using an uncertainty-aware visualization will exhibit 10% lower inappropriate reliance on AI recommendations than clinicians using a confidence-free interface.
- **H2**: Uncertainty visualization will increase 10% decision time under time pressure, relative to the baseline interface.
- **H3**: Uncertainty visualization will improve clinicians' ability to detect incorrect AI recommendations of 10%.

Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).



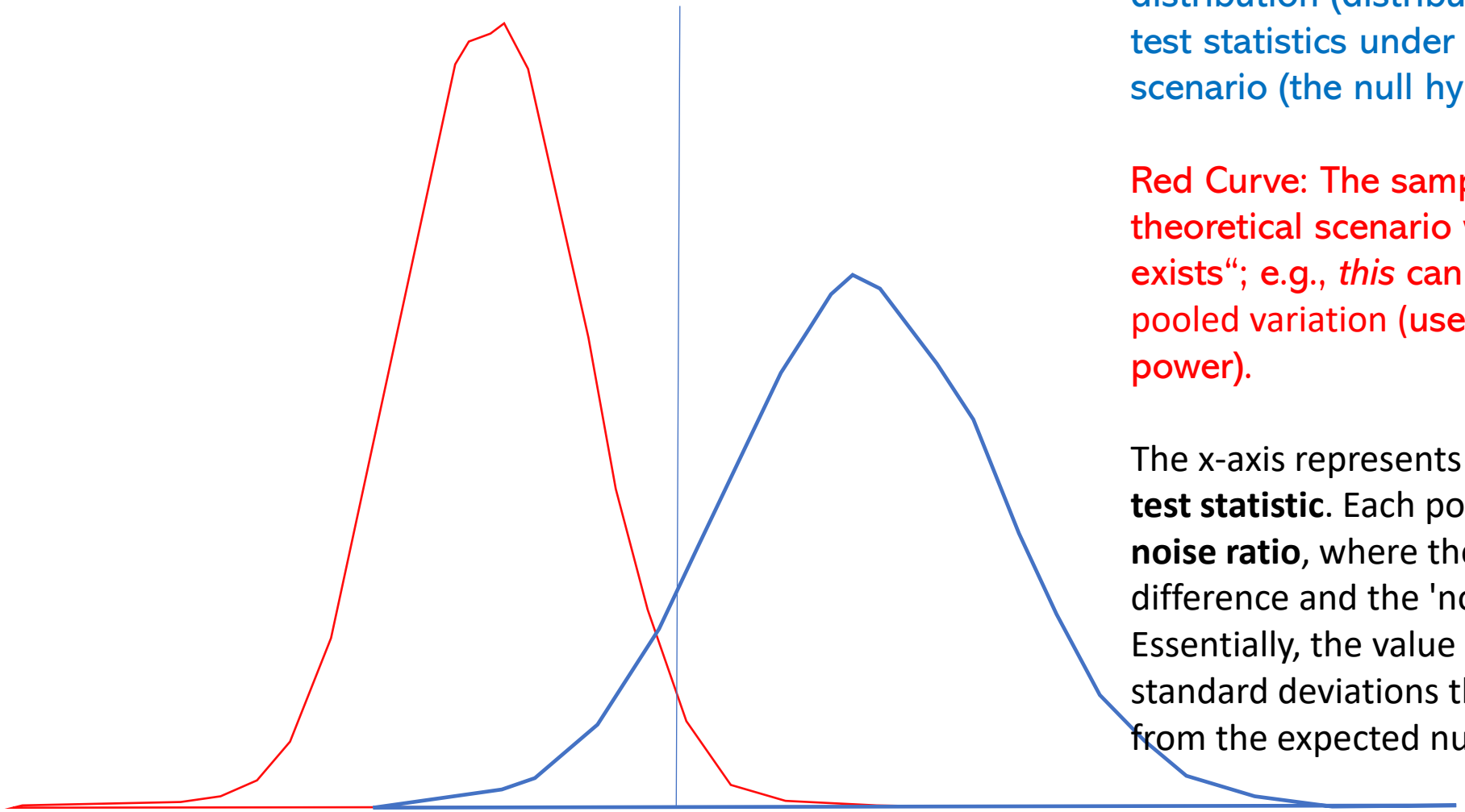
The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.



Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; e.g., *this* can be **−40 units** relative to the pooled variation (used to calculate statistical power).

The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.



Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

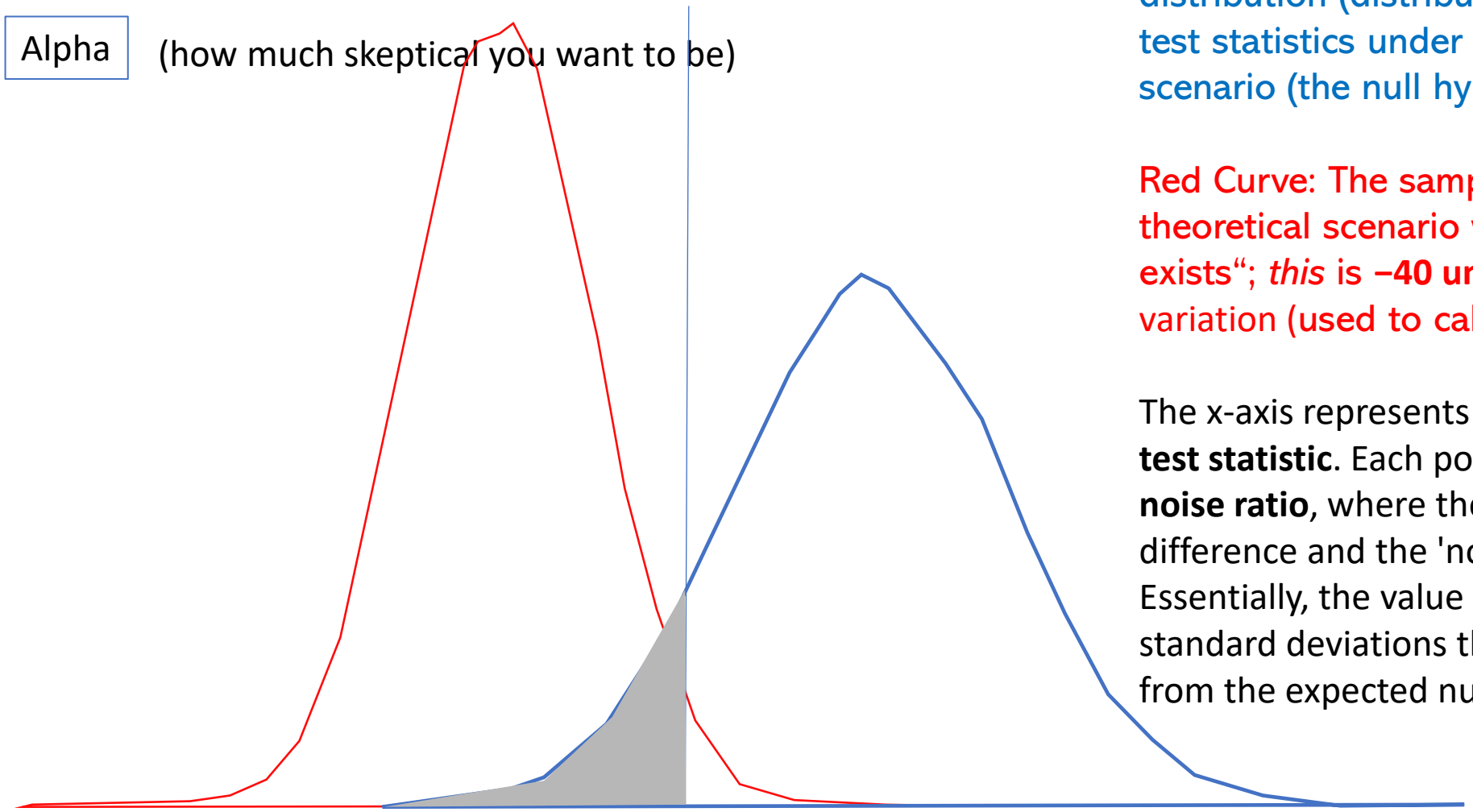
Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; e.g., *this* can be **−40 units** relative to the pooled variation (used to calculate statistical power).

The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.

The test «standardizes» the sampling data, or better yet, **the difference** between our observed data and the null hypothesis.

1) Test and tail

2) Alpha (how much skeptical you want to be)



Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

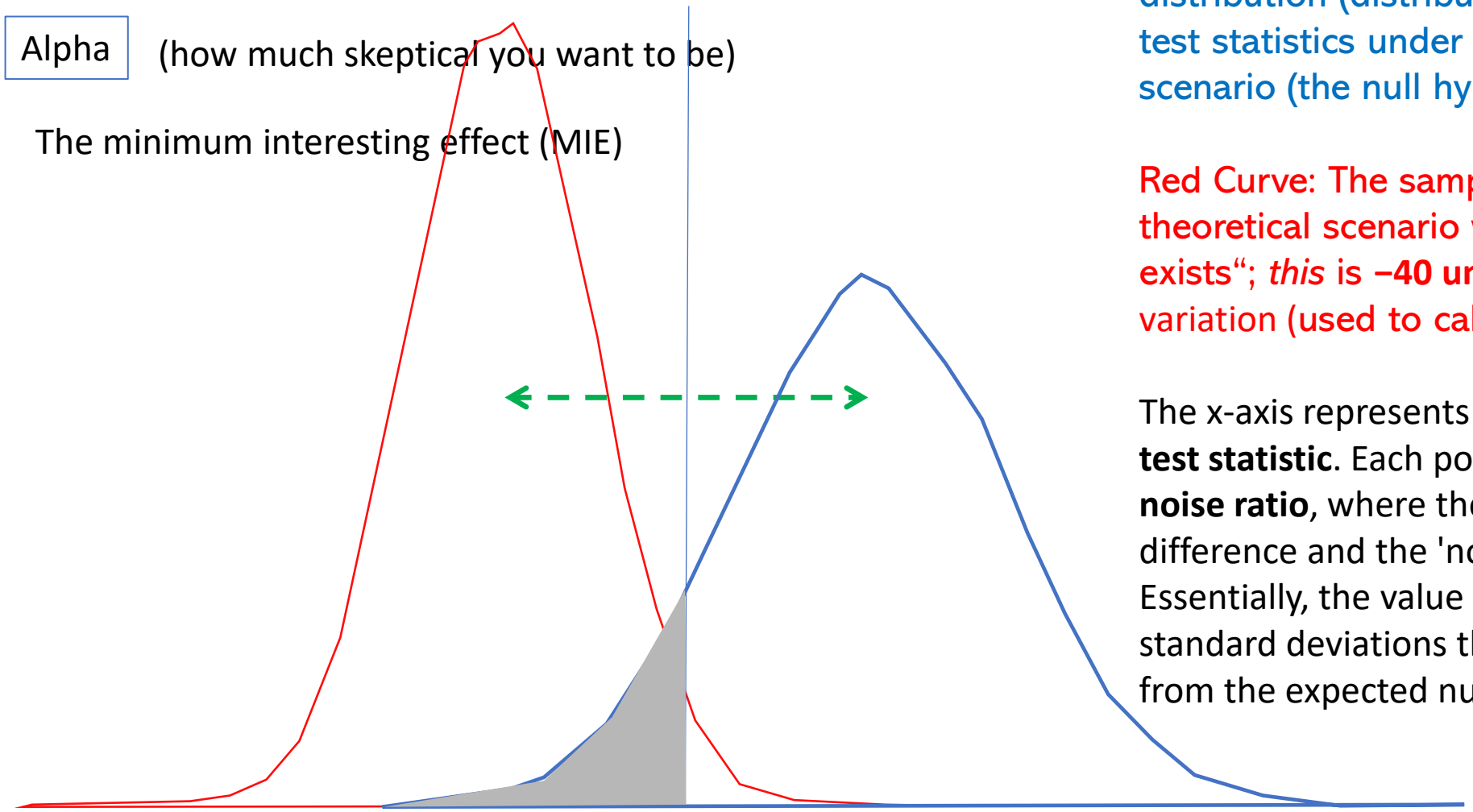
Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; *this* is **−40 units** relative to the pooled variation (used to calculate statistical power).

The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.

1) Test and tail

2) Alpha (how much skeptical you want to be)

3) The minimum interesting effect (MIE)



Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; *this* is **−40 units** relative to the pooled variation (used to calculate statistical power).

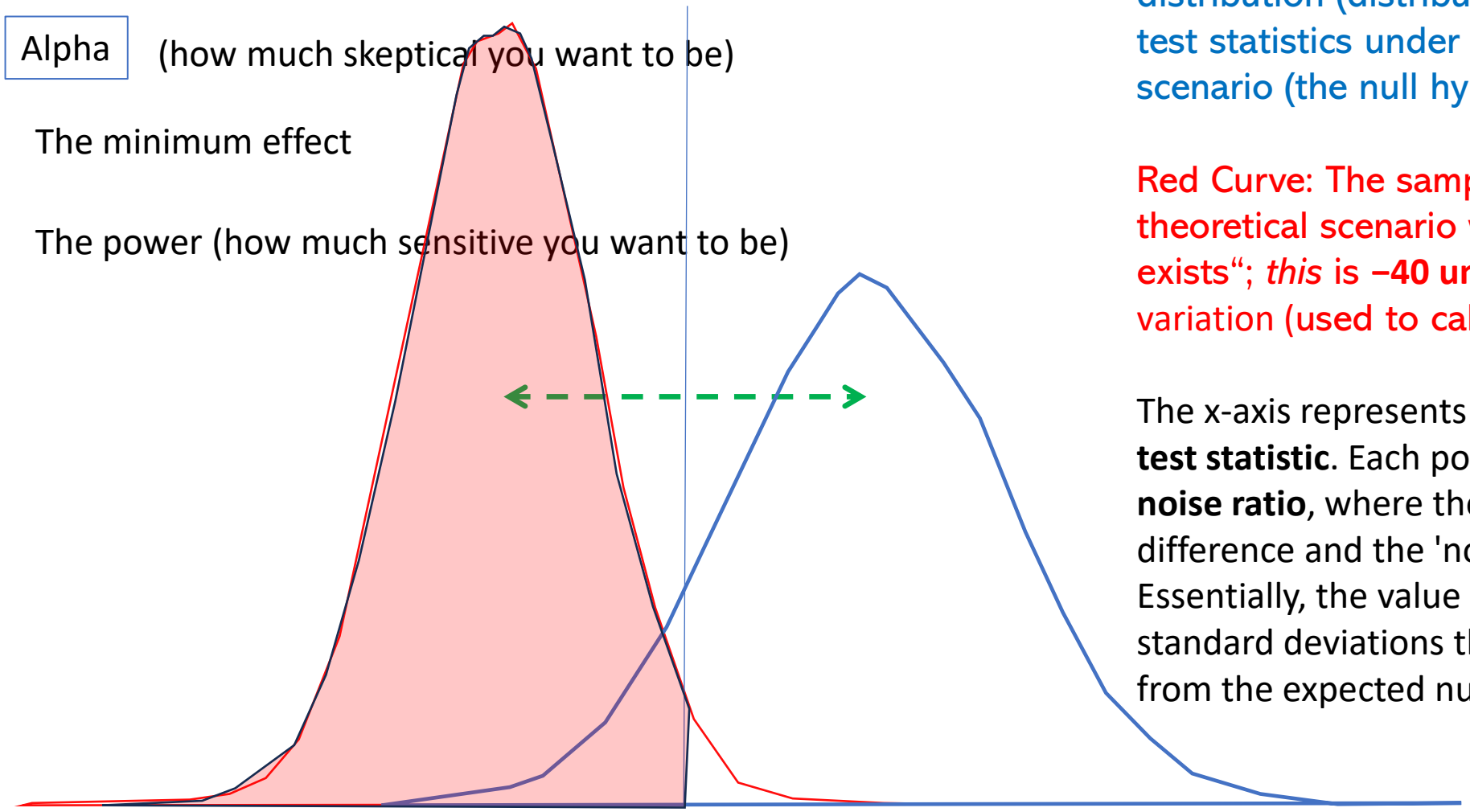
The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.

1) Test and tail

2) Alpha (how much skeptical you want to be)

3) The minimum effect

4) The power (how much sensitive you want to be)



Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; *this* is **−40 units** relative to the pooled variation (used to calculate statistical power).

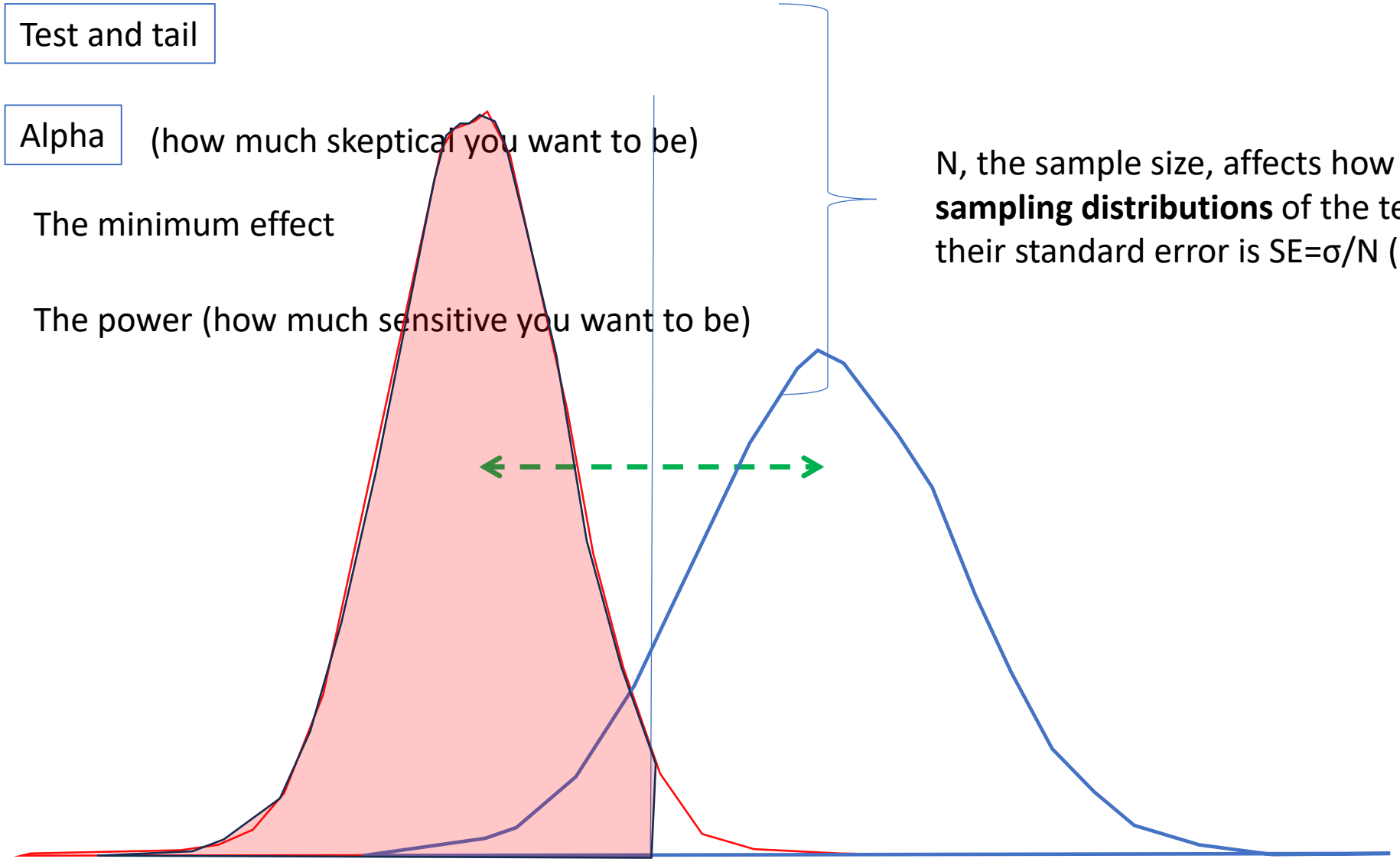
The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.

1) Test and tail

2) Alpha (how much skeptical you want to be)

3) The minimum effect

4) The power (how much sensitive you want to be)



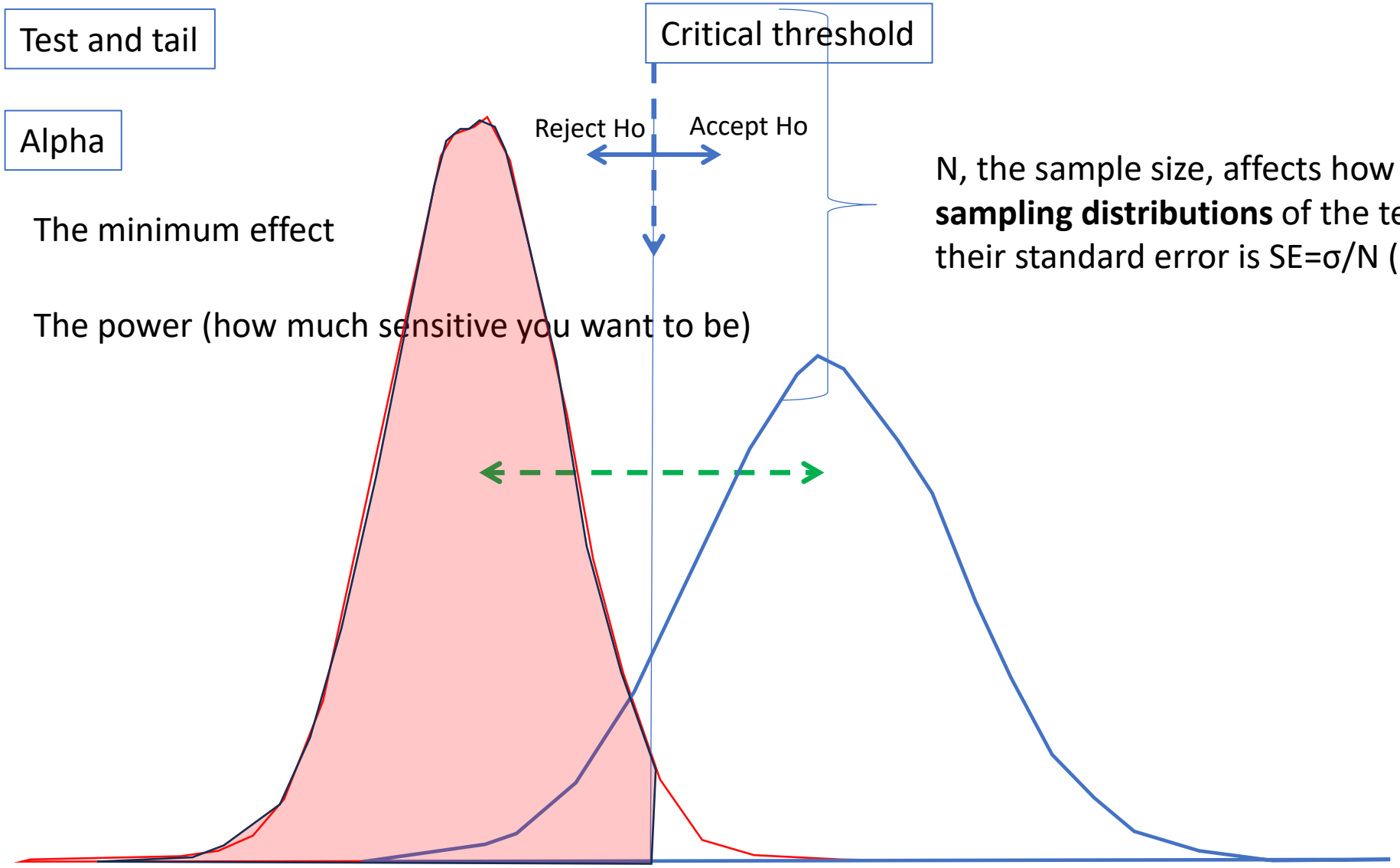
N, the sample size, affects how steep and narrow the **sampling distributions** of the test statistics are, because their standard error is SE=σ/N (and their variance is σ2/N).

1) Test and tail

2) Alpha

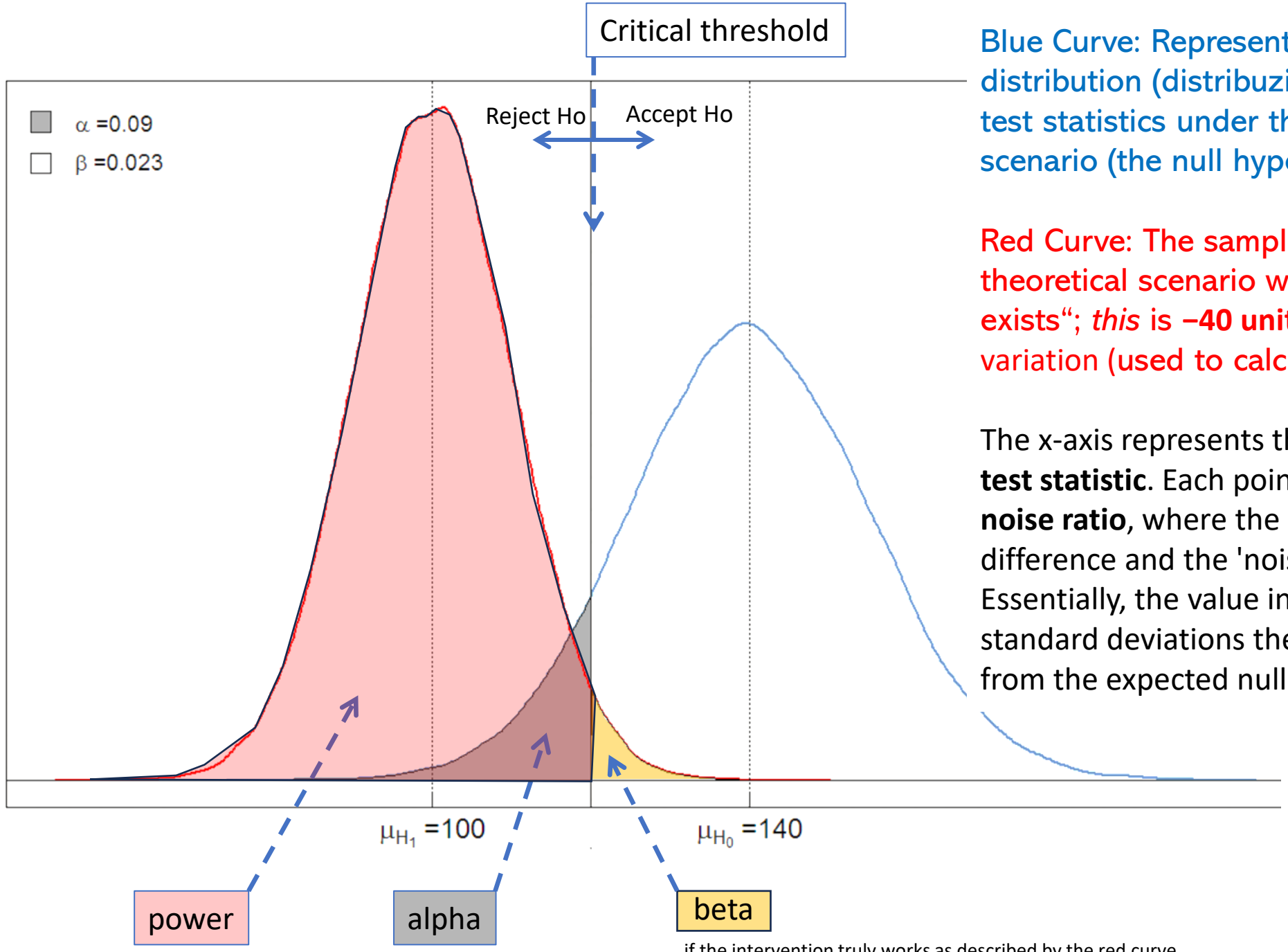
3) The minimum effect

4) The power (how much sensitive you want to be)



N, the sample size, affects how steep and narrow the **sampling distributions** of the test statistics are, because their standard error is SE=σ/N (and their variance is σ2/N).

This output is then compared to the **Blue Curve** to find the p-value, or compared to the **Black Line** (the critical threshold) to make a "Reject/Accept" decision.



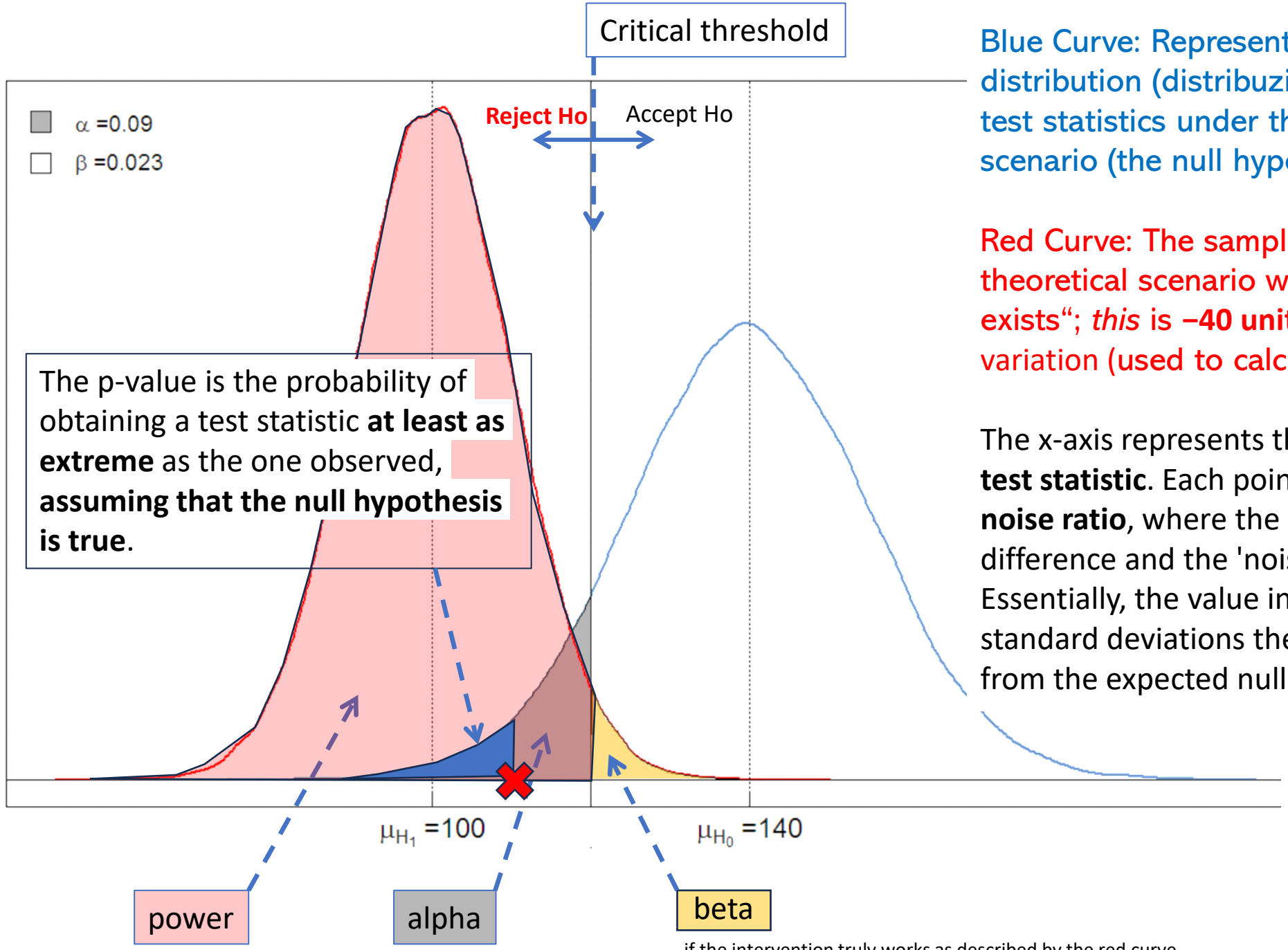
Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; *this* is **−40 units** relative to the pooled variation (used to calculate statistical power).

The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.

H1:
The mean of the population is significantly lower than 140 (μ<140). Specifically, we hypothesize a substantial reduction of ***40 units*****.**

if the intervention truly works as described by the red curve, we will only fail to notice it about 2 times out of 100.



Blue Curve: Represents the sampling distribution (distribuzione campionaria) of the test statistics under the theoretical "no effect" scenario (the null hypothesis).

Red Curve: The sampling distribution under the theoretical scenario where "this specific effect exists"; *this* is **–40 units** relative to the pooled variation (used to calculate statistical power).

The x-axis represents the **sampling space of the test statistic**. Each point on this axis is a **signal-to-noise ratio**, where the 'signal' is the observed difference and the 'noise' is the **standard error**. Essentially, the value indicates how many standard deviations the observed difference lies from the expected null value.

H1:
The mean of the population is significantly lower than 140 (μ<140). Specifically, we hypothesize a substantial reduction of ***40 units*****.**

if the intervention truly works as described by the red curve, we will only fail to notice it about 2 times out of 100.

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

Example

Aim: “To determine whether showing users an AI decision aid’s confidence score leads to more appropriately calibrated reliance during decision-making.”

RQ: “Does presenting a confidence score alongside an explanation change users’ reliance so that they follow the AI more when it is correct and less when it is incorrect?”

H (if any): “Users who see a well-calibrated confidence score will show more appropriate reliance—higher adherence on correct AI recommendations and lower adherence on incorrect ones—than users who do not see confidence.”

Operationalise (IV / DV / controls + metrics): “IV: confidence display condition (none vs uncalibrated score vs calibrated score) alongside the same explanation; DV: appropriate reliance quantified as (adherence when AI correct − adherence when AI wrong), plus decision accuracy and decision time; controls: fixed AI accuracy profile across participants, counterbalanced case order and difficulty, and measured prior domain expertise; metrics: adherence rate by correctness, accuracy (% correct), response time, and self-reported trust/workload (e.g., Likert trust, NASA-TLX).”

Aims → research questions → hypotheses

Make the logic explicit: what you want to know, and what would count as evidence.

One chain per question

Aim

What is the overall goal?

RQ

What specific question will you answer?

H (if any)

Directional prediction + variables

Operationalise

IV / DV / controls + metrics

Rules of thumb

- If you cannot point to the exact analysis that answers an RQ, rewrite the RQ.
- Each hypothesis must be testable with your measures.
- State variables explicitly: what you manipulate vs measure.
- Avoid vague verbs (“explore”) unless using qualitative designs.
- If mixed-methods, map each RQ to a method and output.
- Keep a consistent naming scheme for constructs and metrics.

Study design: specify structure, not labels

Declare the design

- Experimental, correlational, qualitative, mixed-methods…
- Between-subjects / within-subjects / factorial / longitudinal…
- Conditions, factors, and what varies across participants.

Example: 2×2 factorial

	Factor B: Low	Factor B: High
Factor A: Control	C1	C2
Factor A: Treatment	C3	C4

Include a visual representation

- A table/flowchart clarifies factors, conditions, and sequencing.
- Make counterbalancing, randomisation, and timing visible.
- If complex, add a diagram in an appendix and reference it.

Study design: specify structure, not labels

Declare the design

- Experimental, correlational, qualitative, mixed-methods…
- Between-subjects / within-subjects / factorial / longitudinal…
- Conditions, factors, and what varies across participants.

Example: 2×2 factorial

	Factor B: Low	Factor B: High
Factor A: Control	C1	C2

Include a visual representatio

- A table/flowchart clarifies factors
- Make counterbalancing, randomi
- If complex, add a diagram in an a

Between-subjects design (AKA between-groups or independent measures designs),

Within-subjects design

Declare the design

- Experimental, correlational, qualitative, mixed-methods…
- Between-subjects / within-subjects / factorial / longitudinal…
- Conditions, factors, and what varies across participants.

Example: 2×2 factorial

	Factor B: Low	Factor B: High
Factor A: Control	C1	C2

Include a visual representatio

- A table/flowchart clarifies factors
- Make counterbalancing, randomi
- If complex, add a diagram in an ap

Between-subjects design (AKA between-groups or independent measures designs),

1. Each participant experiences only one condition,
2. Researchers compare differences between groups,
3. Can help avoid testing fatigue, and Usually has a control group

Within-subjects design

Declare the design

- Experimental, correlational, qualitative, mixed-methods…
- Between-subjects / within-subjects / factorial / longitudinal…
- Conditions, factors, and what varies across participants.

Example: 2×2 factorial

	Factor B: Low	Factor B: High
Factor A: Control	C1	C2

Include a visual representatio

- A table/flowchart clarifies factors
- Make counterbalancing, randomi
- If complex, add a diagram in an ap

Between-subjects design (AKA between-groups or independent measures designs),

1. Each participant experiences only one condition,
2. Researchers compare differences between groups,
3. Can help avoid testing fatigue, and Usually has a control group

Within-subjects design

1. Participants experience all conditions
2. Researchers compare differences between conditions within the same participants
3. Participants serve as their own control
4. Can help detect real differences between conditions

Declare the design

- Experimental, correlational, qualitati
- Between-subjects / within-subjects /
- Conditions, factors, and what varies a

which experimental design is statistically more powerful
(less prone to false negative errors, Type II errors)

Example: 2×2 factorial

	Factor B: Low	Factor B: High
Factor A: ntrol	C1	C2

Include a visual representatio

- A table/flowchart clarifies factors
- Make counterbalancing, randomi
- If complex, add a diagram in an ap

Between-subjects design (AKA between-groups or independent measures designs),

1. Each participant experiences only one condition,
2. Researchers compare differences between groups,
3. Can help avoid testing fatigue, and Usually has a control group

Within-subjects design

1. Participants experience all conditions
2. Researchers compare differences between conditions within the same participants
3. Participants serve as their own control
4. Can help detect real differences between conditions

Declare the design

- Experimental, correlational, qualitati
- Between-subjects / within-subjects /
- Conditions, factors, and what varies a

Include a visual representatio

- A table/flowchart clarifies factors
- Make counterbalancing, randomi
- If complex, add a diagram in an ap

which experimental design is statistically more powerful
(less prone to false negative errors, Type II errors)

Within-subjects designs are more powerful
- Each participant acts as their own control.
- Inter-subject variability is removed from the error term.
- For the same sample size, confidence intervals are narrower.

Between-subjects design (AKA between-groups or independent measures designs),
1. Each participant experiences only one condition,
2. Researchers compare differences between groups,
3. Can help avoid testing fatigue, and Usually has a control group

Within-subjects design
1. Participants experience all conditions
2. Researchers compare differences between conditions within the same participants
3. Participants serve as their own control
4. Can help detect real differences between conditions

Declare the design

- Experimental, correlational, qualitative, mixed-methods…
- Between-subjects / within-subjects / factorial / longitudinal…
- Conditions, factors, and what varies across participants.

Include a visual representatio

- A table/flowchart clarifies factors
- Make counterbalancing, randomi
- If complex, add a diagram in an ap

When to use each design

1. Between-subjects designs are useful when exposure to one condition might affect responses to another condition.
2. Within-subjects designs are less likely to miss real differences between conditions.
3. Both designs have advantages and disadvantages, and the choice between them depends on the research question and other factors.

Between-subjects de
measures designs),

1. Each participant experiences only one condition,
2. Researchers compare differences between groups,
3. Can help avoid testing fatigue, and Usually has a control group

Within-subjects design

1. Participants experience all conditions
2. Researchers compare differences between conditions within the same participants
3. Participants serve as their own control
4. Can help detect real differences between conditions

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. Planned sample size

1. **Target population**
2. **Recruitment**
3. **Compensation**
4. **Sampling strategy**
5. **Planned sample size**

1. define the population of interest and justify its selection;
2. specify relevant demographics or characteristics and clear inclusion/exclusion criteria.

1. Target population
2. **Recruitment**
3. Compensation
4. Sampling strategy
5. Planned sample size

1. describe how participants will be recruited and
2. how the study invitation will be disseminated.

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. Planned sample size

1. indicate whether participation is compensated or on voluntary basis and
2. specify the type and amount of compensation.

1. Target population
2. Recruitment
3. Compensation
4. **Sampling strategy**
5. Planned sample size

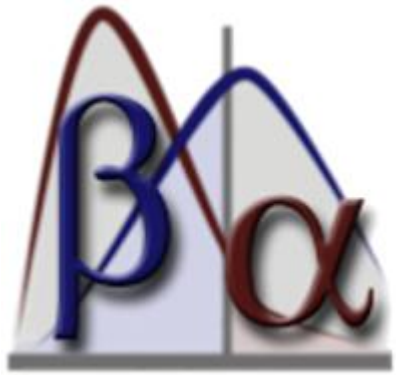
1. state the sampling approach: probabilistic (random, stratified, cluster…) or non-probabilistic convenience, purposive, snowball…)
2. the specific method used
3. constraints and anticipated biases.

1. **Target population**
2. **Recruitment**
3. **Compensation**
4. **Sampling strategy**
5. **Planned sample size**

1. report the expected number of participants
2. provide a justification based on prior studies or an a priori power analysis (effect size, power, alpha).

1. **Target population**
2. **Recruitment**
3. **Compensation**
4. **Sampling strategy**
5. **Planned sample size**

1. report the expected number of participants
2. provide a justification based on prior studies or an a priori power analysis (effect size, power, alpha).

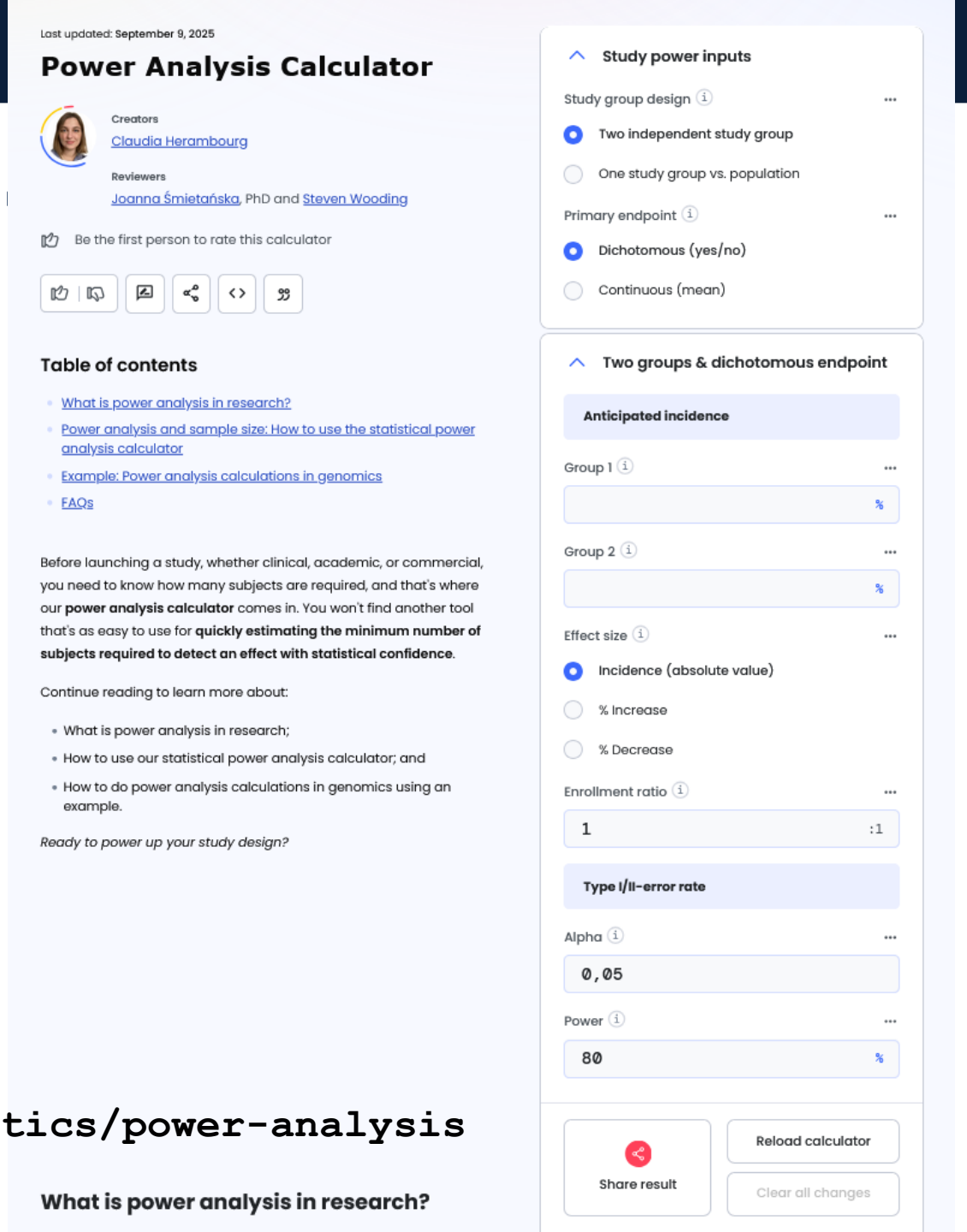


G*Power

https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Participants: who, how many, and why

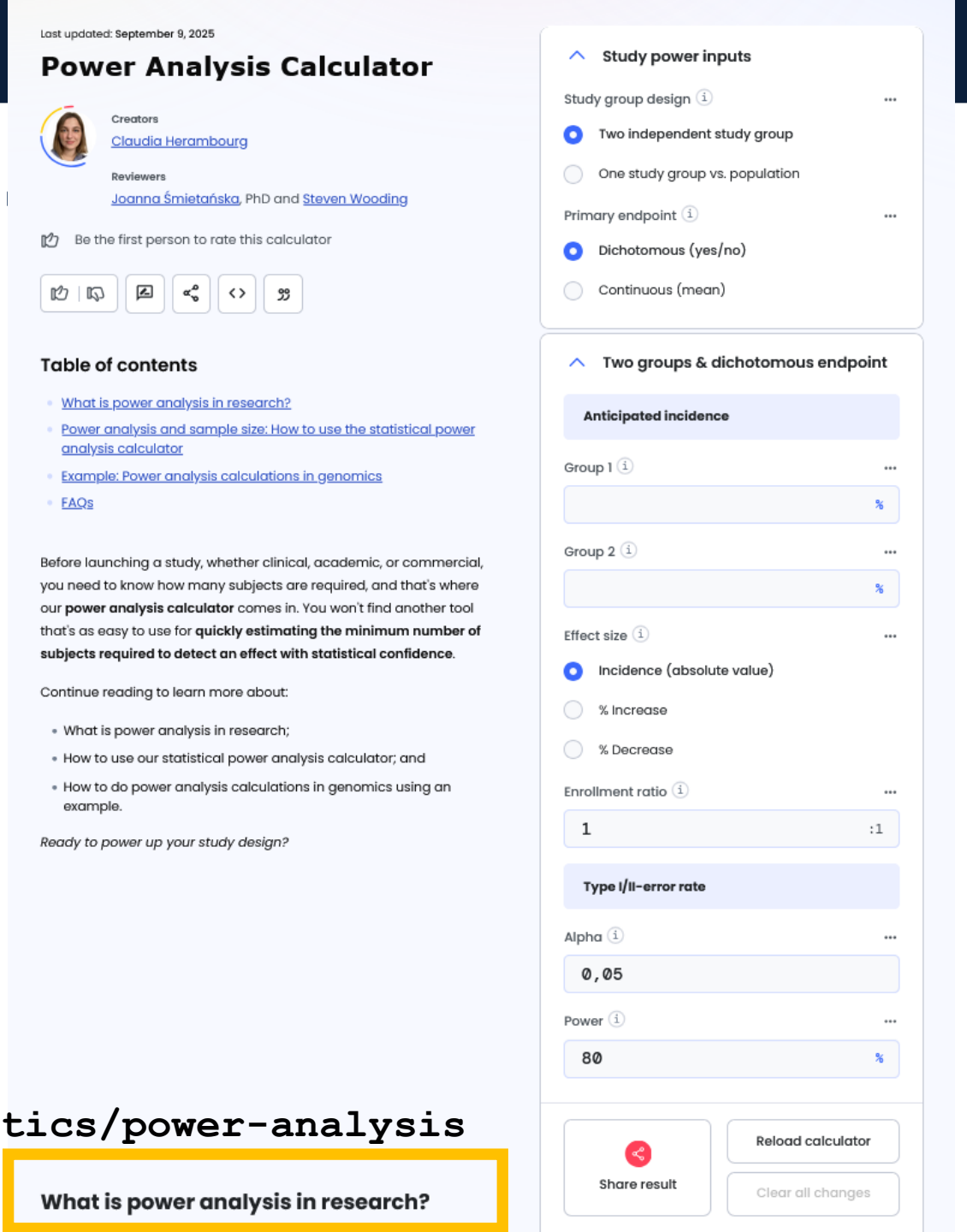
1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**



https://www.omnicalculator.com/statistics/power-analysis

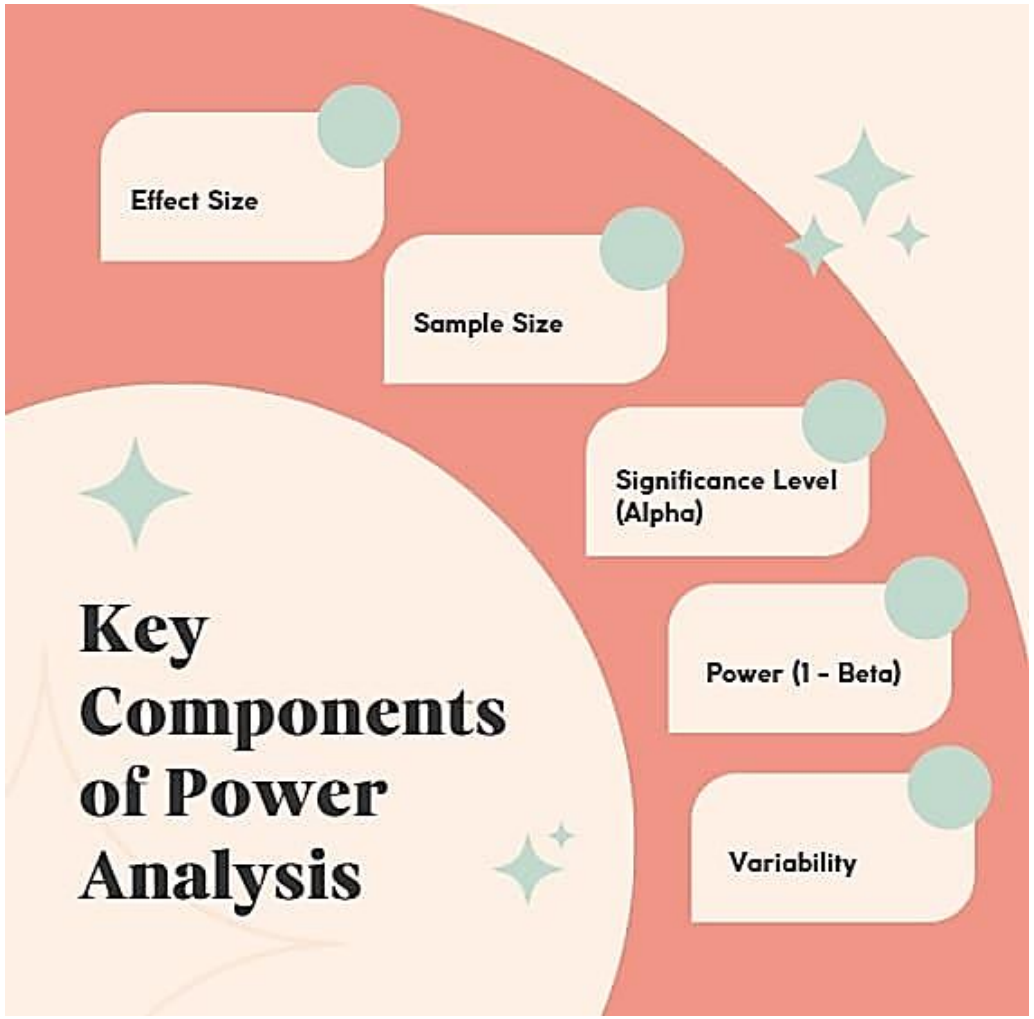
Participants: who, how many, and why

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**



https://www.omnicalculator.com/statistics/power-analysis

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**



What is power analysis in research?

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**

the power analysis is aimed at estimating the **minimum sample size** that is required to achieve a prespecified **statistical power (1 − β)** to detect a prespecified **effect size** at a **chosen significance level α**, under the assumed data-generating model and **variability**.

What is power analysis in research?

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. Planned sample size

Some unconfortable truths on power analysis

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**

Some unconfortable truths on power analysis

Power analysis does not "protect" you from null results — it only protects you against *missing one specific effect size*

Power is always defined **conditional on a single, assumed effect size.**

If the true effect is smaller than the one you assumed (which is extremely common), the nominal power is meaningless.

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**

Some unconfortable truths on power analysis

Power analysis implicitly assumes that the alternative hypothesis is *true*

This is rarely stated explicitly, but it is structurally unavoidable.

Power is defined as:

> P(reject H_0 | H_1 with effect size δ is true)

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**

Some uncomfortable truths on power analysis

Conventional power thresholds (e.g., 80%) have no theoretical justification

The choice of 80% (or $\beta = 0.20$) is **pure convention**, not the result of optimality, decision theory, or statistical necessity.

Historically:

- The threshold emerged from early statistical practice and stuck.
- No general loss function, cost model, or epistemic principle selects 80% as special.

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**

SOME UNCONFORTABLE TRUTHS ON POWER ANALYSIS

For you , is it more relevant alpha (type I errors) or beta (type II errors)?

Conventional power thresholds (e.g., 80%) have no theoretical justification

The choice of 80% (or β = 0.20) is **pure convention**, not the result of optimality, decision theory, or statistical necessity.

Historically:

- The threshold emerged from early statistical practice and stuck.
- No general loss function, cost model, or epistemic principle selects 80% as special.

1. Target population
2. Recruitment
3. Compensation
4. Sampling strategy
5. **Planned sample size**

SOME UNCONFORTABLE TRUTHS ON POWER ANALYSIS

For you , is it more relevant alpha (type I errors) or beta (type II errors)?

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

$$\beta(\delta) = P(\text{fail to reject } H_0 \mid H_1 \text{ with effect } \delta \text{ is true})$$

Conventional power thresholds (e.g., 80%) have no theoretical justification

The choice of 80% (or β = 0.20) is **pure convention**, not the result of optimality, decision theory, or statistical necessity.

Historically:

- The threshold emerged from early statistical practice and stuck.
- No general loss function, cost model, or epistemic principle selects 80% as special.

Ethics checklist (must be explicit)

- Approval status (or plan) with your ethics committee.
- Informed consent: purpose, tasks, duration, contacts.
- Participant rights: autonomy, withdrawal, debriefing.
- Data protection: anonymity/pseudonymisation, access, storage.
- Risks/discomforts (physical/psychological/social) + mitigation.

Pre-registration & sharing

- State whether you will pre-register the study.
- Specify what will be shared: materials, code, data, analysis.
- Use repositories (e.g., OSF; AsPredicted) when appropriate.

Appendices (examples)

- Appendix I: Informed consent template (Italian)
- Appendix II: Data protection information (GDPR-oriented)

In your protocol, treat ethics as part of the method: concrete procedures, not slogans.

Ethics & open research practices

Ethics checklist (must be explicit)

- Approval status (or plan) with your eth committee.
- Informed consent: purpose, tasks, duration, contacts.
- Participant rights: autonomy, with debriefing.
- Data protection: anonymity/pseudonymisation, acces storage.
- Risks/discomforts (physical/psychological/social) + mitigation.

Questions you have to address:

1. What is the purpose of this study?
2. How will the study be conducted?
3. Why are you being invited to participate?
4. Are you required to participate in the study?
5. What steps are required to take part in the study?
6. What will you be asked to do?
7. What are the possible risks and inconveniences of the study?
8. What are the possible benefits of participating in the study?
9. What will happen if information relevant to your health emerges during the study?
10. How is the confidentiality of information ensured?
11. How will your personal data be used?
12. Are there any other important pieces of information you should know?
13. Tell ho is the data controller (titolare, usually the university or research institution) and data processor (responsabile), who processes personal data on behalf of the data controller

- Appendix I: Informed consent template (Italian)
- Appendix II: Data protection information (GDPR-oriented)

In your protocol, treat ethics as part of the method: concrete procedures, not slogans.

Materials: instruments, stimuli, software

Report what others need to reproduce the study (and to judge measurement quality).

Questionnaires / scales

- Full reference and intended population.
- Construct/domain measured and #items.
- Reliability/validity evidence (if available).
- Scoring and interpretation (e.g., subscales).

Stimuli

- Images, videos, text, interfaces, tasks…
- Source (dataset, bank, prior work) OR creation method.
- How stimuli vary across conditions (if applicable).
- Include examples in an appendix.

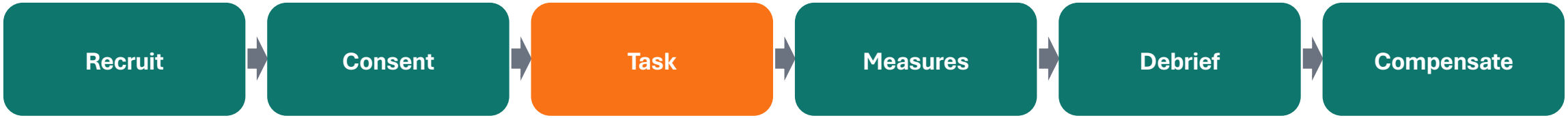
Software

- Tool name, developer, version (e.g., Qualtrics).
- Key features used (timing, randomisation, logging).
- If custom code: what it does + how to access it.
- Document hardware/platform constraints.

Practical: put “version + configuration” in the protocol, not only in your lab notebook.

Procedure: write it like a script

Describe the participant journey step-by-step, including setting, timing, and randomisation.



Minimum detail to include

- Setting: lab / online / field; platform and devices.
- Instructions verbatim or in an appendix.
- Timing: durations, stimulus exposure, breaks, deadlines.
- Randomisation/counterbalancing procedures.
- Conditions: how they are operationalised.

Write for reproducibility

- Use numbered steps and expected outputs.
- Explicitly state what participants see and do.
- Log what matters (events, timestamps, condition IDs).
- If changes occur during piloting, update the protocol.

Your plan should let a reviewer verify that the analyses answer the research questions.

Define variables

- Independent variables (manipulated / predictors).
- Dependent variables (measured outcomes).
- Control/covariates (if needed).
- Operationalisation: coding rules + key metrics.
- Behavioural + self-report measures are both valid—justify each.

Map each RQ to an analysis

RQ	Metric	Model / test
RQ1	DV1	Primary analysis
RQ2	DV2	Secondary / robustness
RQ3	Qual notes	Thematic / mixed-methods

Explicitly label: primary vs exploratory analyses.

Discussion: what you expect and what it would mean

Anticipated results

- Summarise expected patterns (not detailed numbers).
- Relate to prior work: confirm, extend, or contradict.
- State what result would be theoretically informative.

Limitations (and mitigation)

- Constraints of design and sampling.
- Threats to validity and how you reduce them.
- What your study cannot conclude.

Implications & contributions

- HCI theory: what construct, mechanism, or boundary condition?
- Practice: design guidance, evaluation metrics, stakeholder impact.
- Societal impact: fairness, safety, accessibility, wellbeing, etc.
- Future work: concrete follow-ups enabled by this protocol.

A strong discussion section makes clear what *changes* if the study succeeds—or fails.

Discussion: what you expect and what it would mean

Anticipated results

- Summarise expected patterns (not detailed numbers).
- Relate to prior work: confirm, extend, or contradict.
- State what result would be theoretically informative.

Limitations (and mitigation)

- Constraints of design and sampling.
- Threats to validity and how you reduce them.
- What your study cannot conclude.

Implications & contributions

- HCI theory: what construct, mechanism, or boundary condition?
- Practice: design guidance, evaluation metrics, stakeholder impact.
- Societal impact: fairness, safety, accessibility, wellbeing, etc.
- Future work: concrete follow-ups enabled by this protocol.

A strong discussion section makes clear what *changes* if the study succeeds—or fails.

Discussion: what you expect and what it would mean

Anticipated results

- Summarise expected patterns (not detailed numbers).
- Relate to prior work: confirm, extend, or contradict.
- State what result would be theoretically informative.

Limitations (and mitigation)

- Constraints of design and sampling.
- Threats to validity and how you reduce them.
- What your study cannot conclude.

Implications & contributions

- HCI theory: what construct, mechanism, or boundary condition?
- Practice: design guidance, evaluation metrics, stakeholder impact.
- Societal impact: fairness, safety, accessibility, wellbeing, etc.
- Future work: concrete follow-ups enabled by this protocol.

A strong discussion section makes clear what *changes* if the study succeeds—or fails.

Type of Validity
Internal Validity
External Validity
Construct Validity
Content Validity
Face Validity
Ecological Validity
Statistical Conclusion Validity
Conclusion Validity

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity							
External Validity							
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.						
External Validity							
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?					
External Validity							
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).				
External Validity							
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?			
External Validity							
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity							
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.						
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.	Will it work elsewhere?					
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.	Will it work elsewhere?	Real-world relevance.				
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.	Will it work elsewhere?	Real-world relevance.	Will results from a usability study with students hold for professional users in the workplace?			
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.	Will it work elsewhere?	Real-world relevance.	Will results from a usability study with students hold for professional users in the workplace?	Unrepresentative samples, artificial tasks or environments.	Diverse samples, field studies, realistic task settings.	Representativeness of sample, realism of task/context
Construct Validity							
Content Validity							
Face Validity							
Ecological Validity							
Statistical Conclusion Validity							
Conclusion Validity							

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.	Will it work elsewhere?	Real-world relevance.	Will results from a usability study with students hold for professional users in the workplace?	Unrepresentative samples, artificial tasks or environments.	Diverse samples, field studies, realistic task settings.	Representativeness of sample, realism of task/context
Construct Validity	The extent to which the test or measurement accurately represents the theoretical construct it is intended to measure.	Are we measuring the right concept?	Are you measuring what you think you're measuring?	You claim to measure 'user trust in AI' but only ask if users found the system 'useful.'	Poor construct definition, narrow operationalization, mono-method bias.	Precise definitions, multiple indicators, validated instruments.	Theoretical clarity, valid operationalization, convergent measures
Content Validity	The degree to which a test or instrument fully captures the relevant aspects of the construct.	Are we covering all relevant dimensions?	Is the measurement comprehensive?	A usability questionnaire includes items on aesthetics and performance but omits learnability or error recovery.	Omission of key aspects of the construct being measured.	Thorough content analysis, expert review, item comprehensiveness.	Comprehensive item coverage, domain mapping
Face Validity	The extent to which a measure appears, on the surface, to assess what it claims to measure.	Does it look like it measures what it claims to?	Credibility and transparency to participants and stakeholders.	A scale titled 'User Satisfaction' asks about navigation, layout, and visual appealâ€”participants can relate the questions to the label.	Ambiguous or misaligned items that reduce user trust or engagement.	Clear alignment between items and construct, user involvement.	Transparent alignment between measure and construct
Ecological Validity	The degree to which the study setting, task, and context resemble the real world.	Does the context resemble real use?	Does the study simulate realistic use conditions?	A driving AI evaluated on a static simulator lacks realism compared to real road testing.	Unrealistic settings, tasks, or context for intended use.	Naturalistic settings, context-aware design, field evaluation.	Realistic tasks, environments, and interaction contexts
Statistical Conclusion Validity	The extent to which the statistical analysis is appropriate, and the results are not due to chance, insufficient power, or flawed assumptions.	Are the findings statistically reliable and valid?	Are your conclusions based on solid statistics?	You fail to detect an effect because the study is underpowered, even though the effect exists.	Low power, assumption violations, misuse of statistics.	Power analysis, correct statistical tests, assumption checks.	Appropriate statistical methods, power, assumption checks
Conclusion Validity	The overall credibility of the interpretation of the data.	Do the data support the conclusions?	Are the conclusions drawn justified by the evidence?	All data are consistent, but interpretation exaggerates what the evidence actually supports.	Overinterpretation, cherry-picking, ignoring limitations.	Conservative interpretation, triangulation, transparency about limits.	Careful interpretation, integration of multiple forms of evidence

Type of Validity	Definition	Key Question	Concern	Example	Threatened By	Improved By	Controlled By
Internal Validity	The degree to which a causal relationship between variables can be established.	Did X cause Y?	Eliminating alternative explanations (confounds, biases).	Does a new AI interface increase user performance, or was the improvement due to learning effects?	Confounding variables, selection bias, history effects, etc.	Random assignment, control groups, blinding, standard protocols.	Design rigor, random assignment, control of confounds
External Validity	The extent to which findings generalize to other contexts, populations, tasks, or times.	Will it work elsewhere?	Real-world relevance.	Will results from a usability study with students hold for professional users in the workplace?	Unrepresentative samples, artificial tasks or environments.	Diverse samples, field studies, realistic task settings.	Representativeness of sample, realism of task/context
Construct Validity	The extent to which the test or measurement accurately represents the theoretical construct it is intended to measure.	Are we measuring the right concept?	Are you measuring what you think you're measuring?	You claim to measure 'user trust in AI' but only ask if users found the system 'useful.'	Poor construct definition, narrow operationalization, mono-method bias.	Precise definitions, multiple indicators, validated instruments.	Theoretical clarity, valid operationalization, convergent measures
Content Validity	The degree to which a test or instrument fully captures the relevant aspects of the construct.	Are we covering all relevant dimensions?	Is the measurement comprehensive?	A usability questionnaire includes items on aesthetics and performance but omits learnability or error recovery.	Omission of key aspects of the construct being measured.	Thorough content analysis, expert review, item comprehensiveness.	Comprehensive item coverage, domain mapping
Face Validity	The extent to which a measure appears, on the surface, to assess what it claims to measure.	Does it look like it measures what it claims to?	Credibility and transparency to participants and stakeholders.	A scale titled 'User Satisfaction' asks about navigation, layout, and visual appealâ€”participants can relate the questions to the label.	Ambiguous or misaligned items that reduce user trust or engagement.	Clear alignment between items and construct, user involvement.	Transparent alignment between measure and construct
Ecological Validity	The degree to which the study setting, task, and context resemble the real world.	Does the context resemble real use?	Does the study simulate realistic use conditions?	A driving AI evaluated on a static simulator lacks realism compared to real road testing.	Unrealistic settings, tasks, or context for intended use.	Naturalistic settings, context-aware design, field evaluation.	Realistic tasks, environments, and interaction contexts
Statistical Conclusion Validity	The extent to which the statistical analysis is appropriate, and the results are not due to chance, insufficient power, or flawed assumptions.	Are the findings statistically reliable and valid?	Are your conclusions based on solid statistics?	You fail to detect an effect because the study is underpowered, even though the effect exists.	Low power, assumption violations, misuse of statistics.	Power analysis, correct statistical tests, assumption checks.	Appropriate statistical methods, power, assumption checks
Conclusion Validity	The overall credibility of the interpretation of the data.	Do the data support the conclusions?	Are the conclusions drawn justified by the evidence?	All data are consistent, but interpretation exaggerates what the evidence actually supports.	Overinterpretation, cherry-picking, ignoring limitations.	Conservative interpretation, triangulation, transparency about limits.	Careful interpretation, integration of multiple forms of evidence

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
Level 3: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving prospective real-world cases considered by real practitioners in real-world settings.
Level 4: Single experimental study (randomized, controlled) with retrospective real-world cases considered by real practitioners in simulated/laboratory settings.
Level 5: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving retrospective real-world cases considered by real practitioners in simulated/laboratory settings
Level 6: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by real practitioners in simulated/ laboratory settings.
Level 7: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by human participants but not real practitioners in laboratory settings
Level 8: Supervised machine learning train/test studies with external validation (multiple datasets in longitudinal or cross-section/multi-site settings)
Level 9 Supervised machine learning train/test studies with internal validation
Level 10: Consensus opinions of authoritative bodies (e.g., nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 11 (weakest): Opinions of recognized experts and case studies

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
Level 3: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving prospective real-world cases considered by real practitioners in real-world settings.
Level 4: Single experimental study (randomized, controlled) with retrospective real-world cases considered by real practitioners in simulated/laboratory settings.
Level 5: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving retrospective real-world cases considered by real practitioners in simulated/laboratory settings
Level 6: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by real practitioners in simulated/ laboratory settings.
Level 7: Single quasi-experimental study (e.g., nonrandomized, with concurrent or historical controls) involving simulated cases considered by human participants but not real practitioners in laboratory settings
Level 8: Supervised machine learning train/test studies with external validation (multiple datasets in longitudinal or cross-section/multi-site settings)
Level 9 Supervised machine learning train/test studies with internal validation
Level 10: Consensus opinions of authoritative bodies (e.g., nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 11 (weakest): Opinions of recognized experts and case studies

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

Aspect	Experimental (True)	Quasi-Experimental
Random Assignment	Yes	No
Causal Inference	Strong	Moderate to weak
Internal Validity	High	Lower (confound risk)
Manipulation of IV	Yes	Yes
Control of Confounding	Through randomization	Through statistical or design controls
Typical Use	Controlled lab studies, A/B tests	Field studies, natural experiments

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners

Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.

ed, with concurrent or
nsidered by real practi-

with retrospective real-
boratory settings.

nized, with concurrent
ses considered by real

nized, with concurrent
by real practitioners in

ed, with concurrent or
uman participants but

ith external validation
settings)

internal validation

nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)

Level 11 (weakest): Opinions of recognized experts and case studies

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners

Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.

ed, with concurrent or
nsidered by real practi-

with retrospective real-
oratory settings.

nized, with concurrent
ses considered by real

nized, with concurrent
by real practitioners in

guideline groups with robust peer r
organizations)

Level 11 (weakest): Opinions of re

Aspect	Experimental (True)	Quasi-Experimental
Random Assignment	Yes	No
Causal Inference	Strong	Moderate to weak
Internal Validity	High	Lower (confound risk)
Manipulation of IV	Yes	Yes
Control of Confounding	Through randomization	Through statistical or
Typical Use	Controlled lab studies, A/B tests	Field studies, natural

In research, manipulating an independent variable means a researcher actively changes or alters a specific factor (the independent variable) to see how it affects another factor (the dependent variable). This manipulation is done systematically, allowing researchers to observe any causal relationships between the independent and dependent variables. In other words: **the researcher doesn't just observe what naturally occurs; they actively introduce changes to the independent variable.**

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

Aspect	Experimental (True)	Quasi-Experimental
Random Assignment	Yes	No
Control of Confounding	Through randomization	Through statistical or design controls
Typical Use	Controlled lab studies, A/B tests	Field studies, natural experiments

Concept	Definition	Risk to Study Validity
Random Assignment	Equal and independent chance for all participants to be in any group	Ensures internal validity
Non-Random Assignment (e.g., birthdate)	Assignment using a systematic, non-random rule	Risk of unbalanced groups

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
ed, with concurrent or nsidered by real practi-
with retrospective real- oratory settings.
ith external validation settings)
internal validation
nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 11 (weakest): Opinions of recognized experts and case studies

LADDER OF THE LEVELS OF STRENGTH OF EVIDENCE

Level 1: Meta-analyses and systematic reviews of randomized controlled trials or experimental studies involving real practitioners
Level 2: Single experimental study (randomized, controlled) with prospective real-world cases considered by real practitioners in real-world settings.
...ed, with concurrent or ...nsidered by real practi-
...with retrospective real- ...boratory settings.
...nationally recognized guideline groups with robust peer review processes, notified bodies, standardization organizations)
Level 11 (weakest): Opinions of recognized experts and case studies

Aspect	Experimental (True)	Quasi-Experimental
Random Assignment	Yes	No

Concept	Definition	Risk to Study Validity
Random Assignment	Equal and independent chance for all participants to be in any group	Ensures internal validity
Non-Random Assignment (e.g., birthdate)	Assignment using a systematic, non-random rule	Risk of unbalanced groups
Selection Bias	Groups differ systematically at baseline due to non-random assignment	Threat to causal inference
Attrition	Participants drop out unevenly across groups	Biased estimates, reduced generalizability

Bias / Threat

Selection Bias

Attrition Bias

Sampling Bias

Measurement Bias

Hawthorne Effect

Placebo Effect

Experimenter Bias

Confirmation Bias

Maturation

History Effects

Regression to the Mean

Testing Effect

Instrumentation Effects

Ecological Validity Threat

Demand Characteristics

Nonresponse Bias

Human-centered Evidence-based design: the process of basing design decisions about the system on credible user research to make the system more usable [than others, or than it was earlier].

The more valid the research, the more credible it is, the lower the risk of bias or detrimental effects, the more valid the research is.

Bias / Threat	Definition	Effect on Validity	Typical Source	Example in HCI / Human-AI Context
Selection Bias	Systematic differences between groups at baseline due to non-random assignment	**Internal validity**	Self-selection, group assignment by rule	More expert users self-select into the more advanced version of an AI tool
Attrition Bias	Uneven or non-random dropout of participants during the study	**Internal + external validity**	Boredom, difficulty, lack of incentive	Users leave a trial of an AI assistant if they find it too opaque, skewing post-test results
Sampling Bias	Study participants are not representative of the target population	**External validity**	Convenience sampling, online-only recruitment	Only recruiting students or tech workers to test a health app intended for older adults
Measurement Bias	Systematic error in how variables are measured or recorded	**Internal validity**	Poor instrumentation, biased survey design	A trust-in-AI scale is misunderstood by non-native speakers
Hawthorne Effect	Participants alter behavior simply because they know they are being observed	**Internal validity**	Laboratory or observed settings	Users explore features more thoroughly during a usability session than they would in real life
Placebo Effect	Participants show improvement because they believe they are receiving a better or "new" condition	**Internal validity**	Belief in system superiority	Users report higher satisfaction with an AI system labeled “intelligent” even if functionality is equal
Experimenter Bias	Researchers unintentionally influence participants or data interpretation	**Internal validity**	Leading questions, subtle cues, interpretive bias	Interviewer nods approvingly when participants praise the AI interface
Confirmation Bias	Tendency to seek or interpret data in a way that confirms prior beliefs	**Internal validity**	Researcher expectations	An evaluator interprets ambiguous user behavior as validating a design hypothesis
Maturation	Participants change over time due to natural development or fatigue	**Internal validity**	Longitudinal studies	Users become more proficient over time regardless of interface changes
History Effects	External events affect participants during the study	**Internal validity**	Concurrent events outside the experiment	A high-profile AI failure in the news affects trust scores during an ongoing study
Regression to the Mean	Extreme scores tend to move closer to the average on subsequent testing	**Internal validity**	Pre/post-test designs without control group	Users with poor performance initially show improvement even without any system change
Testing Effect	Repeated testing influences participants’ performance	**Internal validity**	Pre-tests, learning effects	Users improve task performance in usability test session 2 due to prior exposure
Instrumentation Effects	Changes in measurement tools or conditions between phases	**Internal validity**	Interface, task, or evaluator changes	Task complexity or measurement criteria differ between A and B groups
Ecological Validity Threat	Lab setting does not reflect real-world context	**External validity**	Highly controlled artificial environments	A chat interface tested in a lab does not account for distractions present in mobile use
Demand Characteristics	Participants infer the purpose of the study and alter behavior accordingly	**Internal validity**	Transparent study design, explicit instructions	Participants overreport trust in AI because they think that’s what the researchers expect
Nonresponse Bias	Certain types of participants fail to respond or participate	**External validity**	Low response rates, selective engagement	Only highly motivated users give feedback on the AI chatbot, skewing satisfaction results

Bias / Threat	Definition	Effect on Validity	Typical Source	Example in HCI / Human-AI Context
Selection Bias	Systematic differences between groups at baseline due to non-random assignment	Internal validity	Self-selection, group assignment by rule	More expert users self-select into the more advanced version of an AI tool
Attrition Bias	Uneven or non-random dropout of participants during the study	Internal + external validity	Boredom, difficulty, lack of incentive	Users leave a trial of an AI assistant if they find it too opaque, skewing post-test results
Sampling Bias	Study participants are not representative of the target population	External validity	Convenience sampling, online-only recruitment	Only recruiting students or tech workers to test a health app intended for older adults
Measurement Bias	Systematic error in how variables are measured or recorded	Internal validity	Poor instrumentation, biased survey design	A trust-in-AI scale is misunderstood by non-native speakers
Hawthorne Effect	Participants alter behavior simply because they know they are being observed	Internal validity	Laboratory or observed settings	Users explore features more thoroughly during a usability session than they would in real life
Placebo Effect	Participants show improvement because they believe they are receiving a better or "new" condition	Internal validity	Belief in system superiority	Users report higher satisfaction with an AI system labeled “intelligent” even if functionality is equal
Experimenter Bias	Researchers unintentionally influence participants or data interpretation	Internal validity	Leading questions, subtle cues, interpretive bias	Interviewer nods approvingly when participants praise the AI interface
Confirmation Bias	Tendency to seek or interpret data in a way that confirms prior beliefs	Internal validity	Researcher expectations	An evaluator interprets ambiguous user behavior as validating a design hypothesis
Maturation	Participants change over time due to natural development or fatigue	Internal validity	Longitudinal studies	Users become more proficient over time regardless of interface changes
History Effects	External events affect participants during the study	Internal validity	Concurrent events outside the experiment	A high-profile AI failure in the news affects trust scores during an ongoing study
Regression to the Mean	Extreme scores tend to move closer to the average on subsequent testing	Internal validity	Pre/post-test designs without control group	Users with poor performance initially show improvement even without any system change
Testing Effect	Repeated testing influences participants’ performance	Internal validity	Pre-tests, learning effects	Users improve task performance in usability test session 2 due to prior exposure
Instrumentation Effects	Changes in measurement tools or conditions between phases	Internal validity	Interface, task, or evaluator changes	Task complexity or measurement criteria differ between A and B groups
Ecological Validity Threat	Lab setting does not reflect real-world context	External validity	Highly controlled artificial environments	A chat interface tested in a lab does not account for distractions present in mobile use
Demand Characteristics	Participants infer the purpose of the study and alter behavior accordingly	Internal validity	Transparent study design, explicit instructions	Participants overreport trust in AI because they think that’s what the researchers expect
Nonresponse Bias	Certain types of participants fail to respond or participate	External validity	Low response rates, selective engagement	Only highly motivated users give feedback on the AI chatbot, skewing satisfaction results

Bias / Threat	Definition	Effect on Validity	Typical Source	Example in HCI / Human-AI Context
Selection Bias	Systematic differences between groups at baseline due to non-random assignment	**Internal validity**	Self-selection, group assignment by rule	More expert users self-select into the more advanced version of an AI tool
Attrition Bias	Uneven or non-random dropout of participants during the study	**Internal + external validity**	Boredom, difficulty, lack of incentive	Users leave a trial of an AI assistant if they find it too opaque, skewing post-test results
Sampling Bias	Study participants are not representative of the target population	**External validity**	Convenience sampling, online-only recruitment	Only recruiting students or tech workers to test a health app intended for older adults
Measurement Bias	Systematic error in how variables are measured or recorded	**Internal validity**	Poor instrumentation, biased survey design	A trust-in-AI scale is misunderstood by non-native speakers
Hawthorne Effect	Participants alter behavior simply because they know they are being observed	**Internal validity**	Laboratory or observed settings	Users explore features more thoroughly during a usability session than they would in real life
Placebo Effect	Participants show improvement because they believe they are receiving a better or "new" condition	**Internal validity**	Belief in system superiority	Users report higher satisfaction with an AI system labeled “intelligent” even if functionality is equal
Experimenter Bias	Researchers unintentionally influence participants or data interpretation	**Internal validity**	Leading questions, subtle cues, interpretive bias	Interviewer nods approvingly when participants praise the AI interface
Confirmation Bias	Tendency to seek or interpret data in a way that confirms prior beliefs	**Internal validity**	Researcher expectations	An evaluator interprets ambiguous user behavior as validating a design hypothesis
Maturation	Participants change over time due to natural development or fatigue	**Internal validity**	Longitudinal studies	Users become more proficient over time regardless of interface changes
History Effects	External events affect participants during the study	**Internal validity**	Concurrent events outside the experiment	A high-profile AI failure in the news affects trust scores during an ongoing study
Regression to the Mean	Extreme scores tend to move closer to the average on subsequent testing	**Internal validity**	Pre/post-test designs without control group	Users with poor performance initially show improvement even without any system change
Testing Effect	Repeated testing influences participants’ performance	**Internal validity**	Pre-tests, learning effects	Users improve task performance in usability test session 2 due to prior exposure
Instrumentation Effects	Changes in measurement tools or conditions between phases	**Internal validity**	Interface, task, or evaluator changes	Task complexity or measurement criteria differ between A and B groups
Ecological Validity Threat	Lab setting does not reflect real-world context	**External validity**	Highly controlled artificial environments	A chat interface tested in a lab does not account for distractions present in mobile use
Demand Characteristics	Participants infer the purpose of the study and alter behavior accordingly	**Internal validity**	Transparent study design, explicit instructions	Participants overreport trust in AI because they think that’s what the researchers expect
Nonresponse Bias	Certain types of participants fail to respond or participate	**External validity**	Low response rates, selective engagement	Only highly motivated users give feedback on the AI chatbot, skewing satisfaction results

Bias / Threat	Definition	Effect on Validity	Typical Source	Example in HCI / Human-AI Context
Selection Bias	Systematic differences between groups at baseline due to non-random assignment	**Internal validity**	Self-selection, group assignment by rule	More expert users self-select into the more advanced version of an AI tool
Attrition Bias	Uneven or non-random dropout of participants during the study	**Internal + external validity**	Boredom, difficulty, lack of incentive	Users leave a trial of an AI assistant if they find it too opaque, skewing post-test results
Sampling Bias	Study participants are not representative of the target population	**External validity**	Convenience sampling, online-only recruitment	Only recruiting students or tech workers to test a health app intended for older adults
Measurement Bias	Systematic error in how variables are measured or recorded	**Internal validity**	Poor instrumentation, biased survey design	A trust-in-AI scale is misunderstood by non-native speakers
Hawthorne Effect	Participants alter behavior simply because they know they are being observed	**Internal validity**	Laboratory or observed settings	Users explore features more thoroughly during a usability session than they would in real life
Placebo Effect	Participants show improvement because they believe they are receiving a better or "new" condition	**Internal validity**	Belief in system superiority	Users report higher satisfaction with an AI system labeled “intelligent” even if functionality is equal
Experimenter Bias				
Confirmation Bias				
Maturation				
History Effects				
Regression to the Me				
Testing Effect				
Instrumentation Effects	Changes in measurement tools or conditions between phases	**Internal validity**	Interface, task, or evaluator changes	Task complexity or measurement criteria differ between A and B groups
Ecological Validity Threat	Lab setting does not reflect real-world context	**External validity**	Highly controlled artificial environments	A chat interface tested in a lab does not account for distractions present in mobile use
Demand Characteristics	Participants infer the purpose of the study and alter behavior accordingly	**Internal validity**	Transparent study design, explicit instructions	Participants overreport trust in AI because they think that’s what the researchers expect
Nonresponse Bias	Certain types of participants fail to respond or participate	**External validity**	Low response rates, selective engagement	Only highly motivated users give feedback on the AI chatbot, skewing satisfaction results

1. Design phase threats (e.g., selection bias, sampling bias)
2. Execution phase threats (e.g., attrition, non-response, maturation)
3. Interpretation phase threats (e.g., confirmation bias)

Bias / Threat	Definition	Effect on Validity	Typical Source	Example in HCI / Human-AI Context
Selection Bias	Systematic differences between groups at baseline due to non-random assignment	**Internal validity**	Self-selection, group assignment by rule	More expert users self-select into the more advanced version of an AI tool
Attrition Bias	Uneven or non-random dropout of participants during the study	**Internal + external validity**	Boredom, difficulty, lack of incentive	Users leave a trial of an AI assistant if they find it too opaque, skewing post-test results
Sampling Bias	Study participants are not representative of the target population	**External validity**	Convenience sampling, online-only recruitment	Only recruiting students or tech workers to test a health app intended for older adults
Measurement Bias	Systematic error in how variables are measured or recorded	**Internal validity**	Poor instrumentation, biased survey design	A trust-in-AI scale is misunderstood by non-native speakers
Hawthorne Effect	Participants alter behavior simply because they know they are being observed	**Internal validity**	Laboratory or observed settings	Users explore features more thoroughly during a usability session than they would in real life
Placebo Effect	Participants show improvement because they believe they are receiving a better or "new" condition	**Internal validity**	Belief in system superiority	Users report higher satisfaction with an AI system labeled “intelligent” even if functionality is equal
Experimenter Bias				
Confirmation Bias				
Maturation				
History Effects				
Regression to the Me				
Testing Effect				
Instrumentation Effec	Changes in measurement tools or			Task complexity or measurement criteria differ between
Ecological Validity Th				
Demand Characterist	and alter behavior accordingly			that's what the researchers expect
Nonresponse Bias	Certain types of participants fail to respond or participate	**External validity**	Low response rates, selective engagement	Only highly motivated users give feedback on the AI chatbot, skewing satisfaction results

1. Design phase threats (e.g., selection bias, sampling bias)
2. Execution phase threats (e.g., attrition, non-response, maturation)
3. Interpretation phase threats (e.g., confirmation bias)

All biases threatens **validity**.

Discussion: what you expect and what it would mean

Anticipated results

- Summarise expected patterns (not detailed numbers).
- Relate to prior work: confirm, extend, or contradict.
- State what result would be theoretically informative.

Limitations (and mitigation)

- Constraints of design and sampling.
- Threats to validity and how you reduce them.
- What your study cannot conclude.

Implications & contributions

- HCI theory: what construct, mechanism, or boundary condition?
- Practice: design guidance, evaluation metrics, stakeholder impact.
- Societal impact: fairness, safety, accessibility, wellbeing, etc.
- Future work: concrete follow-ups enabled by this protocol.

A strong discussion section makes clear what *changes* if the study succeeds—or fails.

References

- Use APA 7th edition consistently.
- Cite key theory and the empirical foundation of your design.
- Prefer primary sources (original methods, validated scales).
- Prefer journals (Q1 & Q2) or A, A*, B conferences
- Consider a reference manager (e.g., Zotero, Mendeley).

Appendices (when needed)

- Full instructions (verbatim).
- Questionnaires, stimuli examples, interface layouts.
- Consent form + data protection information.
- Any complex procedure diagram or codebook.

If you later deviate from the protocol, document deviations transparently.

References

- Use APA 7th edition consistently.
- Cite key theory and the empirical foundation of your design.
- Prefer primary sources (original methods, validated scales).
- Prefer journals (Q1 & Q2) or A, A*, B conferences
- Consider a reference manager (e.g., Zotero, Mendeley).

Appendices (when needed)

- Full instructions (verbatim).
- Questionnaires, stimuli examples, interface layouts.
- Consent form + data protection information.
- Any complex procedure diagram or codebook.

If you later deviate from the protocol, document deviations transparently.

References

- Use APA 7th edition consistently.
- Cite key theory and the empirical foundation of your design.
- Prefer primary sources (original methods, validated scales).
- Consider a reference manager (e.g., Zotero, Mendeley).

Appendices (when needed)

- Full instructions (verbatim).
- Questionnaires, stimuli examples, interface layouts.
- Consent form + data protection information.
- Any complex procedure diagram or codebook.

Before you submit the protocol, verify:

1. Every RQ has measures + an analysis plan (and vice versa).
2. Ethics: consent, risks, data protection, and approvals are explicit.
3. Procedure includes timing, randomisation, and condition definitions.
4. Materials are reproducible (versions, sources, appendices).
5. Sample size has a justification and an plan against participant drop-out and missing data.

One tool, a checklist, lots of implications on classifier performance and utility!

https://www.entechne.com/DiagrAIms

DiagrAIms

https://zenodo.org/record/6451243

https://www.entechne.com/chamaichecklist/

THANK YOU!

federico.cabitza@unimib.it

federicocabitza

www.federicocabitza.net