

# «Crash Course» Empirical Side of Responsible AI

Provided by ENKIS  
Human-Centered Computing (HCC), Institute of Computer Science  
Freie Universität Berlin



With funding from the:



Federal Ministry  
of Research, Technology  
and Space



Finanziert von der  
Europäischen Union  
NextGenerationEU

Senatsverwaltung  
für Wissenschaft, Gesundheit,  
Pflege und Gleichstellung

**BERLIN**



# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

# How to carry out Experimental Research?

## Handout

### 1. Research Questions

Experiment research questions do not only ask, whether a relationship between two variables exists, but also aims at revealing the underlying cause by investigating causality.

**Examples:** "How does display size affect user satisfaction?", "How does text length affect user comprehension?"

### 2. Hypotheses

A hypothesis defines both the variables involved and the relationship between them. For example, A causes B; A is larger, faster, or more enjoyable than B; etc.

**Examples:** "A larger display leads to better performance.", "Longer texts lead to less user comprehension."

### 3. Defining needed variables

- We provide handouts.
- Later on you will work on your own scenario.
- Taking notes might help you!

# Running Example

Before conducting a test you need to ask yourself questions:

- What exactly is the problem?
- What is the real world scenario?
- Who are your potential end-users?
- What intervention/idea do you want to use?



# Running Example

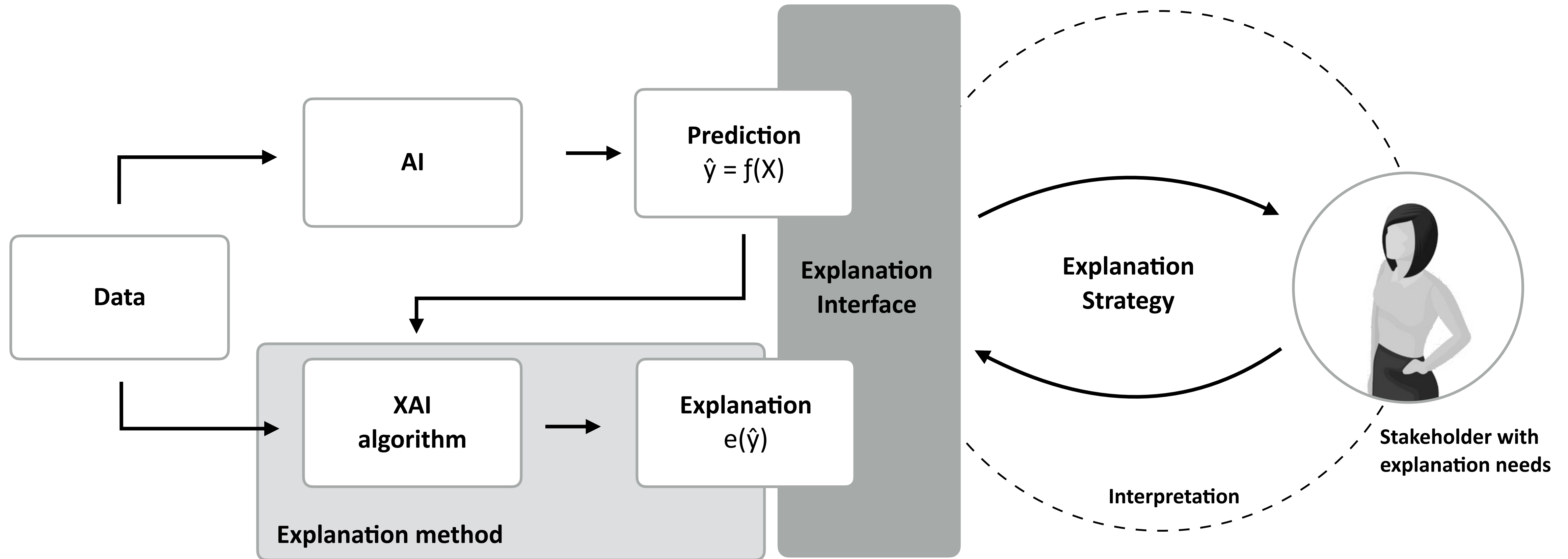
Before conducting a test you need to ask yourself questions:

- What exactly is the problem? —> Improving Human-AI Collaboration
- What is the real world scenario? —> Digital Healthcare, Online Treatment planing
- Who are your potential end-users? —> Trained lay people, Home Office
- What intervention/idea do you want to test? —> Showing explanations (comparing)

# Group comparison: A/B/C Testing

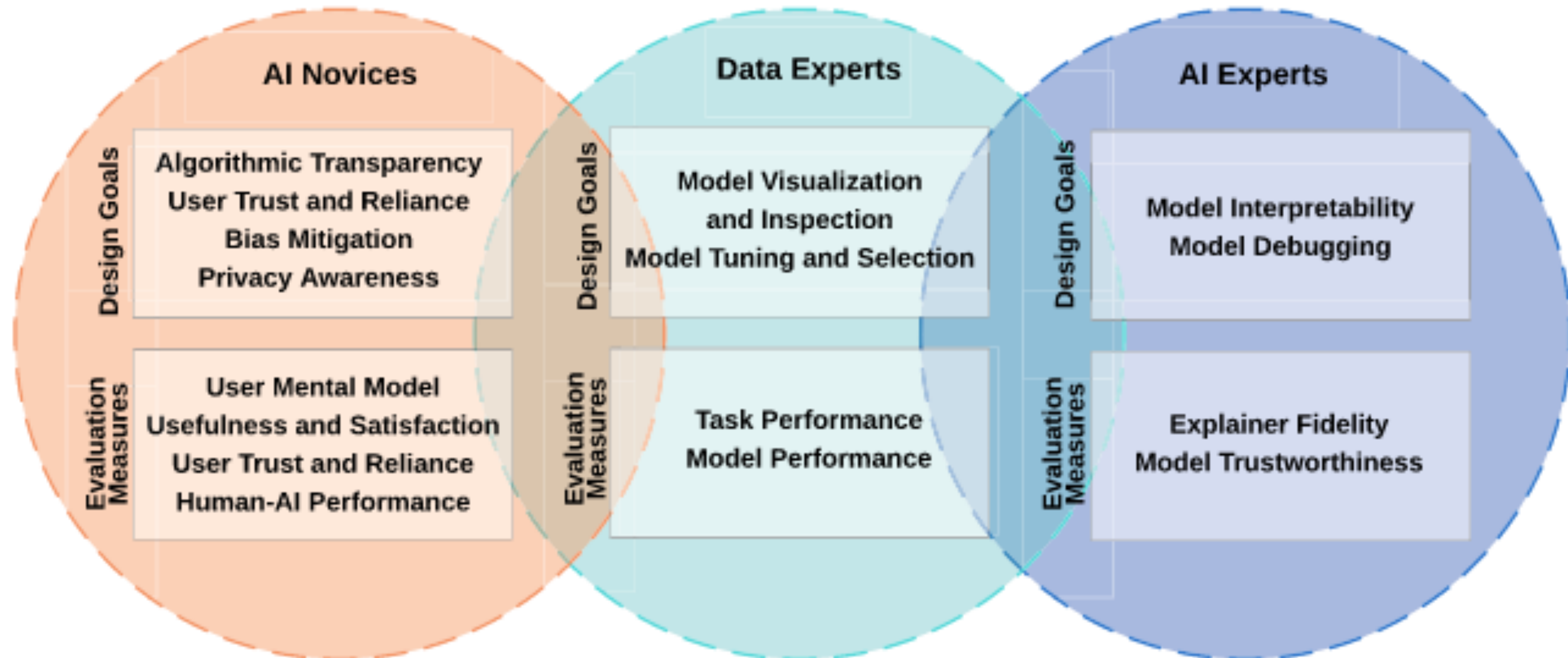


# Running Example





# Who are your potential end-users?





# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

# Research Questions in Experimental Research

Experiment research questions do not only ask, whether a relationship between two variables exists, but also aims at revealing the underlying cause by investigating causality.

Hence, experimental research answers cause-and-effect questions about the relationship between two variables

## Running example:

“How can we explain how understandable our model output is?”

“What is the relationship between task immersion and display size?”

“How does the display size affect task immersion?”

# Research Questions in Experimental Research

Experiment research questions do not only ask, whether a relationship between two variables exists, but also aims at revealing the underlying cause by investigating causality.

Hence, experimental research answers cause-and-effect questions about the relationship between two variables

## Running example:

“What explanation helps best, when interacting with an AI to diagnose?”

“How do the explanations when interacting with an AI to diagnose influence performance?”



# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

# Hypothesis Formulation

Experimental research begins with the development of a statement regarding the predicted cause-and-effect relationship between two variables.

This is known as a **research hypothesis**.

In general, hypotheses clarify and clearly articulate what it is the researcher is aiming to understand.

A hypothesis defines both the variables involved and the relationship between them. For example, A causes B; A is larger, faster, or more enjoyable than B; etc.

# Hypothesis Formulation

Experimental research begins with the development of a statement regarding the predicted cause-and-effect relationship between two variables.

This is known as a **research hypothesis**.

In general, hypotheses clarify and clearly articulate what it is the researcher is aiming to understand.

## Running example:

“Showing **explanation A** leads to **higher performance** than showing **explanation B**.”



# How to carry out Experimental Research?

Research Question

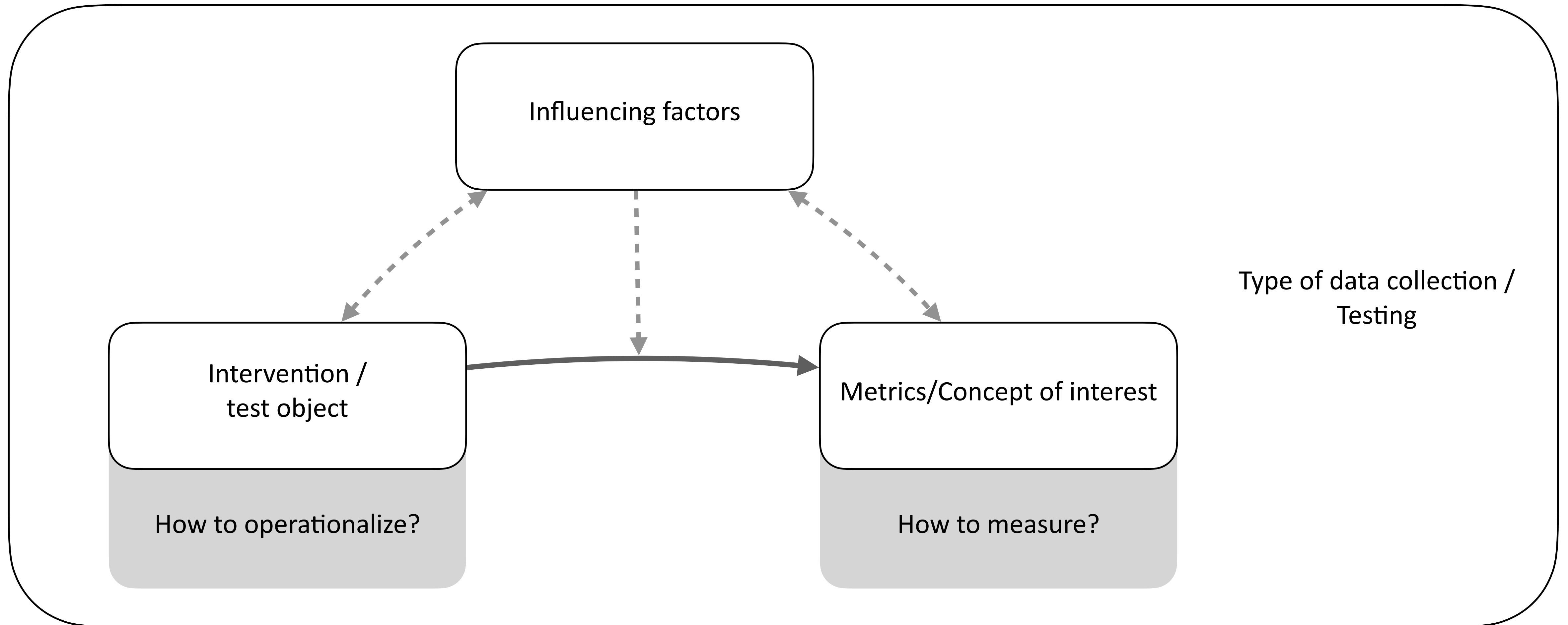
Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

# Defining Needed Variables



# Defining Needed Variables - Operationalization

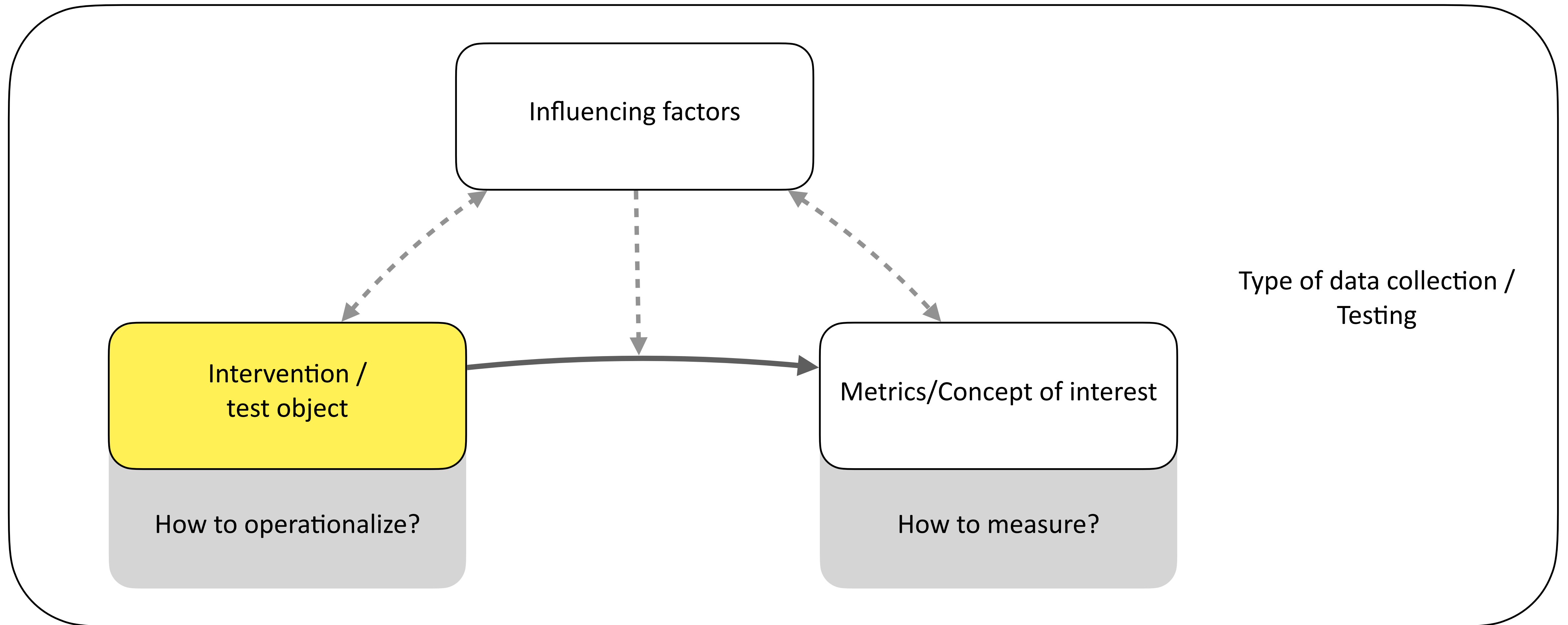
Defining the variables, by translating theoretical concepts into recordable variables

The choice of the right variables can make or break an experiment, hence these need be carefully tested before running an experiment.

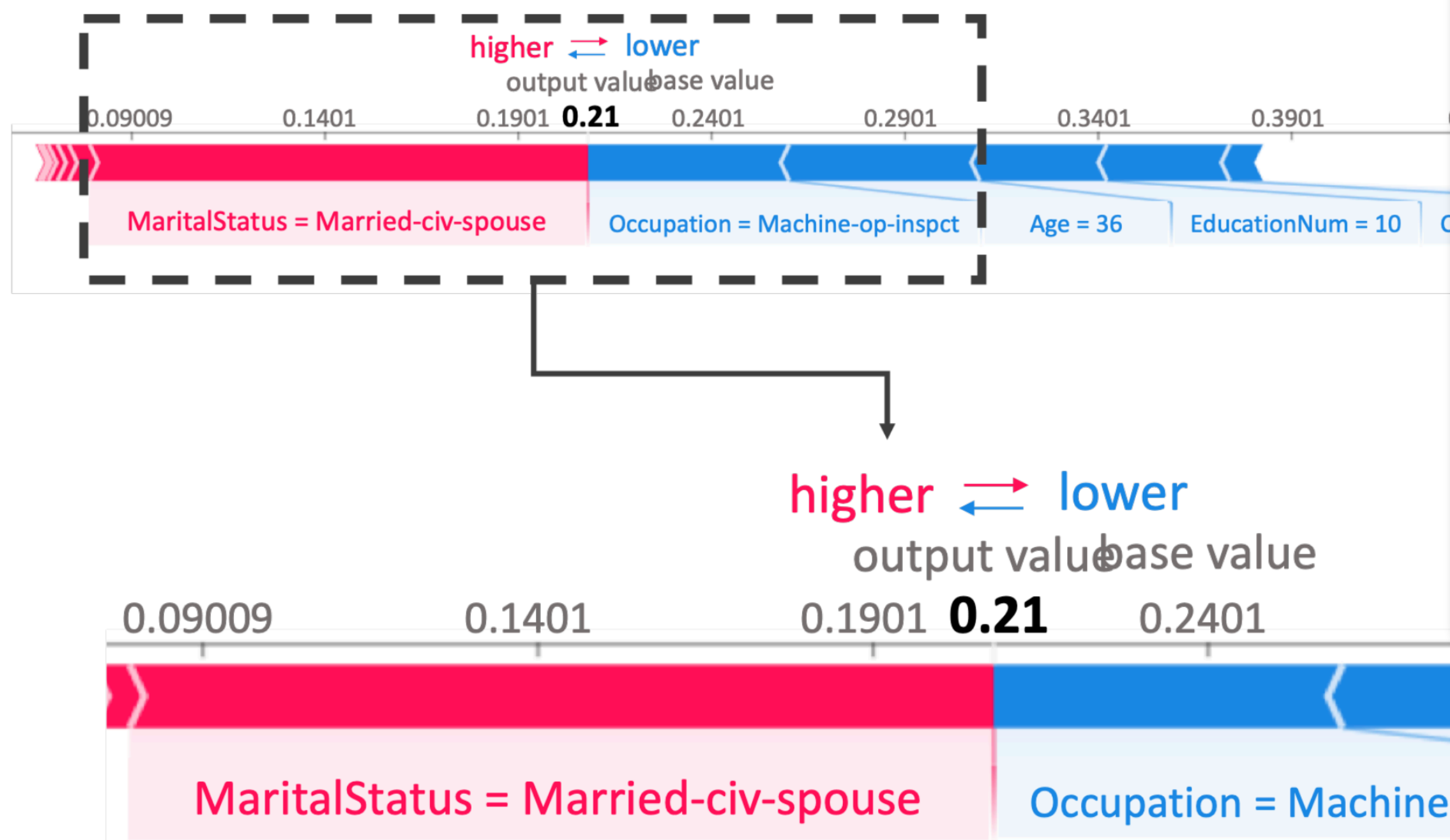
- The following questions guide the operationalization of the two main variables: the **dependent variable** and the *independent variable*.
  - **Manipulation:** How can the independent variable **be changed between the experimental conditions?**
  - **Measurement:** How can a change in the dependent variable be measured?



# Defining Needed Variables



# Defining Needed Variables - Human-AI Interventions

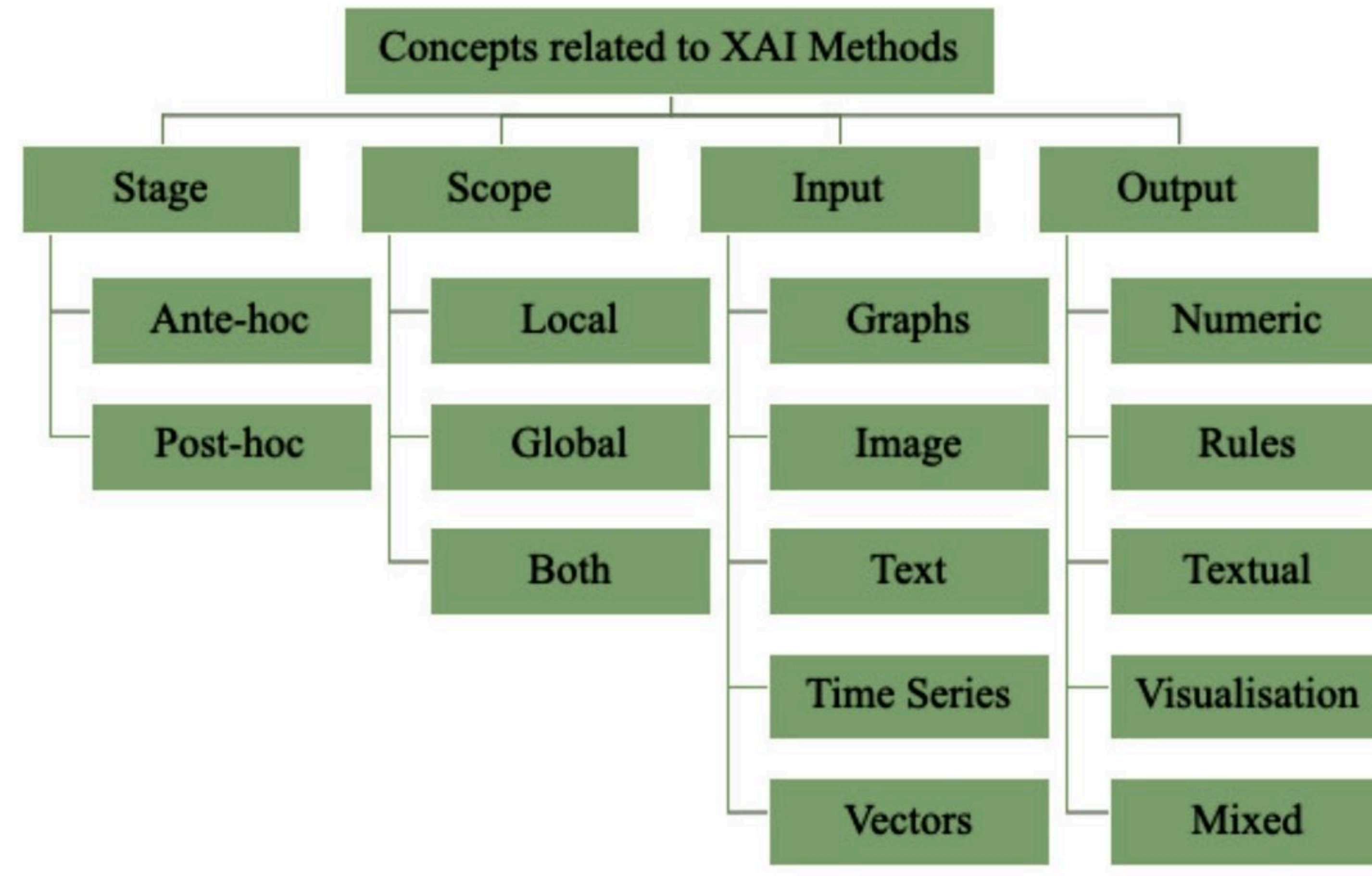


Although interpretability tools are meant to help data scientists understand how ML models work, **some participants used the tools to rationalize suspicious observations instead.** After conducting several exploratory tests on the dataset, P8 said “Test of means says the same thing as SHAP about Age. All’s good!” (P8, SHAP), and gave confidence ratings of 7 (extremely).

In contrast, two participants **under-used the tools** because they did not provide explanations with the content or clarity that they expected. P7 noted that “This is not an explanation system. It’s a visualization. There was no interpretation provided here” (P7, GAMs). Similarly, P4 became skeptical when they did not fully understand how SHAP’s importance scores values were being calculated, eventually leading to disuse [17, 52]:



# Defining Needed Variables - Human-AI Interventions





# Which interventions did we use?



# Which interventions did we use?



Clinical Guideline

It is a **severe** nail fungus case if ...

- the nail bed is affected
- or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...

- the nail bed is not affected
- and
- the nail is less than 50% affected

The AI classifies:

**Mild**

Patient ID: XXX



Clinical Guideline

It is a **severe** nail fungus case if ...

- the nail bed is affected
- or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...

- the nail bed is not affected
- and
- the nail is less than 50% affected

The AI classifies:

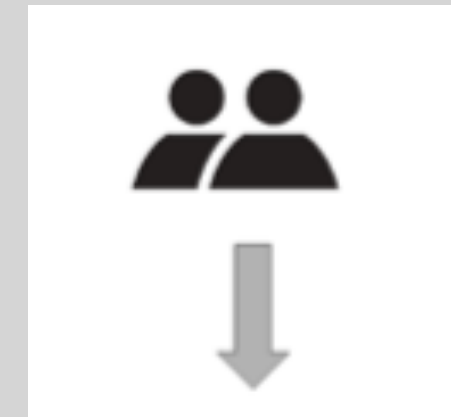
**Mild**

Patient ID: XXX

In 100 samples like this, AI would predict ...

... 75 to be **severe** (75%).

... 25 to be **mild** (25%).



Clinical Guideline

It is a **severe** nail fungus case if ...

- the nail bed is affected
- or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...

- the nail bed is not affected
- and
- the nail is less than 50% affected

The AI classifies:

**Mild**

Patient ID: XXX

In 100 samples like this, AI would predict ...

... 75 to be **severe** (75%).

... 25 to be **mild** (25%).

Capabilities

⊖ AI limits

Unable to consider a wider context.

Inflexible with untypical, low quality, or rotated images

⊕ Human strengths

Able to consider a wider context.

Can adapt based on new information.

⊖ Human limits

Less experience with nail fungus images.

Affected by surroundings.

⊕ AI strengths

Trained on many images chosen by experts.

Unaffected by surroundings.

# Which interventions did we use?



## Uncertainty / Confidence depiction

- XAI method
- Local Explanation
- Static
- Understanding the model
- Text and visualisation

# Which interventions did we use?



## Guidance

- Explanation in a wider sense
- Educating the User



# Which interventions did we use?



## Guidance

- Explanation in a wider sense
- Educating the User

# How to carry out Experimental Research?

Research Question

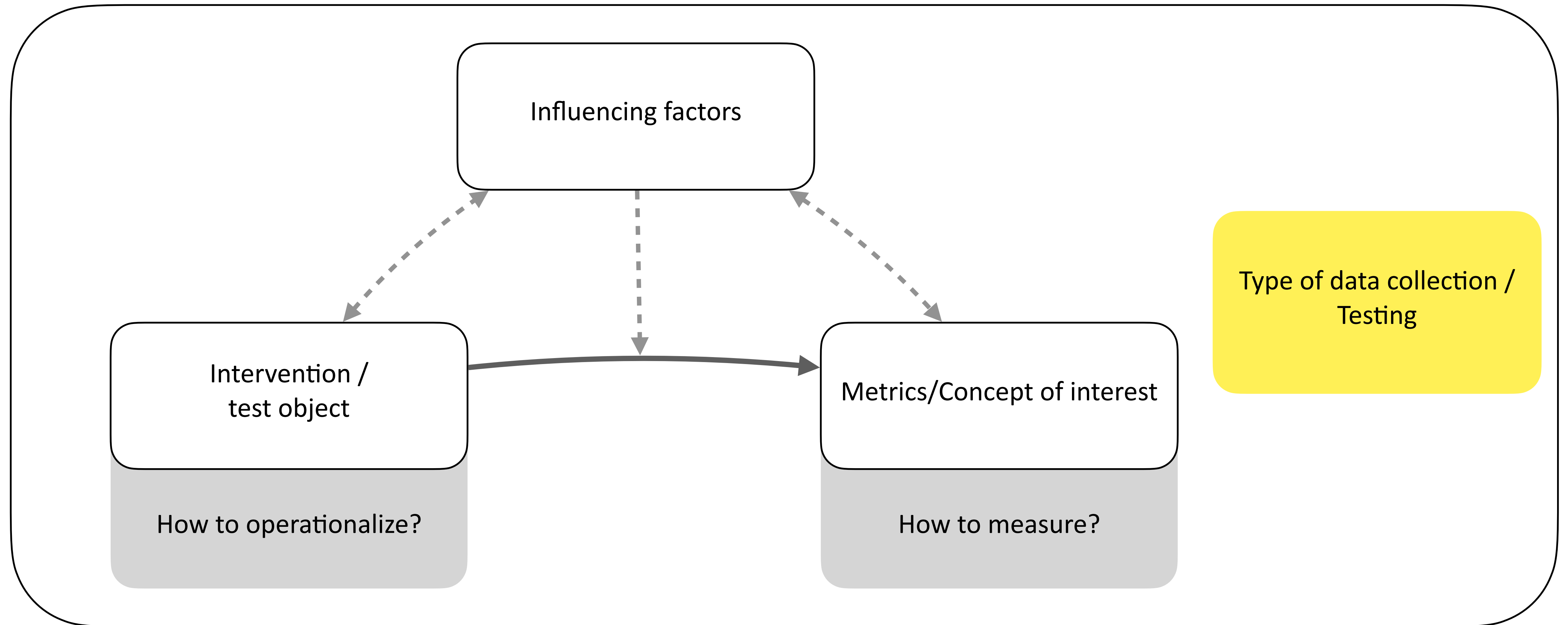
Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

# Specifying Your Research Design



# Specifying Your Research Design - Experimental Design

## User Testing

- » Improve products
- » Few participants
- » Results inform design
- » Usually not completely replicable
- » Conditions controlled as much as possible
- » Procedure planned
- » Results reported to developers

## Experiments for Research

- » Discover knowledge
- » Many participants
- » Results validated statistically
- » Must be replicable
- » Strongly controlled conditions
- » Experimental design
- » Scientific report to scientific community



# Specifying Your Research Design - Experimental Design

There are two basic research designs, for obtaining different treatment groups:

- **Within-subjects design:** Getting measurements from the participant group before and after receiving the treatment



- **Between subjects design:** Differing the treatment between two participants groups



# Specifying Your Research Design - Experimental Design

## Between-subjects design

---

- When there are small individual differences, but large expected differences across conditions
  - When learning and carryover effects are likely to influence performance
  - When fatigue may be an issue
- 

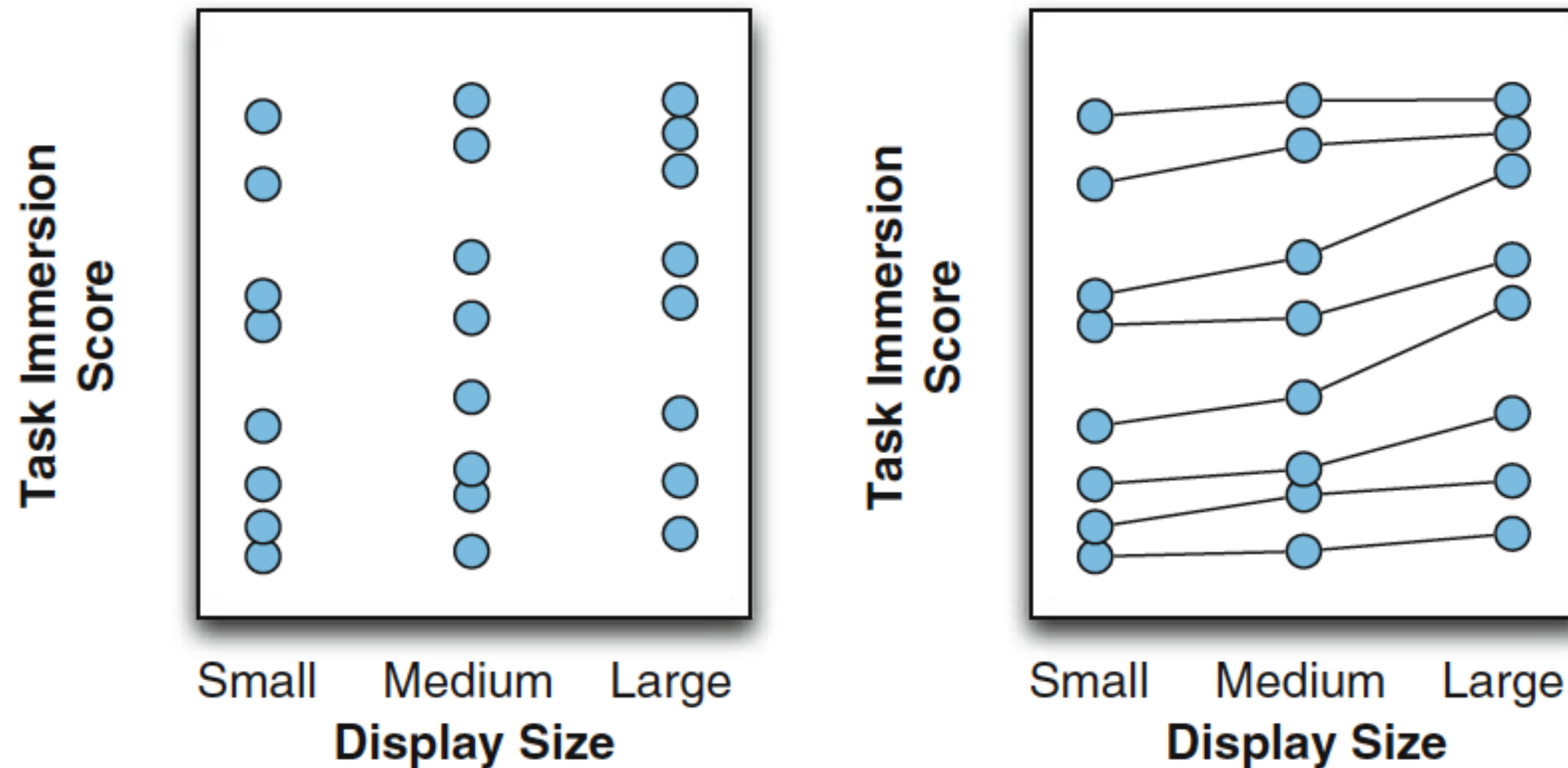
## Within-subjects design

---

- When there are large individual differences (i.e., high variance across participants with respect to the dependent variable(s) of interest)
  - When tasks are unlikely to be affected by learning and carryover effects are unlikely to occur
  - When working with rare or hard to reach populations
-



# Specifying Your Research Design - Experimental Design



Example  
subjects design (left) when there are large individual differences in participants' scores

compared to a between-

# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

Defining Needed Variables

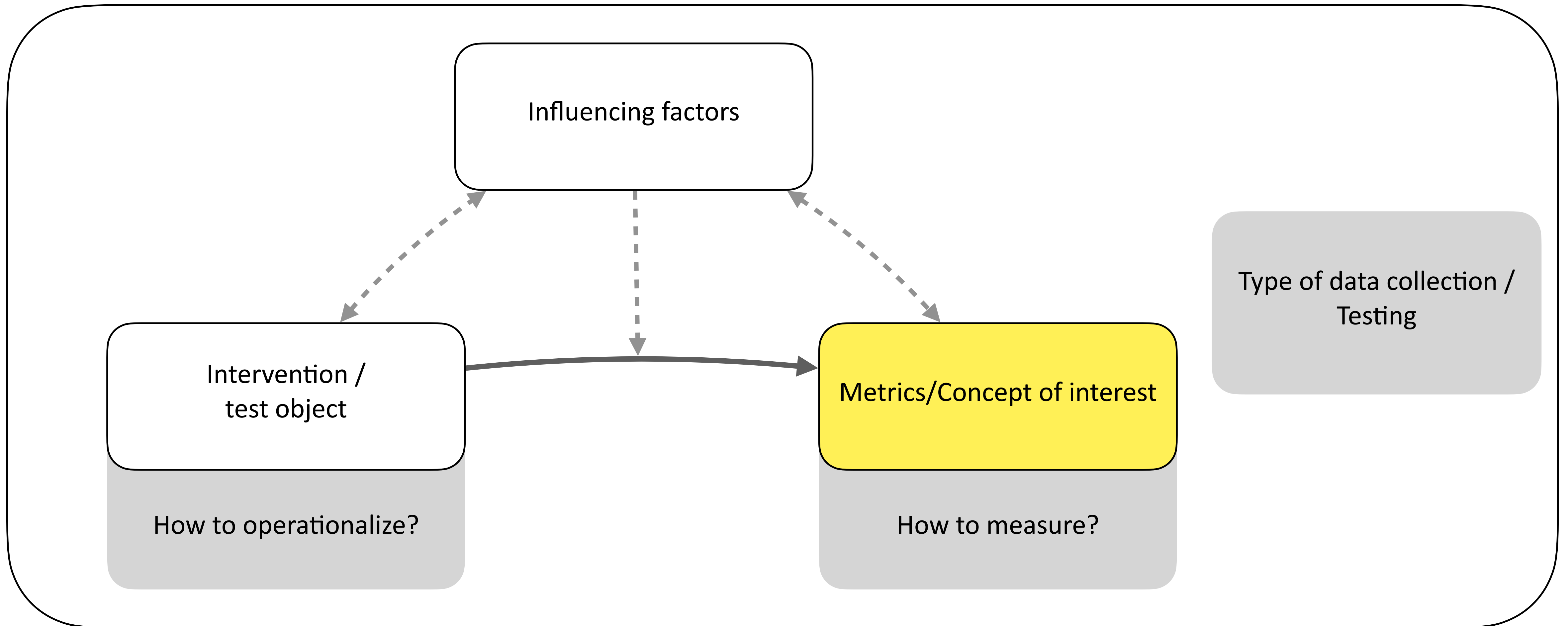
Specifying Your Research Design

Conducting Your Statistical Analysis

Gergle, D., & Tan, D. S. (2014). Experimental research in HCI. In Ways of Knowing in HCI (pp. 191-227). Springer, New York, NY.



# Defining Needed Variables



# Defining Needed Variables - Metrics

With funding from the:

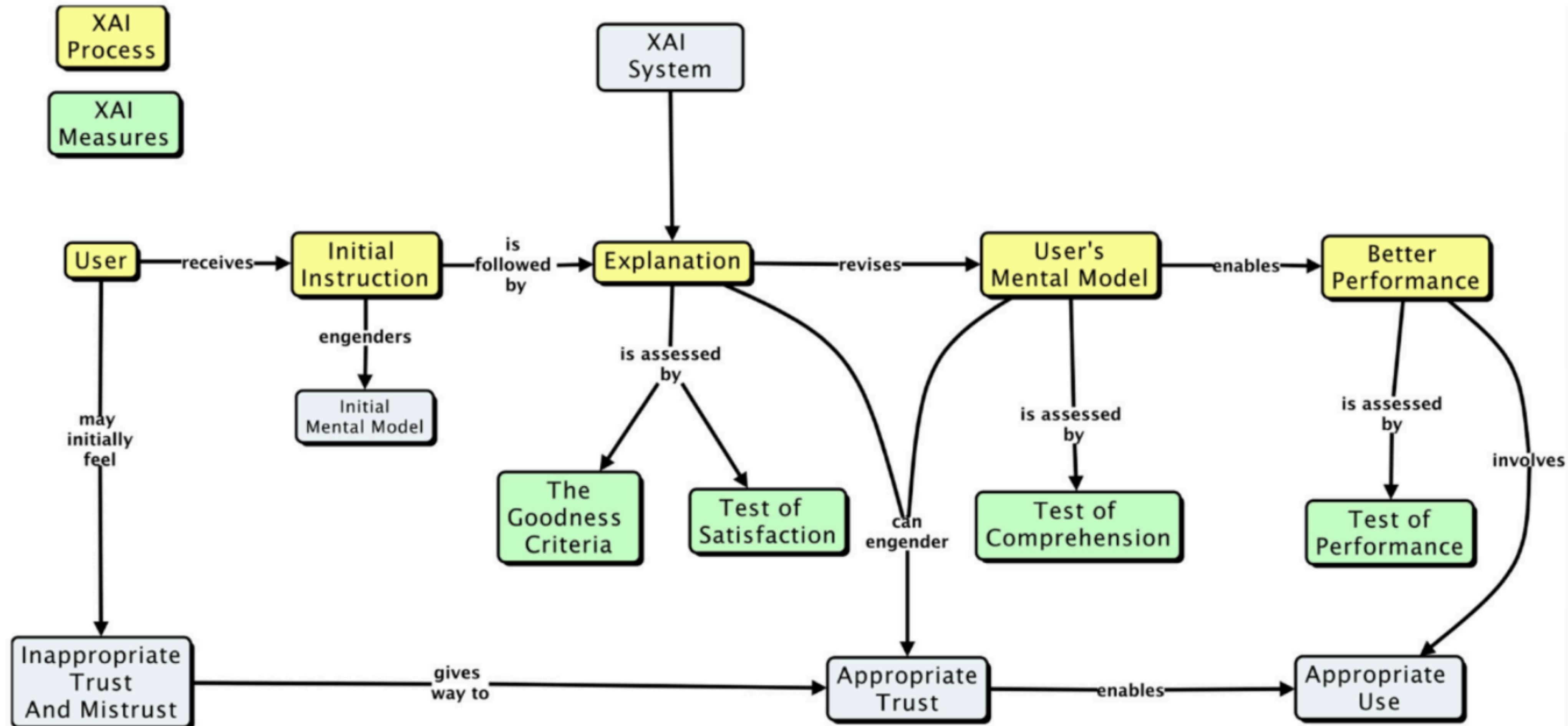


FIGURE 1

A conceptual model of the process of explaining, in the context of evaluating XAI systems.

# Defining Needed Variables - Metrics

## User Trust and Reliance, such as

- » Subjective Measures (self-explanation and interview, likert-scale questionnaire)
- » Objective Measures (user perceived system competence, user compliance with system, user perceived understandability)

## Human-AI Task Performance, such as

- » User Performance (task performance, task throughput, model failure prediction)
- » Model Performance (model accuracy, model tuning and selection)

## Computational Measures, such as

- » Explainer Fidelity (simulated experiments, sanity checks, comparative evaluation)
- » Model Trustworthiness (debugging model and training, human-grounded evaluation)

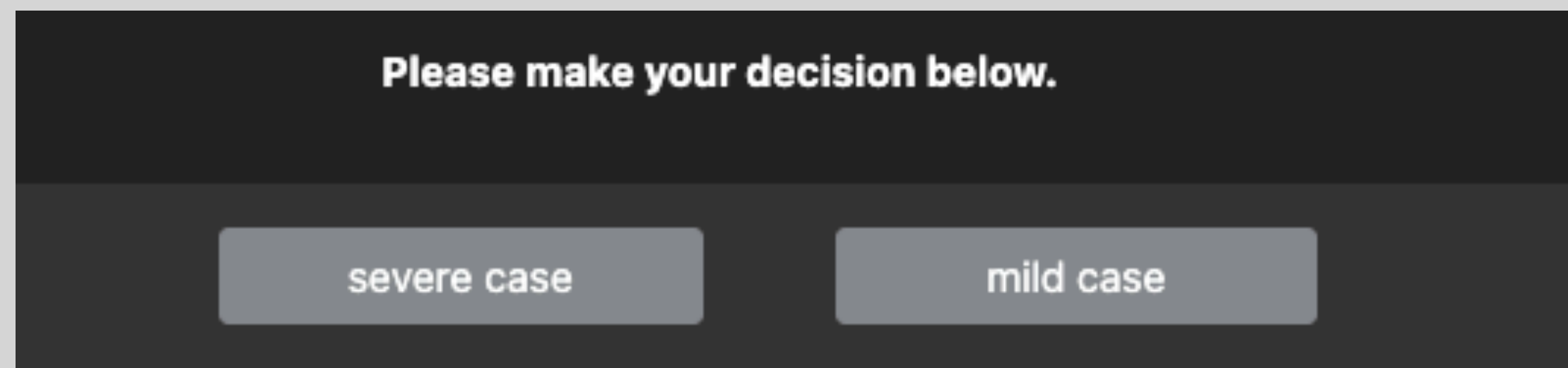


# Which metrics did we use?





# Which metrics did we use?

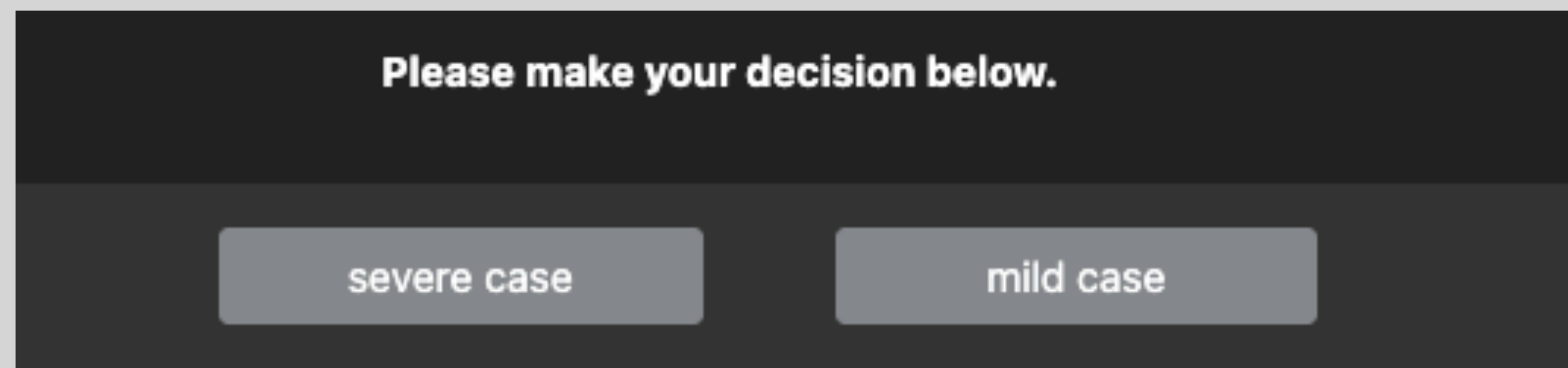


## Human-AI Task Performance, such as

- » User Performance (task performance, task throughput, model failure prediction)
- » Model Performance (model accuracy, model tuning and selection)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 1–45. <https://doi.org/10.1145/3387166>

# Which metrics did we use?



## Human-AI Task Performance, such as

- » Amount of correct decisions

## Human-AI Reliance, such as

- » Amount of correct decisions when the AI recommendation is wrong

## Human-AI Task Performance, such as

- » User Performance (task performance, task throughput, model failure prediction)
- » Model Performance (model accuracy, model tuning and selection)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 1–45. <https://doi.org/10.1145/3387166>

# Defining Needed Variables - Metrics

## XAI Evaluation Measures

### Explanation Usefulness and Satisfaction, such as

- » User Satisfaction (Interview and Self-report, likert-scale questionnaire, expert case study)
- » Explanation Usefulness (task duration, cognitive load, engagement with explanations)

### Mental Model, such as

- » User Understanding of Model (interview, self-explanations)
- » Model Output Prediction (user prediction of model output)
- » Model Failure Prediction (user prediction of model failure)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 1–45. <https://doi.org/10.1145/3387166>

# Which metrics did we use?

In which phase do you think you made the best recommendations?

- ☐ With the AI recommendation
- ☐ Deciding on your own
- ☐ With the AI recommendation and the confidence explanation

## Explanation Usefulness and Satisfaction, such as

- » User Satisfaction (Interview and Self-report, likert-scale questionnaire, expert case study)
- » Explanation Usefulness (task duration, cognitive load, engagement with explanations)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 1–45. <https://doi.org/10.1145/3387166>



# Defining Needed Variables - Metrics

## XAI Evaluation Measures

“Explanation Usefulness and Satisfaction, such as

- » User Satisfaction (Interview and Self-report, likert-scale questionnaire, expert case study)
- » Explanation Usefulness (task duration, cognitive load, engagement with explanations)

**Mental models are internal representations that people build based on real world experiences. These models allow people to understand, explain, and predict phenomena.**

- » User Understanding of Model (interview, self-explanations)
- » Model Output Prediction (user prediction of model output)
- » Model Failure Prediction (user prediction of model failure)

Johnson-Laird (1983)

# Which metrics did we use?

What would your decision be in this case?

☐ severe case

☐ mild case

What do you think the AI would recommend in this case?

☐ severe case

☐ mild case

## Mental Model, such as

- » User Understanding of Model (interview, self-explanations)
- » Model Output Prediction (user prediction of model output)
- » Model Failure Prediction (user prediction of model failure)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 1–45. <https://doi.org/10.1145/3387166>

# Which metrics did we use?

How confident were you in the recommendations you made to the patients?  very confident not confident at all

How well do you think you are able to decide whether this is a mild or severe case of nail fungus?  "Very capable" "Not capable at all"

How well do you think the AI is able to decide whether this is a mild or severe case of nail fungus?  "Very capable" "Not capable at all"

Explain why you think the AI would recommend this:

## Mental Model, such as

- » User Understanding of Model (interview, self-explanations)
- » Model Output Prediction (user prediction of model output)
- » Model Failure Prediction (user prediction of model failure)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–45. <https://doi.org/10.1145/3387166>

# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

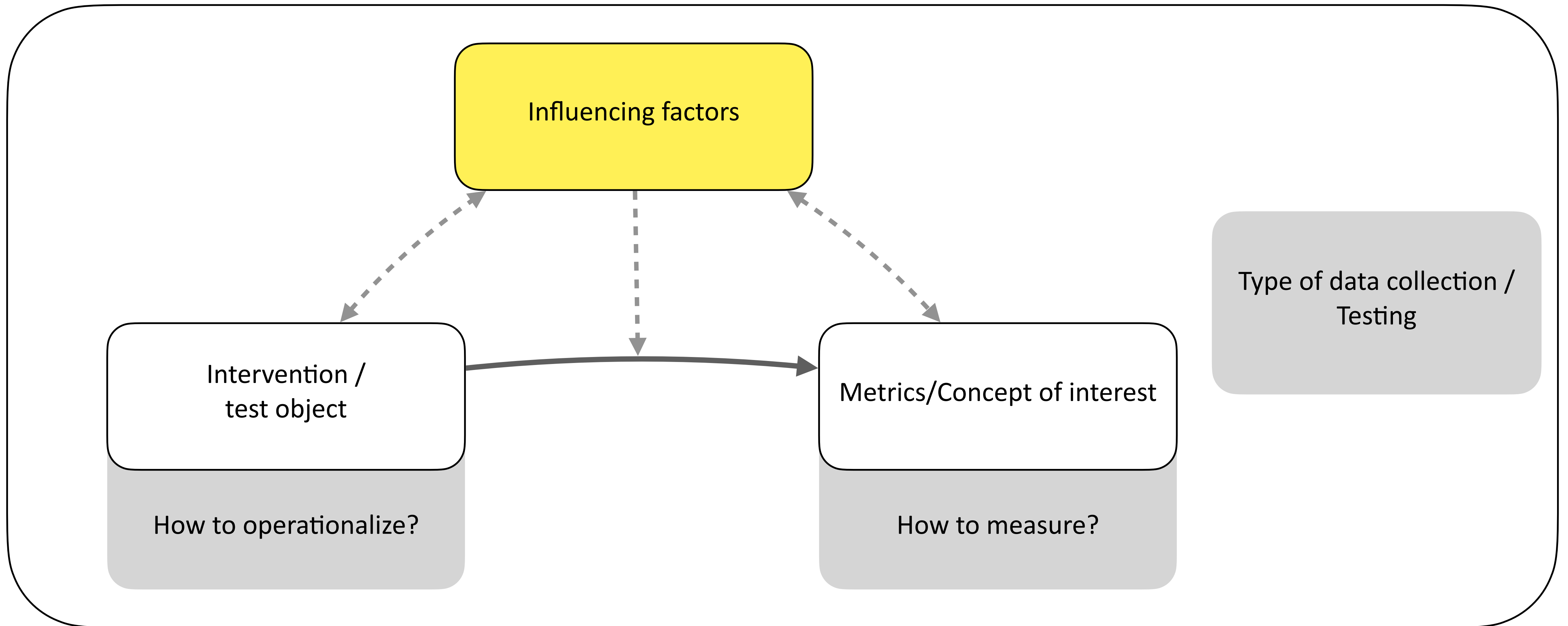
Defining Needed Variables

Specifying Your Research Design

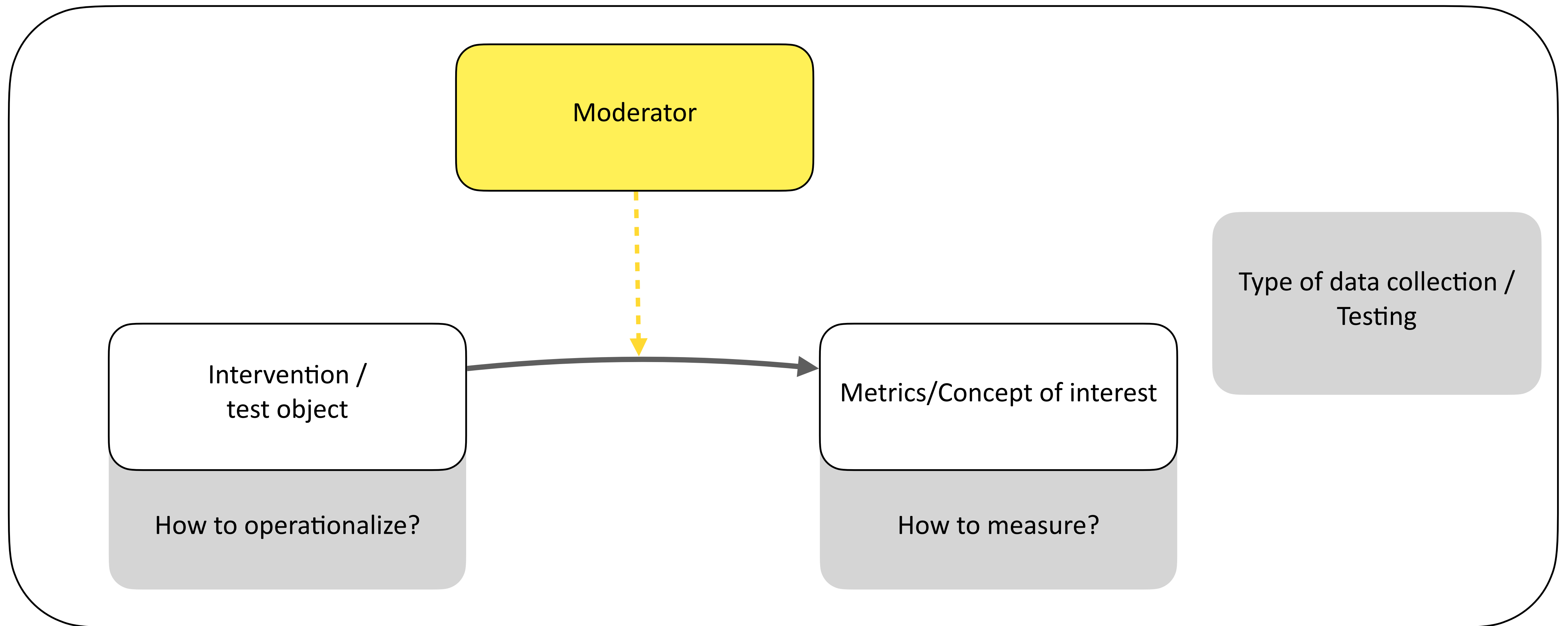
Conducting Your Statistical Analysis



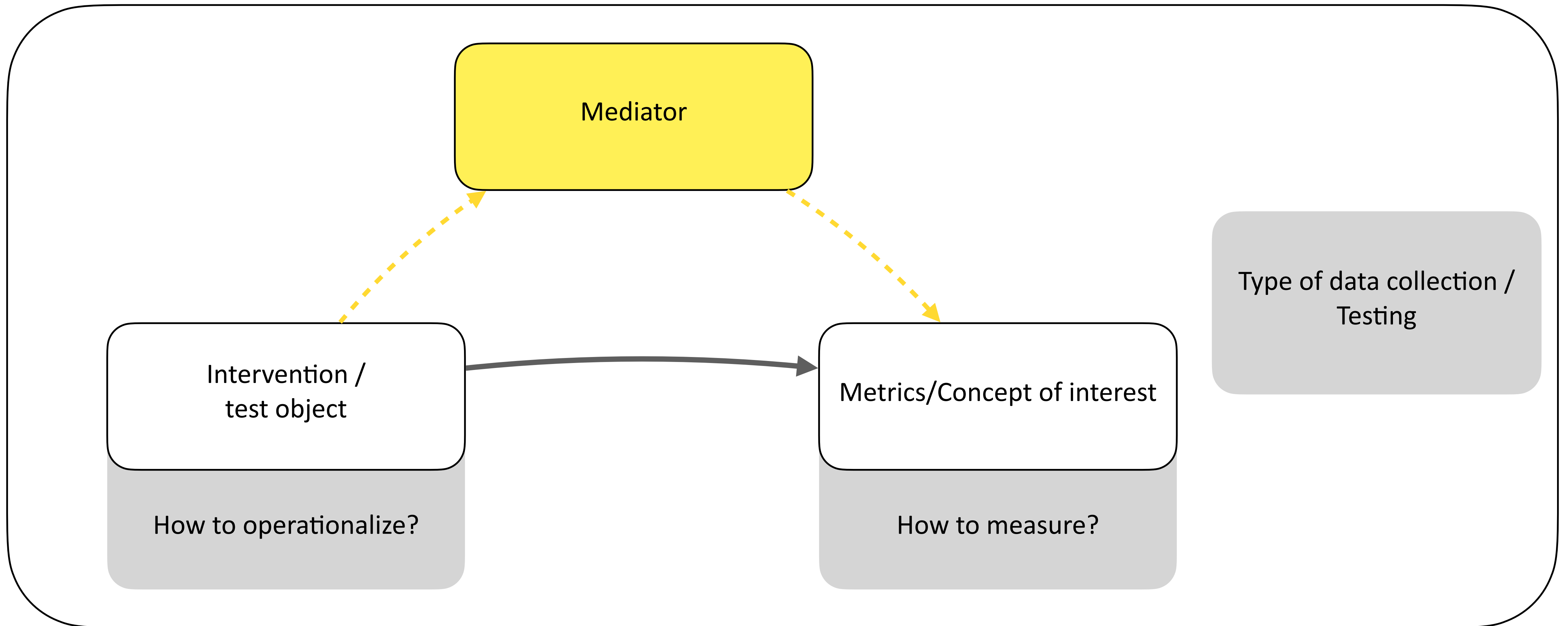
# Defining Needed Variables



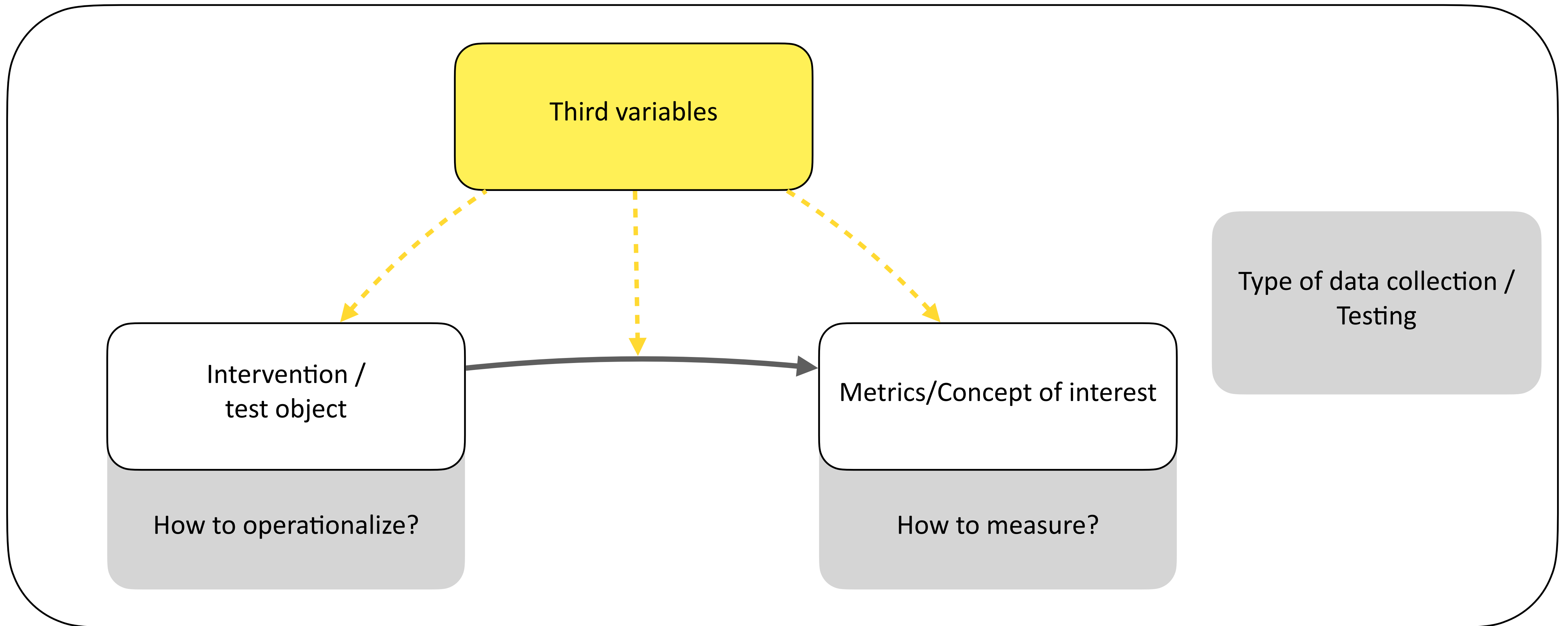
# Defining Needed Variables - Influencing factors



# Defining Needed Variables - Influencing factors



# Defining Needed Variables - Influencing factors





# Control Variable

A control variable is a potential independent variable that is held constant.

For example, when running reaction time studies you need to control lighting, temperature, and noise levels and ensure that they are constant across participants.

Holding these variables constant is the best way to minimize their effects on the dependent variable.

## Running Example:

While the screen size changes between experimental conditions, the frame rate is kept constant.

# Covariates

Covariates are additional variables that may influence the value of the dependent variable but that are not controlled by the researcher and therefore are allowed to naturally vary.

Often demographic variables.

The idea is that they need to be controlled because random assignment is not perfect, particularly in small samples.

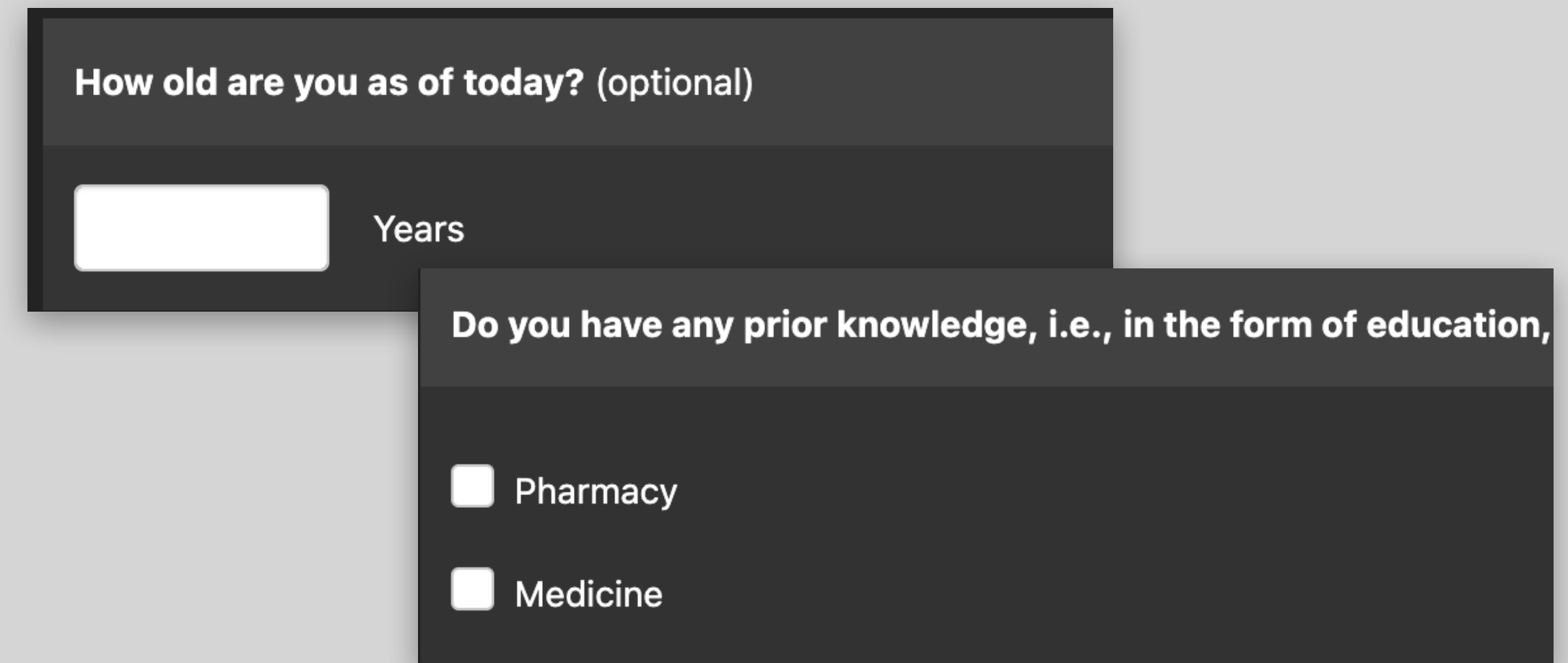
## Running Example:

Having impaired eyesight can have an influence on the task immersion. It should therefore be included as covariate for the analysis.

# Defining Needed Variables - Human factors

## Human Factors, such as

- » Age (Xiang et al., 2022 on trust)
- » Prior experience (Holstein et al., 2023 on explanation use)



How old are you as of today? (optional)

Years

Do you have any prior knowledge, i.e., in the form of education,

☐ Pharmacy

☐ Medicine

Xiang, H., Zhou, J., & Xie, B. (2022). AI tools for debunking online spam reviews? Trust of younger and older adults in AI detection criteria. *Behaviour & Information Technology*, 42, 478 - 497.

Holstein, K., De-Arteaga, M., Tumati, L., & Cheng, Y. (2023). Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1-20. <https://dl.acm.org/doi/pdf/10.1145/3579628>



# Defining Needed Variables - Human factors

## Human Factors, such as

- » Age (Xiang et al., 2022 on trust)
- » Prior experience (Holstein et al., 2023 on explanation use)
- » Need for Cognition (Buçinca et al., 2021; Chiesi et al., 2018)
- » Workload, e.g., TLX, SWAT, WT (Rubio et al., 2004)

TITLE	ENDPOINTS	DESCRIPTIONS
MENTAL DEMAND	Low/High	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	Low/High	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack
TEMPORAL DEMAND	Low/High	

I would prefer complex to simple problems.

I like to have the responsibility of handling a situation  
that requires a lot of thinking.

Thinking is not my idea of fun

Chiesi, F., Morsanyi, K., Donati, M. A., & Primi, C. (2018). Applying item response theory to develop a shortened version of the need for cognition scale. *Advances in cognitive psychology*, 14(3), 75. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7171511/pdf/acp-14-3-242.pdf>

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1-21.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied psychology*, 53(1), 61-86.

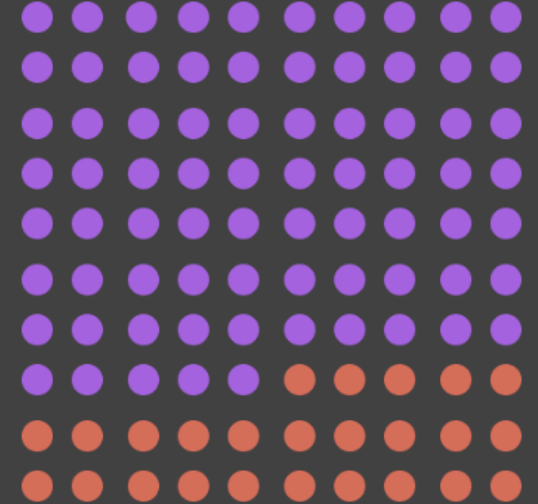
# Defining Needed Variables - Operationalization factors

**Did the user understand what he needs to understand, so that their data is valuable?**

» Did they understand the task? Or may they have guessed?

» ...

What did the visualization you were shown at the end indicate?



[Choose one of the following answers](#)

- ☐ Similarity of the image to the training data
- ☐ Confidence of the AI in each recommendation
- ☐ Amount of criteria met for a specific recommendation

Please name at least one limitation of the AI:

# Defining Needed Variables - Operationalization factors

**Did the user understand what he needs to understand, so that their data is valuable?**

- » Did they understand the task? Or may they have guessed?
- » Did they understand it language wise?
- » ...

How would you rate your English skills?

- ☐ None
- ☐ Beginner
- ☐ Intermediate

# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

Gergle, D., & Tan, D. S. (2014). Experimental research in HCI. In Ways of Knowing in HCI (pp. 191-227). Springer, New York, NY.



# General How-Tos - before the actual testing

## Checklist

- » Prepare consent forms for the participants.
  - » Check your local law (e.g., GDPR).
  - » Give the participants an overview of what is happening.
  - » Describe the origin of the project, contacts, compensation, risks, ...
- » If you want to publish the data, check if an ethics committee is available.

### Participation and Study description

- Your participation is voluntary.
- You may withdraw from partaking in the study at any time without giving reasons, without any study.
- The purpose of this study is to investigate different decision system designs and how they affect how humans and AI can work together to decide how nail fungus cases should be treated.
- This study will take
- First, you will be asked to choose not to. Next, you have to decide

### Information about your personal rights (GDPR)

- Right to gain access to the stored personal data (Article 15 GDPR).
- Right to rectification if data concerning your person is incorrect or incomplete.
- Right to deletion of data concerning your person, insofar as one of the legal requirements is met.
- Right to restriction of processing, in particular if the correctness of data is disputed, also instead of deletion of the data (Article 18 GDPR).

☐ I consent to participate.

☐ I do not consent to participate.

# General How-Tos - during testing

- » Explain your scenario, set the scene.
- » Introduce the user to the task of interest.
- » Give the Opportunity to practice.

# General How-Tos - What to consider during testing

For our study, we ask you to **imagine yourself in the following scenario.**

Imagine you want to earn extra money, and you have found a flexible remote job with the digital health startup called **eHEALTH**. This company aims to make digital health applications more efficient and affordable.

Your role is to work with eHEALTH's new treatment recommendation tool called **TreatMe**.

- » **Explain your scenario, set the scene.**
- » Introduce the user to the task of interest.
- » Give the Opportunity to practice.



# General How-Tos - What to consider during testing

- » Explain your scenario, set the scene.
- » **Introduce the user to the task of interest.**
- » Give the Opportunity to practice.

For our study, we ask you to **imagine yourself in the following scenario.**

Imagine you want to earn extra money, and you have found a flexible remote job with the digital health startup called **eHEALTH**. This company aims to make digital health applications more efficient and affordable.

Your role is to work with eHEALTH's new treatment recommendation tool called **TreatMe**.

You will now be introduced to the **clinical guidelines for determining whether a case of nail fungus is mild or severe**.

The interface displays one patient's case at a time on the left. Patients are instructed to upload a single, high-quality image that shows only one unique nail. The clinical guidelines are also provided to help guide your decision making.

The interface shows you the **clinical guidelines** and **the case of nail fungus**. The following example shows a severe case:

Your TreatMe Dashboard



# General How-Tos - What to consider during testing

- » Explain your scenario, set the scene.
- » Introduce the user to the task of interest.
- » Give the Opportunity to practice.

For our study, we ask you to **imagine yourself in the following scenario.**

Imagine you want to earn extra money, and you have found a flexible remote job with the digital health startup called **eHEALTH**. This company aims to make digital health applications more efficient and affordable.

Your role is to work with eHEALTH's new treatment recommendation tool called **TreatMe**.

You will now be introduced to the **clinical guidelines for determining whether a case of nail fungus is mild or severe**.

The interface displays one patient's case at a time on the left. Patients are instructed to upload a single, high-quality image that shows only one unique nail. The clinical guidelines are also provided to help guide your decision making.

The interface shows you the **clinical guidelines** and **the case of nail fungus**. The following example shows a severe case:

## Your TreatMe Dashboard

Mark the correct answers for the described cases below. For now, we will give you a description of

1. **Imagine a nail image showing a nail that is about 25% affected in the upper half and on t**  
**What would be your decision?**

Example:

### Clinical Guideline

It is a **severe** nail fungus case if ...

- the nail bed is affected  
or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...

- the nail bed is not affected  
and
- the nail is less than 50% affected

# General How-Tos - What to consider during testing

- » Explain your scenario, set the scene.
- » Introduce the user to the task of interest.
- » Give the Opportunity to practice.
- » **Use attention checks to assure quality.**  
**(for surveys)**

The color test you are about to take part in is very simple. When asked your favorite color you must select 'Grey.' This is an attention check.  
What is your favorite color?

- ☐ Blue
- ☐ Green
- ☐ Grey
- ☐ Red
- ☐ Brown

# How to carry out Experimental Research?

Research Question

Defining and Evaluating Hypothesis

Defining Needed Variables

Specifying Your Research Design

Conducting Your Statistical Analysis

Gergle, D., & Tan, D. S. (2014). Experimental research in HCI. In Ways of Knowing in HCI (pp. 191-227). Springer, New York, NY.





# Conducting Your Statistical Analysis

Just as important as the research design is planning the statistical analysis ahead of time in a way that ensures you can draw the appropriate conclusions from your experiments.

Simple means, medians, standard deviations, etc. are not usually sufficient – especially for comparative studies.

Resources:

<https://stats.oarc.ucla.edu/other/dae/>. | <https://stats.oarc.ucla.edu/other/examples/>

[https://www.methodenberatung.uzh.ch/de/datenanalyse\\_spss.html](https://www.methodenberatung.uzh.ch/de/datenanalyse_spss.html)

<http://www.stat.fu-berlin.de/en/index.html>

Gergle, D., & Tan, D. S. (2014). Experimental research in HCI. In Ways of Knowing in HCI (pp. 191-227). Springer, New York, NY.

# Determining Sample Size

When designing an experimental study it is important to plan for the number of participants needed.

Ideally you want an estimate that will allow you to reach a conclusion that is accurate with sufficient confidence.

A systematic approach to determining sample size depends on the particular experimental design, number of conditions, desired level of statistical confidence.

Existing web resources can help: <https://www.surveysystem.com/sscalc.htm> , [G\\*Power](#)

# Our example summarised

# Controlled Online Experiments

Controlled online experiments (known as A/B testing) are used at large Internet companies such as Google, Microsoft, or Facebook to generate design insights and stimulate innovation.

Experimental techniques are also widely used in usability testing

- » to help reveal flaws in existing designs or user interfaces
- » to evaluate if one user interface design is better than another
- » to show how a new recommender system algorithm influences social interaction
- » to assess the quality, utility, or excitement of an existing design



# Group comparison: A/B/C Testing



Clinical Guideline

It is a **severe** nail fungus case if ...


- the nail bed is affected
- or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...


- the nail bed is not affected
- and
- the nail is less than 50% affected

AI Support

The AI classifies:

 Mild

Patient ID: XXX





Clinical Guideline

It is a **severe** nail fungus case if ...


- the nail bed is affected
- or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...


- the nail bed is not affected
- and
- the nail is less than 50% affected

AI Support

The AI classifies:

 Mild

Patient ID: XXX



In 100 samples like this, AI would predict ...

...

75

to be severe (75%).

...

25

to be mild (25%).



Clinical Guideline

It is a **severe** nail fungus case if ...

- the nail bed is affected
- or
- the nail is at least 50% affected

It is a **mild** nail fungus case if ...

- the nail bed is not affected
- and
- the nail is less than 50% affected

AI Support

The AI classifies:

 Mild

Patient ID: XXX



In 100 samples like this, AI would predict ...

...

75

to be severe (75%).

...

25

to be mild (25%).

Capabilities

⊖ AI limits

Unable to consider a wider context.

Inflexible with untypical, low quality, or rotated images

⊕ Human strengths

Able to consider a wider context.

Can adapt based on new information.

⊖ Human limits

Less experience with nail fungus images.

Affected by surroundings.

⊕ AI strengths

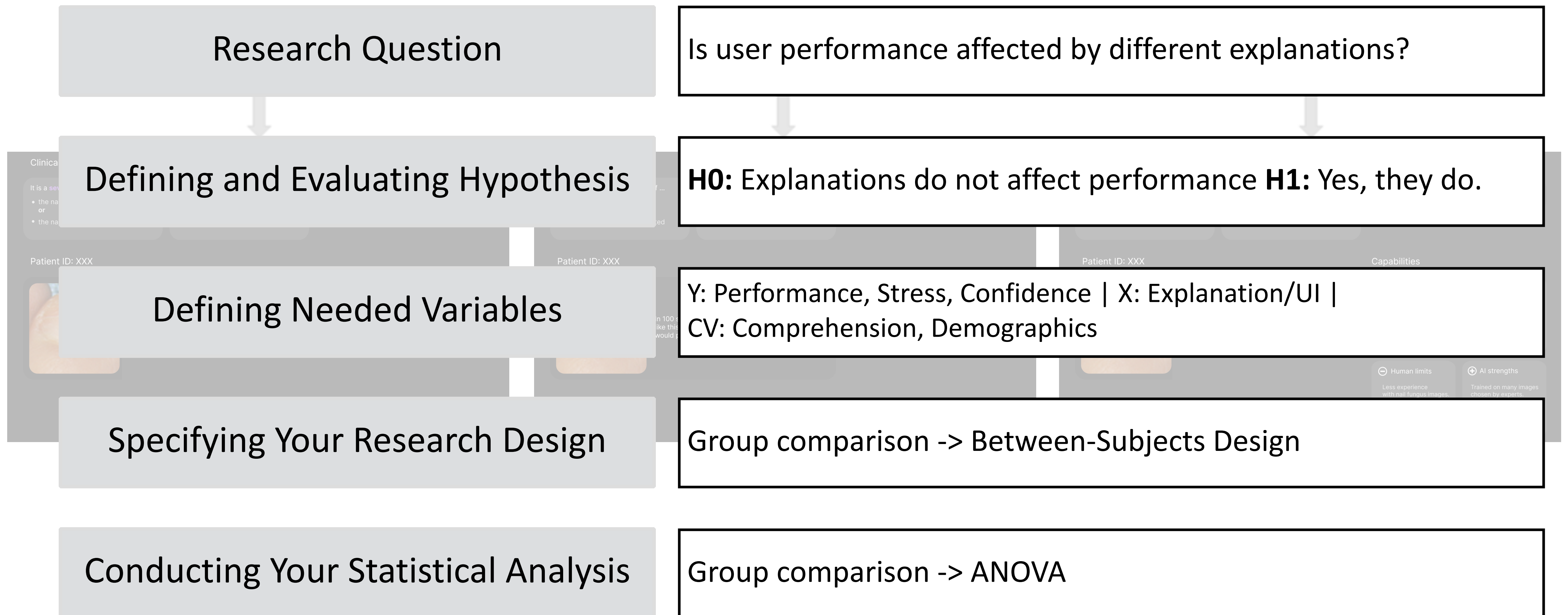
Trained on many images chosen by experts.

Unaffected by surroundings.

# Group comparison: A/B/C Testing



# Group comparison: A/B/C Testing (Warning - extremely simplified)



# Now to you — Present

## Scenario 1

You are working in a consulting firm. The company is developing a new LLM-based health assistant. They are using LLMs to help laypeople understand medical information. The LLM is based on a curated database including medical literature and patient data for many diseases. Often patients do not grasp medical information. The goal is to present patient with simple, everyday language so that they can understand what their options are. The developers at better health have different ideas on how to present the information. A first user test showed that people often struggle to understand the information. The goal is to be the best.

## Handout

### 1. Research Questions

Experiment research questions do not only ask, whether a relationship between two variables exists, but also aims at revealing the underlying cause by investigating causality.

**Examples:** "How does display size affect user satisfaction?", "How does text length affect user comprehension?"

### 3. Defining needed variables

## Your turn

Your task is to translate the scenario into an experiment plan in 5 steps (see example).

### Research Questions

### Hypotheses



# Scenario 1

You are part of a consulting firm tasked with assisting Health4All, an organization dedicated to improving patient understanding of their medical diagnoses. Health4All has developed a large language model (LLM) that utilizes a curated database containing information about symptoms, possible treatments, and disease implications. The goal is to translate complex medical terminology into simple, everyday language, making it easier for patients to comprehend their diagnoses and treatment options. Despite the promising design, initial user testing revealed significant challenges: many patients struggled to understand the information presented by the LLM. Health4All is eager to identify the underlying reasons for this confusion and determine which LLM variant performs best in communicating effectively with laypeople.

# Scenario 2

You are working for the government and have identified a significant challenge: many employees need to understand statistics related to demographics in Germany. For example, this understanding is crucial for planning hospital capacities, as they must assess the risks of health complications and predict future needs. Currently, employees often create bar plots, line plots, and box plots to visualize this data. However, you have observed that some users struggle with interpreting box plots accurately, leading to potential errors in their analysis. To address this issue, you are tasked with investigating how different visualization types impact users' understanding. Your goal is to determine which visualization method leads to the most accurate interpretation of data and, ultimately, to improve the quality of future decisions made by government employees.



# Scenario 3

You recently joined a company specializing in high-quality kitchen appliances, which has seen an unexpected surge in sales. You suspect this may stem from a flaw in the product recommendation algorithm. Currently, customers see items that are complementary in color to their purchases; for example, a customer with a violet couch is shown only violet and yellow kitchen appliance designs. The original algorithm aimed to recommend items based on what others bought, which did not boost sales. This raises questions about which algorithm is better for a small customer base and its scalability. To determine which recommendation algorithm drives sales more effectively, you plan to conduct an experiment to compare the current complementary color algorithm with the original one based on customer purchase patterns.

# Now to you — Exercise

**20 min:**

In groups of 2-3 choose an example and fill in the blanks.

**20 min:**

We will discuss your answer!

What have you decided on? What may be challenging?



# Thank you for your attention!

