Threats to credible research: Where are we at?

Sarah von Grebmer zu Wolfsthurn

09/11/2026

Prerequisites

> **!** Important
>
> Before completing this session, please carefully read about the prerequisites.

Prerequisite	Description	Link/Where to find it
UNESCO Recommendation on Open Science	For a first overview of Open Research, particularly pp 6-19	Download here

Presenter Notes: These are the **presenter notes**. You will find a script for the presenter for every slide. In presentation mode, your audience will not be able to see these presenter notes, they are only visible to the presenter.

Since this is the introductory session, there are no formal prerequisites for this session. If you will, today we are all starting on a blank page. The UNESCO Recommendation on Open Science is an optional prerequisite which sets the scene for what this course will be all about. The UNESCO Recommendation on Open Science is a landmark international framework adopted by UNESCO Member States in November 2021. It represents the first global standard-setting instrument for open research. The recommendation provides a shared definition, core values, guiding principles, and a set of action areas designed to help shape policies and practices that make science more accessible, transparent, inclusive, and equitable. Reading at least part of the recommendations before starting this course can provide a global context and shared language for understanding the concepts and goals that underlie open science practices. This helps ensure discussions are grounded in internationally recognized standards. It also clarifies the broad scope of open science, which goes beyond open access to publications and data to include cultural change, policy, education, societal engagement, and ethical considerations. The recommendations highlight the rationale and potential impact

of open science, particularly in terms of equity, inclusiveness, and the capacity to address complex global challenges through enhanced collaboration and information sharing.

Instructor Notes: There are also **instructor notes**. For some slides, there will be pedagogical tips, suggestions for activities and troubleshooting tips for issues your audience might run into. You can find these notes underneath the presenter notes.

Since this is the introductory session, there are no formal prerequisites. However, to get a small kick-start into the topic, learners are advised to have skimmed the UNESCO Recommendations on Open Science, particularly pages 6-19.

Threats to credible research - Intro Survey

Presenter Notes: Before we get started, I would like to explore your previous knowledge around Open Research. Perhaps some of you have read the recommendations by UNESCO, perhaps some have never heard of the term Open Research before. On the following slides you will find some questions which aim to capture how much you know already. You can scan the QR code, which will take you to a survey tool where you can click your way through the questions. We will then briefly discuss the questions in the plenum. Some of these questions I will ask again towards the end of this session, to get an understanding of whether anything has changed in terms of your knowledge or understanding during this session.

Instructor Notes: The pre-session survey serves to examine students' prior knowledge about the session's topic. Use free survey software such as Particify to establish the following questions (shown on separate slides). *Note*: This survey component does not work in an asynchronous setting.

Trust levels

- **Based on your experience so far, how would you currently rate your trust in published scientific findings on a scale from 1 - 5?**
  - 1 = trusting virtually **none** of the findings
  - 2 = trusting only **some** findings
  - 3 = trusting about **half** of the findings
  - 4 = trusting the **majority** of the findings
  - 5 = trusting virtually **all** findings

Challenges to research trust

- **Based on your experience so far, which challenges to research trustworthiness do you see? Name a few keywords.**
  - Display Wordcloud answer.

Open Research

- **What is your level of familiarity with Open Research practices in general (e.g., basic concepts, terminology, or tools)?**
  a. I am unfamiliar with the concept of Open Research practices.
  b. I have heard of them but I would not know how they apply to my work.
  c. I have basic understanding and experience with Open Research practices in my own work/research/studies.
  d. I am very familiar with Open Research practices and routinely apply them in my daily work/research/study routines.

Discussion of survey results

What do we see in the results?

Presenter Notes: Let us have a brief look at where we are at.

Instructor Notes: Briefly examine the answers given to each question interactively with the group. Use visuals from the survey to highlight specific answers.

Your current understanding of key terms (6 min)

Think (2 min)

- For yourself: Write down a definition for those terms: *Replicability*, *Replication*, *Reproduction*, *p-hacking* (without internet search - your current understanding)

Pair (2 min)

- Discuss your definitions with your neighbour

Share (2 min)

- Share your definitions with the larger group

This is not a test! These terms probably have not been covered in your courses yet. We just want to see what your current understanding of these terms is.

Covered in this session

- **Do** people **trust** in science?
- **Should** people **trust** in science? –> Replicability and reproducibilty
- Threats to replicability: **p-hacking, errors, and biases** in research

Presenter Notes: This slide provides an overview of the main topics we will cover in this session. Do not worry if some of these concepts make no sense to you at this point, we will get to all of them by the end of this session. We will begin by looking at research and the research cycle. We will look at how knowledge is produced, from forming a research question, through data collection and analysis, to publication and dissemination. Next, we will discuss trust in research and what it means. Trust is fundamental to science and research because it shapes how researchers trust each other's work, how institutions make decisions based on evidence, and how society at large engages with scientific findings. We will then move on to replicability and reproducibility, two key concepts for assessing the robustness of research. We will clarify the difference between them and examine why the inability to replicate or reproduce results has become a major concern across many disciplines. Following that, we will learn about biases, inconsistencies, and mistakes in research. These issues are often unintentional but can arise from methodological choices, data handling, publication pressures, or systemic incentives. Recognizing these challenges helps us better understand where and why research can go wrong.

Learning objectives

By the end of this session, you will be able to:

- **Reflect** on the role of trust in scientific research
- **Recognize** different types of challenges in research: research biases, statistical insecurities, errors and where they can arise from
- **Analyze** research scenarios to identify potential research biases and questionable research practices
- **Reflect** on the current threats for research and on the need for alternative approaches to conducting research

Presenter Notes: These are the learning objectives for today's session. By the end of the session, you should be able to define and distinguish key concepts related to research and open research, including the research cycle, trust in research, replicability, and reproducibility. You should also be able to recognize common challenges in research, such as biases, statistical issues, and errors, and understand where they can arise. In addition, you will analyze research scenarios to identify potential biases and questionable research practices. Finally, you will reflect on current challenges facing research and consider why alternative approaches to conducting research have been proposed.

Do people trust in research?

What is public "trust" in research?

- "*Society trusts that scientific research results are an honest and accurate reflection of a researcher's work.*" (Committee on Science, Engineering and Public Policy 2009: ix)
- "*The public must be able to trust the science and scientific process informing public policy decisions.*" (Obama 2009)

Resnik, D. B. (2011). Scientific research and the public trust. *Science and engineering ethics, 17*(3), 399-409.

Presenter Notes: What do we mean by trust in research? In their paper, Resnik argues that maintaining public trust is essential for the scientific enterprise, and he highlights how U.S. policy leaders have emphasized this responsibility. For example, he references the 2009 Committee on Science and Technology report, which highlights that the public expects science to be conducted responsibly and that preserving trust requires strong ethical norms, oversight, and systems that prevent misconduct. The part on "honest reflection of a researcher's work is particularly central here: society should be able to get a reliable and truthful insight into the methods and outcomes of the research process. They also cite President Obama's 2009 statement on scientific integrity, which stresses that government science must be guided by facts, transparency, and freedom from political manipulation. Resnik uses this to illustrate that public trust in science and research is not automatic. Going even further, they argue that trust needs to be earned and protected, given that science and research play a critical role in informing policies relevant to the public.

Instructor Notes: Beware that depending on the background of your learners, trust can have many facets. At its core, trust refers to the confidence that research processes, findings, and actors are reliable, ethical, and credible. However, how this trust is understood and emphasized can vary by discipline and fields. One important dimension of trust relates to research methods.

Across scientific disciplines, trust is built when appropriate methods are chosen, data are collected and analyzed rigorously, and decisions are clearly justified. In **experimental and quantitative fields**, this often involves standardized procedures, controls, and statistical robustness. In qualitative and interpretive disciplines, trust may instead be grounded in transparency, reflexivity, and the coherence of the research design. Another dimension is trust in researchers and research communities. This includes expectations of integrity, honesty, and adherence to ethical standards, as well as openness about limitations and potential conflicts of interest. In **research involving human participants or sensitive topics**, trust is closely linked to ethical responsibility and respect for participants. Trust also concerns research results and findings. This involves confidence that conclusions are supported by evidence and have not been distorted through selective reporting or questionable practices. Concepts such as replicability and reproducibility play a key role here, particularly in fields where results are expected to be independently verified. In **applied research areas such as medicine, engineering, or economics**, trust in findings is especially critical because research outcomes directly influence decisions, policies, and interventions. These dimensions are good to keep in mind when interacting with your learners.

Measuring trust in research

Probe your prejudices: In which profession do you have the highest trust to tell the truth? And the lowest?

. . .

Data from 1,015 UK adults, 2024

Sources: Veracity Index 2024, Ipsos; Seyd, B. (2025). What is trust (in science and scientists) and is it in crisis?. *Current Opinion in Psychology, 67*. https://doi.org/10.1016/j.copsyc.2025.102201

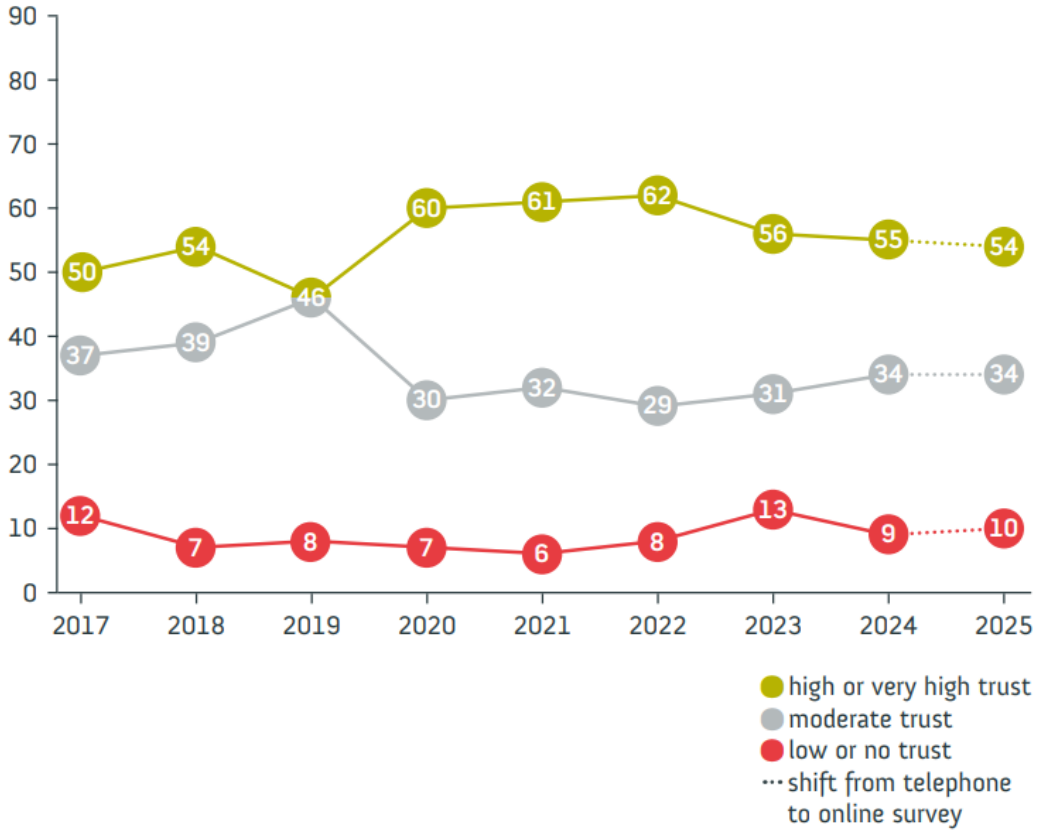
Presenter Notes: This chart shows results from the Ipsos Veracity Index 2024, which measures how much the UK public trusts different professions to tell the truth. The findings are based on a nationally representative sample of around 1,000 UK adults. Respondents were asked how much they trust people in each profession to tell the truth, and the chart reports the percentage who say they trust them (either "a great deal" or "a fair amount").

Starting at the top, nurses are the most trusted profession, with very high trust levels. They are followed closely by engineers and doctors, and then teachers and professors. Scientists also score relatively high, just behind judges and museum curators. These top professions share a common theme: they are generally seen as skilled, evidence-based, and working in the public interest, which likely contributes to their high trust. Moving to the middle of the chart, we see professions like police, judges, and business leaders, where trust is more moderate and mixed. At the lower end, journalists and advertising executives have relatively low trust levels. Finally, politicians and government ministers are the least trusted, with a significant gap between them and the top-ranked professions.

Public trust in research

German survey "Wissenschaftsbarometer"

"Wie sehr vertrauen Sie in Wissenschaft und Forschung?"



Source: *Wissenschaft im Dialog/Verian*: www.sciencebarometer.com. Minimum of 1,000 respondents in each survey wave.

Presenter Notes: Let us have a look at trust in research from the perspective of the general public. The Science Barometer is an annual survey of the German general public's attitudes toward science and research. It has been conducted every year since 2014 by Wissenschaft im Dialog (WiD). It focuses on a number of different elements, such as the expectations towards science, the relevance of science but also the trustworthiness of research and science. It is conducted annually and targets the German-speaking population aged 14 and over living in private households in Germany. The survey was traditionally conducted over the phone, but in 2025 shifted to an online panel to enhance the reach. In this, the data collection was done via a computer-assisted web interviewing (CAWI) during a defined period (July 4–18, 2025) with 2,011 respondents participating.

As mentioned, one aspect they enquired about was people's trust in science. The results suggested the following: Trust in science and research remains stable. Also in 2025, slightly more than half of respondents indicate having high or very high trust in science and research (54 per cent). One in ten trusts science and research little or not at all. In the telephone reference survey, 53 per cent said they trusted science and seven per cent had little or no trust. As in previous years, it is primarily younger people who place a high level of trust in science and research: 68 per cent of those under 30 and 60 per cent of 30–39-year-olds indicate trusting science. Among the older age groups, it is between 46 and 48 per cent of respondents.

Important, there are clear differences in trust in science and research with regard to the respondents' formal education level. Among respondents with high formal education levels, 72 per cent this year express high or very high trust. This means this proportion is below 75 per cent for the first time since the outbreak of the coronavirus-pandemic, but is still higher than before 2020. For respondents with medium and low formal education levels, however, the level of trust has returned to the value before the outbreak of the pandemic: 49 per cent of respondents with medium and 37 per cent of respondents with low formal education level indicate trusting science and research in 2025.

Instructor Notes: Highlight that while the results demonstrate that about half of the respondents had high or very high trust in research, the other half only had moderate, low or no trust in science. Highlight that one in two people from the respondents' sample did not trust science very much.

As an additional good to know from *Seyd (2025)* to show that the percentages in Germany are comparable to the rest of the world: A recent 68-nation study found a positive distribution of people's trust in scientists, with a global mean of 3.62 on a 1 (low trust) to 5 (high trust) scale (Cologna et al., 2025), with no country showing a trust score below the mid-point of 3. A 32-country survey conducted by Ipsos in 2024 found on average that 56 % of people trusted scientists, while just 15 % expressed distrust. However, the same study highlighted substantial differences in trust between national populations. Thus, while 70 % of Argentinians and Indonesians expressed trust in scientists, just 43 % of Japanese did likewise; see also the Ipsos 2024 survey.

Share the website of the Science Barometer with this report if asked by your learners: https://wissenschaft-im-dialog.de/documents/528/WIBA_2025_en_Web.pdf

---

Take aways from multiple international studies

- Trust in scientists remains **relatively high** in many countries
- Overall trust levels in research **remain stable**
- **But**: public **skepticism** about scientists' **integrity** and **transparency**

Seyd, B. (2025). What is trust (in science and scientists) and is it in crisis?. *Current Opinion in Psychology, 67*. https://doi.org/10.1016/j.copsyc.2025.102201

The used car dealer (10 min)

a.k.a. "Trust in a situation with asymmetric information"

- Form two groups: the *buyers* and the *dealers*.
- Given the bad reputation of car dealers:
  - *Buyers*: What strategies could you use to ensure you are making a trustworthy purchase?
  - *Dealers*: What would convince buyers that you are not one of the bad apples and your car has *really* good quality?
- Discuss in groups; present your strategies





Inspired by Vazire, S. (2017). Quality Uncertainty Erodes Trust in Science. *Collabra: Psychology, 3(1)*, 1. https://doi.org/10.1525/collabra.74. Images from ChatGPT.

Presenter Notes: Imagine you want to buy a used car from a car dealer. The dealer knows much more about the history and condition of the car than you do. As a buyer, you have to make a decision based on limited information, and you need ways to determine whether you can trust the seller and the product.

We will split into two groups: buyers and dealers. If you are in the buyer group, think about this question: Given the negative reputation that car dealers sometimes have, what strategies could you use to reduce uncertainty and make sure you are making a trustworthy purchase? If you are in the dealer group, think about the opposite perspective: How could you demonstrate that you are a trustworthy dealer and that your car is genuinely high quality? What evidence, guarantees, or practices could convince buyers that you are not one of the bad examples?

Discuss your ideas within your groups and then share your strategies with the rest of the class.

Afterwards, we will connect these ideas to research. Just as buyers need ways to evaluate the quality of a car when they cannot directly observe everything about it, researchers and the public need ways to assess the reliability and credibility of research findings.

Used cars vs. science

- Parallel in science: product = manuscript; seller = author; buyer = reader

  "When there are asymmetries in the information that the seller and the buyer have, the buyers cannot be certain about the quality of the products"

  "by keeping vital information private – the raw data, the original design and analysis plan, the exploratory analyses that were conducted along the way to the final analysis authors are hiding valuable information and preventing consumers of their manuscript from being certain about its quality. Just like sellers of used cars keep things [...] from the buyers."

  "For many non-scientists, learning that transparency is not the norm in science comes as a surprise. [...] it must seem obvious that **scientists should be held to a higher standard than used car salespeople**."

Quotes from Vazire, S. (2017).

Presenter Notes: The used car example can be directly connected to how trust works in science. The same problem of asymmetric information can occur when research findings are communicated.

In this analogy, the product is the research manuscript, the seller is the author or researcher, and the buyer is the reader, such as other researchers, policymakers, or the public.

The author has access to much more information about how the study was conducted than the reader does. They know the details of the research process: the original research question, the study design, the data collected, the analyses that were considered, and the decisions made along the way. The reader, however, often only sees the final published manuscript.

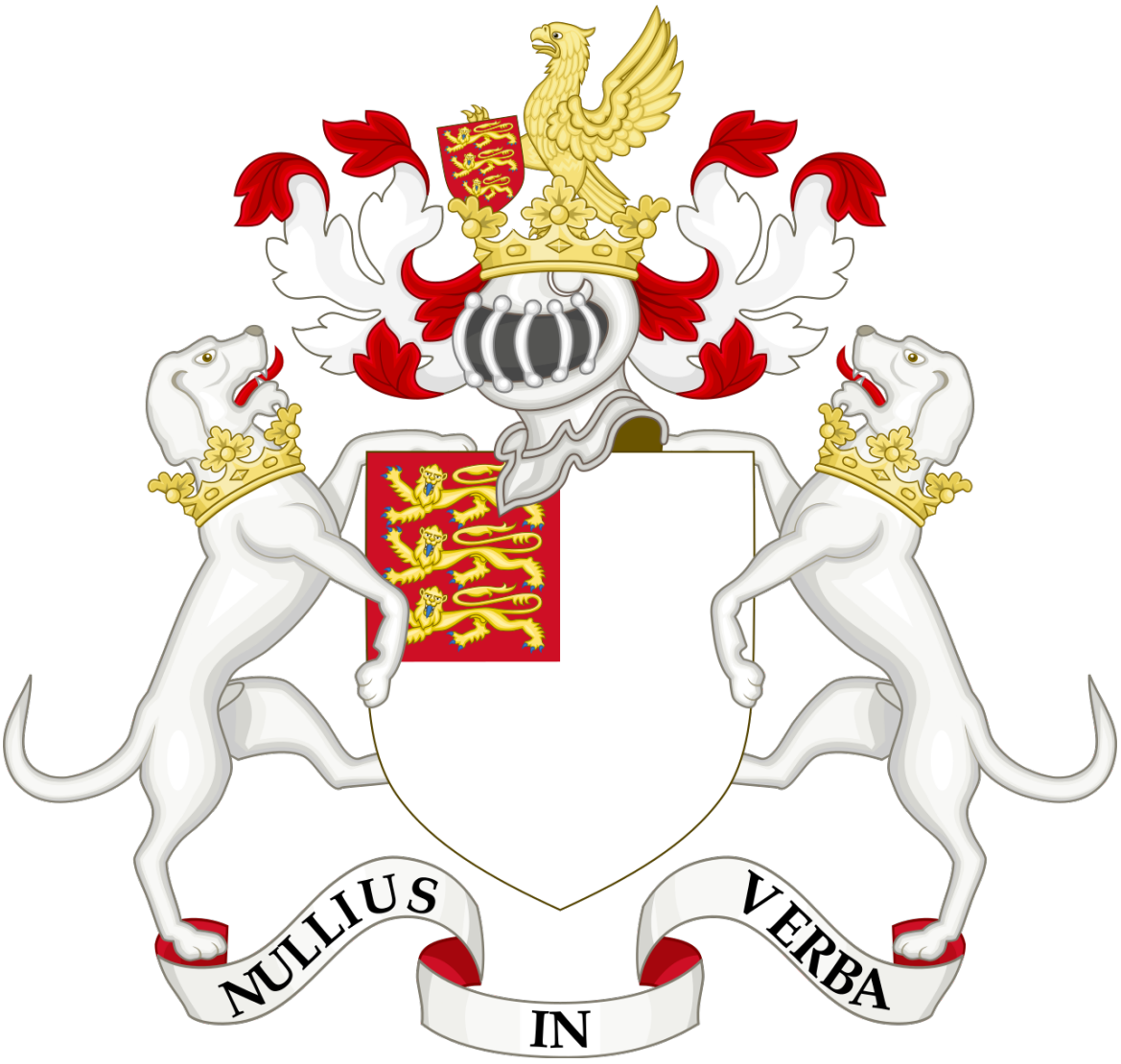
When important information is not shared, such as the raw data, the original analysis plan, or alternative analyses that were explored, it becomes difficult for readers to fully evaluate the quality and reliability of the research. In the same way that a used car buyer cannot be certain about the quality of a car if important information about its history is hidden, research consumers cannot be fully certain about the quality of findings if key parts of the research process remain invisible.

For many people outside of science, it may be surprising that transparency is not always the default in research. Given the role that scientific findings play in decision-making, one could argue that scientists should be held to an even higher standard of transparency and accountability than used car sellers.

A stark contrast?



Nullius in verba ("take nobody's word for it")



The motto of the Royal Society, the oldest scientific society in the world

Royal Society logo by MostEpic - Own work, CC BY-SA 4.0, Link; Car dealer by ChatGPT.

Should **we trust in science?**

Definition of key terms

Replicability and reproducibility

- **Replicability**: The extent to which design, implementation, analysis, and reporting of a study enable a third party to repeat the study and assess its findings.
  - Replicability is a property of the original study - it makes no statement about whether a replication has been attempted yet or whether that was successful or not.
  - Conceived as a continuum (hence the *extent* of replicability)
- **Replication**: A study that repeats all or part of another study *including new data collection* and allows researchers to compare their findings. Can be successful or not successful.
- **Reproduction**: The act of re-computing some numerical results using the *original data* (with or without the original code).
- **Reproducibility**: The extent to which the results of an original study agree with those of replication or reproduction studies.

Definitions of Replicability and Replication are from the iRISE glossary v2 by Voelkl et al. (2025). Note that no consensus about the meaning of these words exists - they are used differently in different disciplines.

Presenter Notes: One way to check if you are on the right track with your approach and to assess the relibility and credibility of your pasta experiment is to consider replicability and reproducibility.

Replicability refers to the ability to repeat an experiment using the same methods and obtain similar results. The focus here is on performing the experiment again, often with new participants or new samples, but following the same procedures, protocols, and conditions. Essentially, it tests whether the findings are robust when the experiment is repeated under similar conditions. For example, in psychology, this might mean running the same study with a new group of participants to see if the results hold.

Reproducibility, on the other hand, is about whether you can obtain consistent results using the original data and code. Here, no new data are collected; instead, you re-analyze the existing dataset with the same scripts or analysis methods to confirm that the reported results can be reproduced exactly. This is common in computational research, where sharing the dataset and analysis code allows others to verify the findings.

The **key distinction** is that replicability tests the experiment itself, while reproducibility tests the analysis of the original data. Both are essential for building trust in research because they help ensure that results are robust, transparent, and reliable. Understanding the difference between these two concepts helps researchers and readers critically evaluate the strength and credibility of scientific findings.

Instructor Notes: Make sure to spend enough time on this slide and provide an example scenario from the extra assignment at the end of the slide deck if you notice that the distinction between the terms remains too vague and/or abstract.

Prototypes of replication studies

- **Exact/direct replication**: (Virtually) Everything is inherited from the original study except the sample itself
- **Close replication**: Similar to direct replication, but with some (necessary) modifications to the methods or procedures; e.g. translations
- **Conceptual replication**: The same substantive research question is tested, but with different operationalizations
- **Replication and extension**: Doing a direct replication and adding new elements (also called "direct+" replication)

Note: The terminology around replication varies between fields, and the boundaries between these types are blurry.

Presenter Notes: There are different types of replication studies.

- **Exact/direct replication**: The goal is to reproduce the original study as closely as possible. The main thing that changes is the sample, for example, recruiting participants from a different population or location.
- **Close replication**: We still try to follow the original study closely, but make practical or necessary adjustments. An example is translating the materials into another language while keeping the underlying procedure the same.
- **Conceptual replication**: Here, we are testing the same underlying research question, but we approach it differently, for example, by using different measures, tasks, or manipulations.
- **Replication and extension**: This combines replication with something new. We first reproduce the original study and then add an additional element, such as a new condition, measure, or research question.
- **Terminology more generally**: These categories are useful as a guide, but different fields use different terms and sometimes the lines between the different concepts are blurry.

Better approach: Define exactly what changed/stayed identical

A prototypical **exact/direct replication**:

	At least one change	Identical	Unknown/ uncontrolled	Not applicable	Reason for change
Empirical Data	•				
Participant Population		•			
Independent Variable(s) and/or Intervention(s)		•			
Dependent Variable(s) and/or Outcome Measure(s)		•			
Procedure and Setting (incl. stimulus materials)		•			
Data Analysis Methods		•			
Computational Code / Scripts	•				Pragmatically motivated
Research Question, Hypotheses, and Theoretical Frameworks		•			

Presenter Notes: This table here visualizes an exact or direct replication. Since we would collect new data for a replication study, there would be at least one change in the empirical data. The rest would all be the same: design, stimuli, materials, analysis procedures, participant pool, research questions, hypotheses etc., except for scripts and analysis code.

Better approach: Define exactly what changed/stayed identical

A **realistic example**:

	At least one change	Identical	Unknown/ uncontrolled	Not applicable	Reason for change
Empirical Data	•				Theoretically motivated
Participant Population		•			
Independent Variable(s) and/or Intervention(s)	•				Pragmatically motivated
Dependent Variable(s) and/or Outcome Measure(s)		•			
Procedure and Setting (incl. stimulus materials)			•		
Data Analysis Methods		•			
Computational Code / Scripts	•				Pragmatically motivated
Research Question, Hypotheses, and Theoretical Frameworks		•			

The qualifiers “direct”, “conceptual”, “close” etc. replications are prototypical configurations of this matrix.

Presenter Notes: In reality, this is what a more realistic table could look like for a replication study: You may additionally also have to change the independent variables or the intervention you are giving to participants, and your procedure may be predefined by setting in which the study is taking place. Your intervention or independent variables may change because in the original study they depended on a specific software, or an intervention that is no longer available.

---

p-hacking and QRPs

***p*-hacking** (*n.*). Tune your data analysis in a way that you achieve a significant *p*-value in situations where it would have been non-significant.

Questionable research practices (**QRPs**) (*n.*). Practices of data collection and data analysis that are not outright fraud, but also not really kosher.

Presenter Notes: We will move to another factor which contributes to the low reproducibility and replicability rates across fields: accidental p-hacking. P-hacking is when researchers manipulate their data or analyses, intentionally or not, to get a statistically significant result (usually $p < 0.05$). In simple terms, it's like trying many different tests, methods, or subsets of data until something looks significant, and then reporting only that "lucky" finding. So, p-hacking is basically "data fishing" or "cherry-picking" results to make the research seem more impressive than it really is. It is important to highlight here at this point that p-hacking are considered questionable research practices, a.k.a. approaches which are not yet considered fraud by the scientific community, but are not considered "clean" either.

Replicability and reproducibility

> Additional exercises
>
> Want to practice how to distinguish the two? Skip to the end of the slides for additional exercises on replicability vs. reproducibility.

Presenter Notes: Want to practice how to distinguish the two? As an extra assignment, you can skip to the end of the slides for additional exercises on replicability vs. reproducibility.

Instructor Notes: Use the extra excercises as a way to challenge learners which may already have some prior knowledge about these two concepts.

Reproducibility Project: Psychology (RP:P, 2015)

The first large-scale replication project in the field

- **Exact and close replications** of 100 studies from 3 top journals from psychology; aiming for high power (larger sample sizes than original studies)
- Different ways to judge **replication success**: evaluated based on effect sizes, p-values, subjective assessment of replication teams
- Contacted original study authors when necessary
- Note: With today's understanding of the terms, the project should have been called "Replicability Project:Psychology"

Presenter Notes: The Reproducibility Project by the Open Science Collaboration (2015) involved a large-scale, collaborative effort by over 270 researchers. They selected 100 experimental studies published between 2008 and 2012 in three top psychology journals: Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition.

For each study, they carefully reviewed the original materials, methods, and analyses, and often contacted the original authors for clarification or access to materials and protocols. Each replication aimed to match the original study's sample size, experimental design, and analysis plan as closely as possible, while making only minimal changes when exact materials were unavailable. The replications were preregistered to specify hypotheses, design, and analysis before data collection, reducing bias and "researcher degrees of freedom." Data collection was conducted independently from the original research teams, often in multiple labs to ensure rigor. After data collection, they analyzed the results using the same statistical methods as the original studies to determine whether the findings could be reproduced. The project also evaluated effect sizes, not just statistical significance, to compare the magnitude of the original and replicated effects.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Reproducibility Project: Psychology (RP:P, 2015)

Results

- Only **36%** of the replications produced **statistically significant** results compared to 97% of original studies
- **39% of effects** were considered to have **replicated** the original results

- The mean **effect size** in replications was approx. **half** compared to the original results

"More work to do to verify whether we know what we think we know".

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

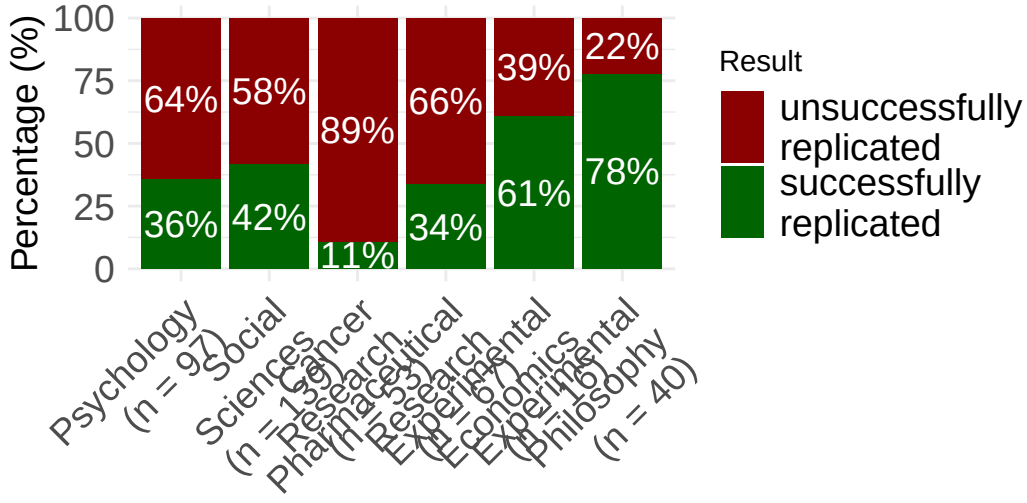
Presenter Notes: Only 36% of the replications produced statistically significant results compared to the original 97%. Only 39% of effects were considered to have replicated the original results. The mean effect size in replications was about half that of the original studies, highlighting the reproducibility challenges in psychology. However, it is still important to remember that replication is a continuum, hence the "extent of" replication. A failed replication does not automatically mean that the original finding was false.

In sum, many replication studies produced weaker evidence for the original findings despite using materials provided by the original authors. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know.

Replication rates across disciplines

Up to 2018

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.2.1     v readr     2.2.0
v forcats   1.0.1     v stringr   1.6.0
v lubridate 1.9.5     v tibble    3.3.1
v purrr     1.2.2     v tidyr     1.3.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```



Begley, C. G., & Ellis, L. M. (2012); Camerer et al (2016); Chang & Li (2015); Cova et al. (2018); Open Science Collaboration (2015); Social Science: Combined sample of systematically sampled projects (RPP, SSRP, EERP); Prinz, F., Schlange, T., & Asadullah, K. (2011); Protzko et al. (2023)

Presenter Notes: Different studies looked at the replicability of findings. They provide us with interesting and probably alarming results about whether independent researchers can repeat the study using new data and get the same overall findings or conclusions.

As you can see here, the dark red bars represents the percentage of studies which were not successfully replicated. Psychology is quite up there, about 64% of all studies we conduct do not replicate. What is equally as worrying, if not more, are that 89% of studies from cancer research do not replicate either. The "best" field really is experimental philosophy, so from this, they have the most robust and credible scientific approach.

Instructor Notes: Keep in mind that a lot of aspects of these studies can be discussed in a critical light. Learners might point this out. For example: Were the studies selected randomly (or did they only select the studies that looked fishy from the start?). How do you even measure the "replicability"? Is simply looking for significance in the replications study sufficient or even useful? All of these are good points, and there is a whole emerging field that tackles these questions. It is called meta-science, and it takes a scientific view onto the scientific enterprise itself.

Replication rates in the social and behavioural sciences

- Tyner et al. (2026) attempted **replications of 274 claims of positive results** from 164 quantitative papers (published between 2009 and 2018)
- **~55% of positive claims replicated successfully**

Discipline	Replication attempts (successful / total)	Percentage successful
Business	17 / 36	47.2%
Economics	10.2 / 24	42.5%
Education	8.2 / 13	63.1%
Political science	7.8 / 15	52.0%
Psychology	28.4 / 58	49.0%
Sociology	9.2 / 18	51.1%

Tyner, S. K., Abatayo, A. L., Dayley, M., et al. (2026). Investigating the replicability of the social and behavioural sciences. Nature. https://doi.org/10.1038/s41586-025-10078-y

Presenter Notes: - Authors attempted replications of 274 claims of positive results from 164 quantitative papers (published between 2009 and 2018) in social and behavioural sciences, including psychology, sociology, education, business, economics, and political science. - Original materials used when available - ~55% of positive claims replicated successfully

Notes:
- Papers are weighted combinations of claims, accounting for multiple claims per paper replicated in some cases.
- The left column shows the number of successful replications out of total papers with replication attempts.

Not a recent issue

Adapted from Dr. Malika Ihle: https: https://osf.io/u3znx

Presenter Notes: Replication issues are not a new problem in research; they have existed across disciplines for decades. Historically, scientists have encountered difficulties when trying to repeat experiments, even when methods were clearly described. This has been seen in fields ranging from psychology and biology to medicine and economics. What is more recent is the systematic awareness and large-scale study of replication problems, often referred to as the "replication crisis." Researchers today are documenting and quantifying these issues, but the underlying challenges, e.g., methodological complexity, variability, and human error, have been part of the research landscape for a long time.

Why did not more studies replicate?

human error

lack of resources

cognitive bias

publication pressure

??

lack of statistical knowledge

Presenter Notes: The question now becomes: **Why did not more studies replicate?** Several factors contribute to these long-standing replication challenges. We will now have a closer look at some of these factors.

Instructor Notes: Clearly state the switch into the origin for the replication challenges and get learners curious as to why not more studies replicated.

Low power and low p(H1)

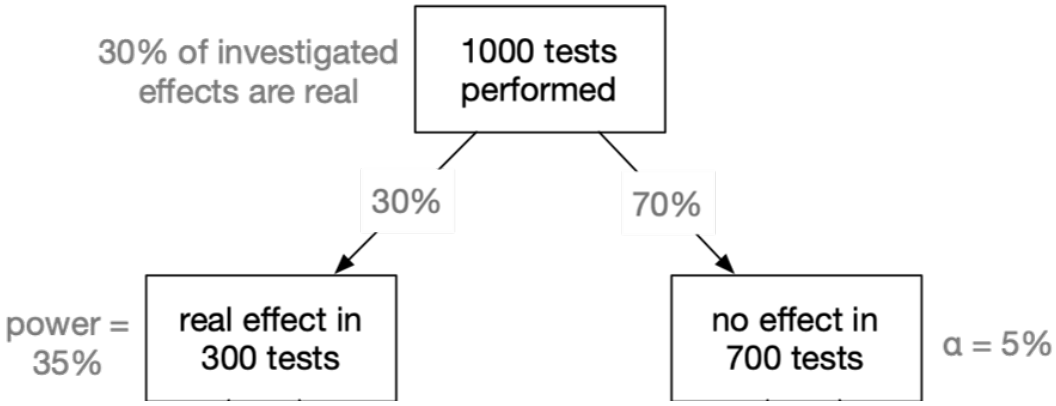
Live Poll

Given that in a given study $p < .05$: What is the probability that a real effect exists in the population $p(H_1|D)$?

Presenter Notes: Time for a quick poll! Given that in a given study $p < .05$: What is the probability that a real effect exists in the population $p(H_1|D)$?

Instructor Notes: Prepare the poll in advance using software such as Particify.

False discovery rates (I)



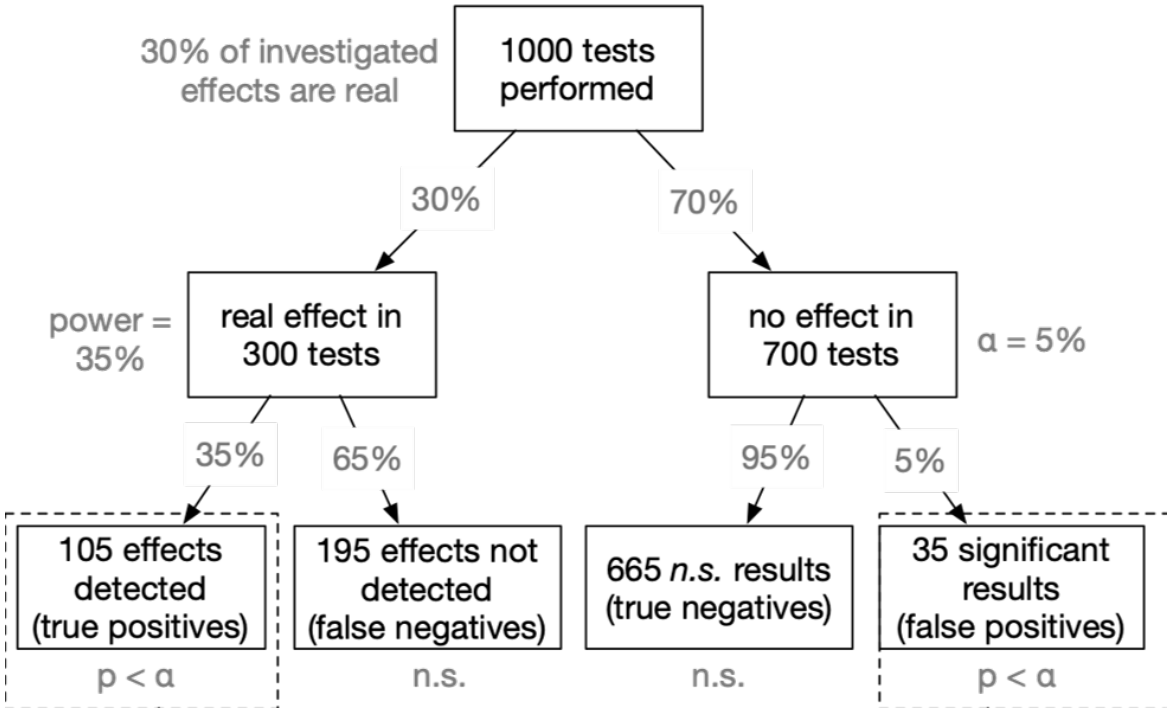
Nuzzo, R. (2014). Statistical errors. Nature. Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1(3), 140216–140216. http://doi.org/10.1073/pnas.1313476110

Presenter Notes: This figure illustrates an important problem with interpreting a p-value of 0.05 as proof that a finding is real.

Let's imagine that we perform 1,000 statistical tests. In this example, we assume that 30% of the tests (so 300 of these tests) are investigating effects that are actually real, while the remaining 70% have no real effect.

The key issue is statistical power. Here, the power is only 35%, meaning that when a real effect exists, the test detects it only about 35% of the time. This is an example of an underpowered study, which Nuzzo highlights as an important source of statistical error.

False discovery rates (II)



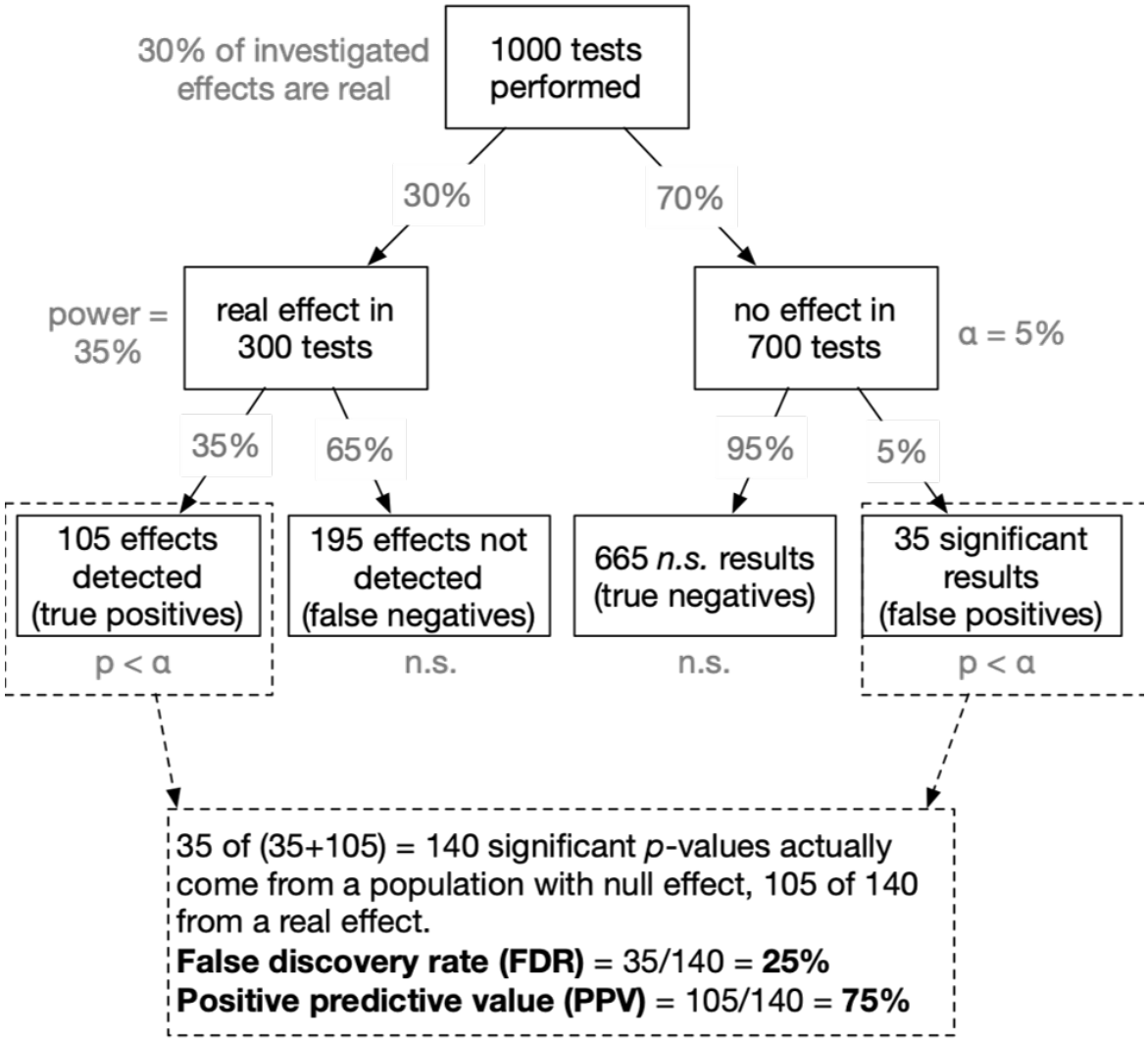
Nuzzo, R. (2014). Statistical errors. Nature. Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1(3), 140216–140216. http://doi.org/10.1073/pnas.1313476110

Presenter Notes: Recapping what we discussed before: Out of the 1,000 tests, 300 are testing real effects, and 700 are testing effects where there is actually no effect.

On the left-hand side, we have the 300 tests where a real effect exists. The statistical power of the test is 35%. Power is the probability of detecting an effect when that effect is genuinely present. So, with 35% power, we detect the real effect in 35% of those 300 tests. That gives us 105 true positives. These are findings where the test is statistically significant, p is less than alpha of 0.05, and there really is an underlying effect.

However, the remaining 65% of the 300 tests, or 195 tests, fail to detect the real effect. These are false negatives. The effect is genuinely there, but the statistical test does not identify it as significant. This illustrates one of the problems associated with low statistical power, which Nuzzo (2014) discusses: underpowered studies have a greater risk of missing genuine effects.

False discovery rates (III)



Nuzzo, R. (2014). Statistical errors. Nature. Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1(3), 140216–140216. http://doi.org/10.1073/pnas.1313476110

Presenter Notes: With an alpha level of 5%, we expect 5% of these tests to be statistically significant just by chance. That gives us 35 significant results that are actually false positives. The remaining 665 are true negatives.

Now we can focus specifically on the results that are statistically significant. There are 105 true positives plus 35 false positives, giving us a total of 140 significant p-values.

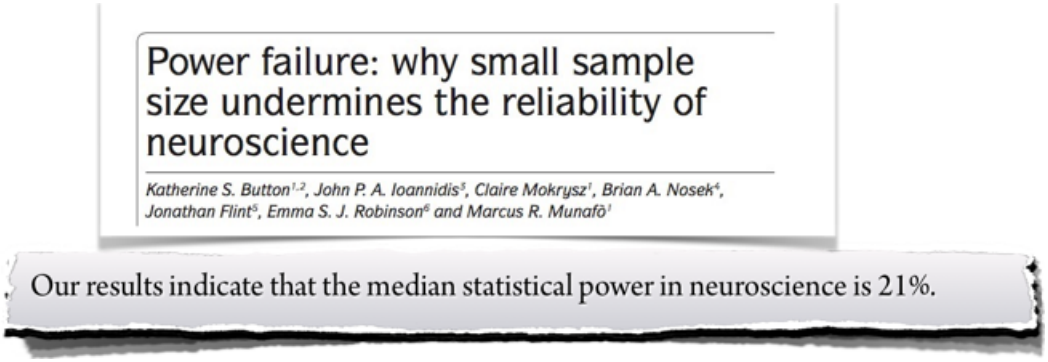
The important question is: How many of these 140 significant results actually represent a real effect? We know that 105 of them do, while 35 do not. Therefore, the false discovery rate, or

FDR, is: $35 \div 140 = 25\%$. So, despite using the conventional 5% significance threshold, 25% of the statistically significant findings in this example are false discoveries.

Another way to express the same idea is through the positive predictive value, or PPV. This tells us the probability that a statistically significant result actually corresponds to a real effect. Here, the PPV is: $105 \div 140 = 75\%$. So, when we obtain a statistically significant result in this particular scenario, there is a 75% probability that it represents a real effect, meaning there is still a 25% probability that it is a false positive.

So, the key takeaway from this example is that '$p < 0.05$' should not be interpreted as 'there is only a 5% chance that this result is false.' In this scenario, the actual false discovery rate among significant results is 25%.

Applied to neuroscience



- Assumed that our tested hypothesis are true in 30% of all cases (which is a not too risky research scenario):
  - A typical neuroscience study must "fail" ($p > \alpha$) in 90% of all cases
  - In the most likely outcome of $p > .05$, we have no idea whether a) the effect does not exist, or b) we simply missed the effect. Virtually no knowledge has been gained.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci, 14(5), 365–376. doi:10.1038/nrn3475

Presenter Notes: In the example of neuroscience, where studies are often underpowered, typical study may fail to reach statistical significance in around 90% of cases. When we get a non-significant result, or p > , we cannot tell whether there is genuinely no effect or whether the study simply lacked enough power to detect the effect.

The evidential strength of underpowered studies

> "When a study is underpowered it most likely provides only weak inference. Even before a single participant is assessed, it is highly unlikely that an underpowered study provides an informative result."
>
> "Consequently, research unlikely to produce diagnostic outcomes is inefficient and can even be considered unethical. Why sacrifice people's time, animals' lives, and societies' resources on an experiment that is highly unlikely to be informative?"

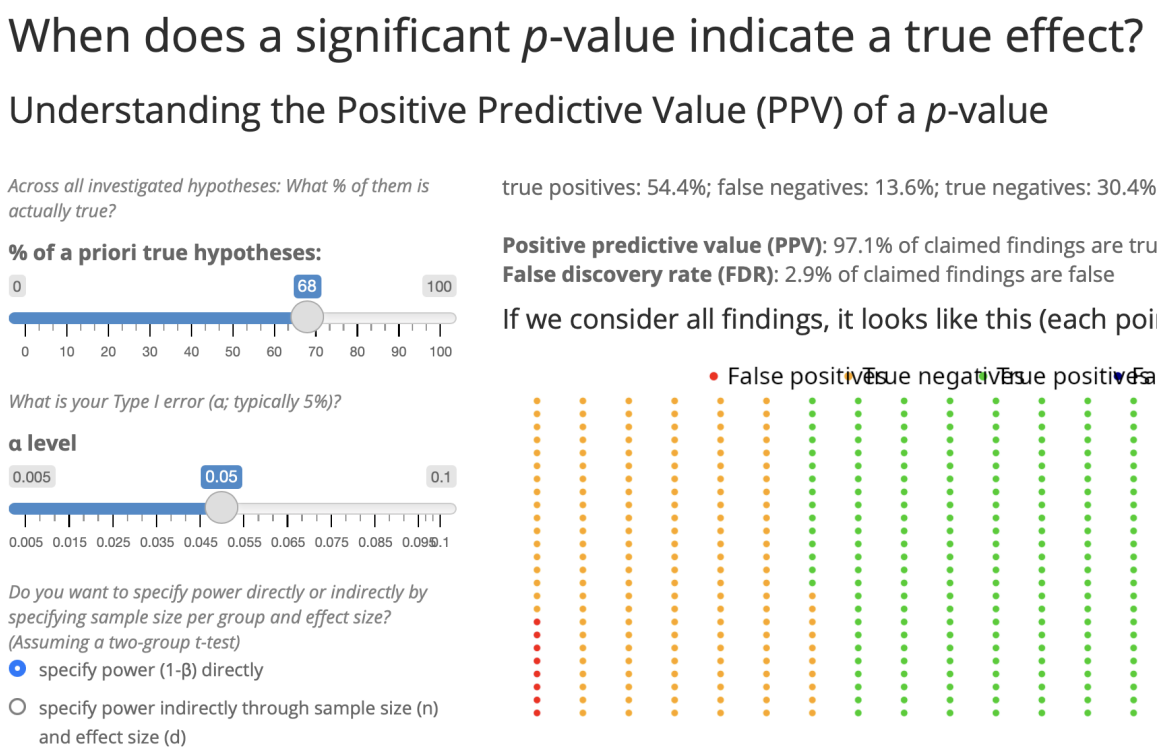
Schönbrodt, F. D. & Wagenmakers, E.-J. (2018). Bayes Factor Design Analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*, 128-142. doi:10.3758/s13423-017-1230-y. Open Access

Presenter Notes: Add.

Conclusion

Given that in a given study $p < .05$: What is the probability that a real effect exists in the population $p(H_1|D)$?

- $\rightarrow$ In the typical situation in psychology ($p(H_1) = 10\%$; power = 35%): The probability that a significant p-value indicates a true effect is only 44%.
- Try it out yourself! https://shiny.psy.lmu.de/felix/PPV



Career incentives in science

Presenter Notes: This goes back to the question: why is replicability and reproducibility such a big theme across all disciplines?

How to become a professor?

- What is the single most important **thing you need** to become a **professor**?
- Survey among N = 1453 psychology researchers, 66% were actually members of a professorship hiring committee

Actual (not desired) relevance in professorship hiring committees	**Rank**
Number of peer-reviewed publications	1
Fit of research profile to the hiring department	2
Quality of research talks	3

Actual (not desired) relevance in professorship hiring committees	Rank
Number of publications	4
Volume of acquired third party funding	5
Number of first authorships	6

Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? *Psychologische Rundschau, 67*(4), 250–261. http://doi.org/10.1026/0033-3042/a000335

Presenter Notes: There has been research which tried to look at this element of importance of published papers. This study by Abele-Brehm et al. (2016) looked at criteria in the hiring process and what the desired criteria are vs. which criteria candidates are selected on in real life. The growth of a scientific field depends on the people working in it. Choosing the right people for professorships is therefore very important. This study looks for the first time at how psychologists view the process of hiring professors. It explores how important they find different signs of suitability, how big the gap is between what they want and what actually matters in these procedures, and what they think about different ways of designing them.

A total of 3,784 members of the German Psychological Society (DGPs) were invited to take part in an online survey, and 1,453 answered at least some of the questions.

The take-home from this study was that among the top six criteria based on which candidates are selected. Three are about publications and first authorships. We have number of peer reviewed publications in the first spot, more generally number of publications in spot 4, and the number of first authorships, also indirectly connected to how much you published. So, based on this, we could say great, in order to advance in academia, all I need to do is to publish as many papers as possible, preferably where I am the first author.

Instructor Notes: Highlight how quantity is rewarded of quality of the research.

How to become a professor?

! Publication Pressure

"Researchers are not rewarded for being right, but rather for **publishing a lot**."

⚠ Consequence: "Publish or Perish" culture

Researchers try to **publish as much as they can** and to outperform their peers (Schmidt et al., 2021).

Nelson, Simmons, & Simonsohn (2012); Nosek, Spies, Motyl (2012); Munafo (2016)

Presenter Notes: First, we have the issue of publication pressure. Publications are currently *the* go-to measure of academic success. Papers help researchers being considered for permanent positions, for grants, and for becoming and remaining visible in the respective research fields. However, a consequence of published papers being so important is that researchers try to publish as much as they can.

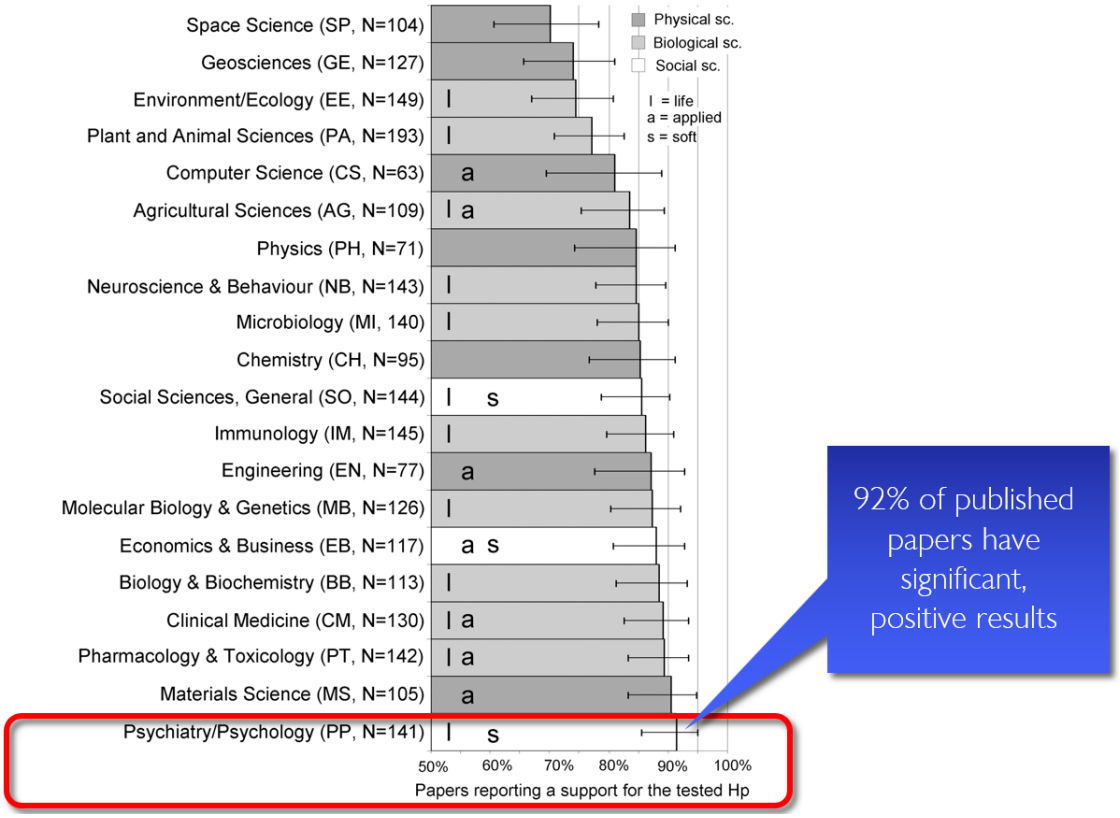
Why published papers are useful:

- Getting a **job** in academia
- Being awarded **grant money** for your research
- Being **visible** in the respective research field

Instructor Notes: Check if learners are aware of the publication process in broad terms and check if they understand what *grants* means in this context.

OK, I need a lot of publications ...

How do I get lots of publications?



Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS ONE, 5*, e10068. https://doi.org/10.1371/journal.pone.0010068

Presenter Notes: Following this logic, the next question immediately becomes: *well, if I need papers, how do I get papers published?* Turns out that journals love positive results. In other words, significant results. This study looked at published work across all these different fields, and found that in space science, about 75% of published papers had significant results. In other words, 75% of papers were able to confirm their tested hypothesis. Great right? Then we have statistics for many other research field, e.g., the field of psychology or psychiatry, which can be found all the way at he bottom of this graph. Psychology doing even better than space science, because in their papers, 92% of results are significant. Is that not great? 92% of their papers support the hypothesis researchers had generated. And clearly what this also means is that, if I want to get published in a psychological journal, all I need to do is bring significant, positive results. Seems easy enough, but how do we do that?

Instructor Notes: Depending on your learners, highlight the relevant field or a related field.

Publication bias

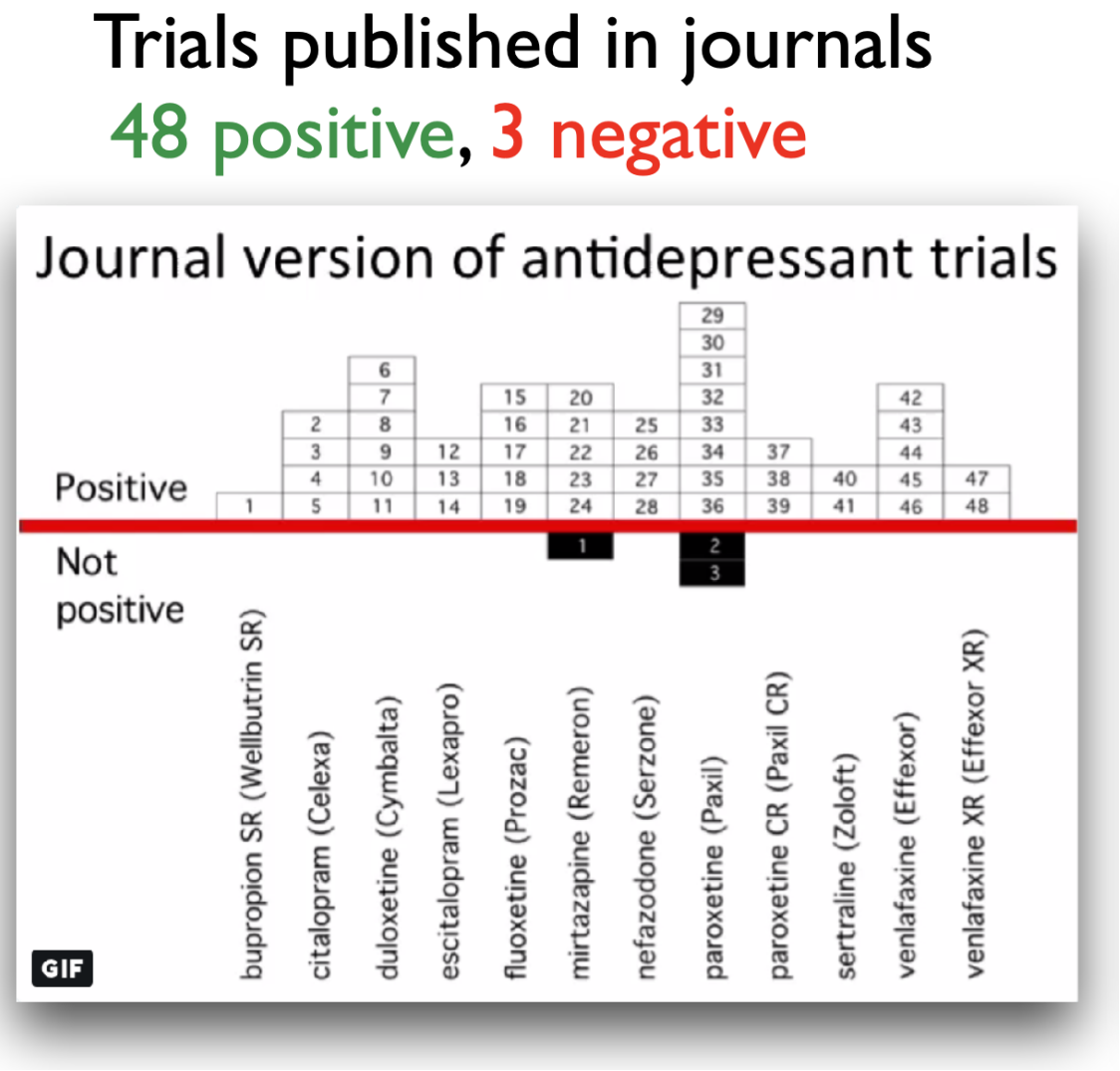
"If my study *works*, I can publish it. *If it does not, let's hide it the drawer.*"

Publication bias happens when studies with *positive* or *significant* results are much more likely to be published than studies with negative or non-significant results.

Song, F., Hooper, L., & Loke, Y. K. (2013). Publication bias: what is it? How do we measure it? How do we avoid it?. *Open Access Journal of Clinical Trials*, 71-81. https://doi.org/10.2147/OAJCT.S34419

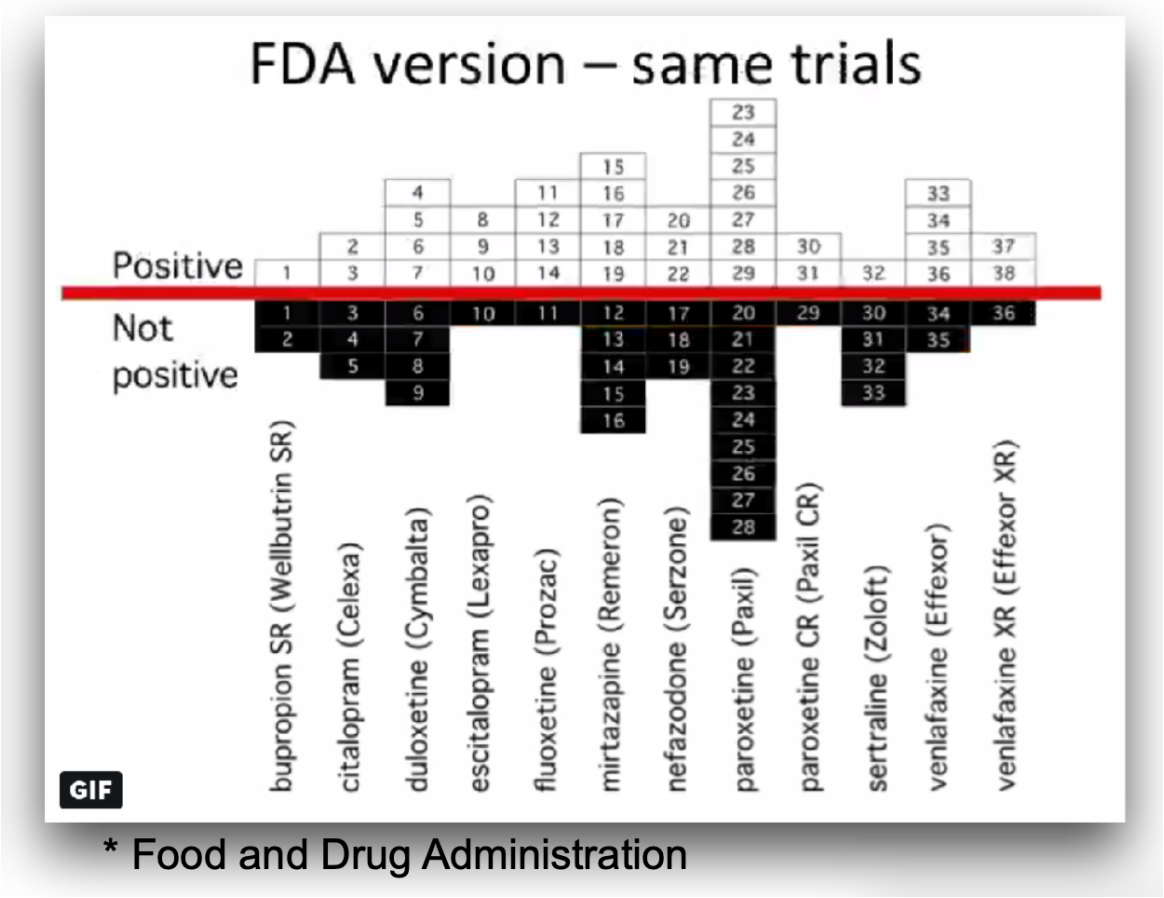
Presenter Notes: The simplest way to obtain many papers with positive results is to, well, publish the papers with positive results. The logic behind it is: If my study works and the results are good (=positive) I can publish. If not, let's pretend the study never happened. This can be captured under the term *publication bias*. Publication bias happens when studies with positive or significant results are much more likely to be published than studies with negative or non-significant results.

Example publication bias



Trials preregistered at FDA*

38 positive, 36 negative



Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252-260. https://doi.org/10.1056/NEJMsa065779

Presenter Notes: Turner and colleagues (2008) studied how selective publication bias affects research on antidepressant drugs. They compared antidepressant trials registered with the FDA to what was actually published in journals. They found that studies with positive results were much more likely to be published than those with negative or uncertain results. This made antidepressants seem more effective than they really were. Specifically, their analysis showed that about half of all FDA-registered trials did not show clear benefits, but almost all positive studies were published. In contrast, most negative studies were either not published at all or were written in a way that made the results look better than they were. This bias gave an inflated view of how well antidepressants work.

Instructor Notes: Additional details about the study and the conclusions: In this study, the authors obtained the archive of clinical trials for 12 antidepressant drugs submitted the the US Food and Drug Administration (FDA). In total, the trials involved more than 12000 patients. They performed a systematic search to identify which of these trials had been published in the literature. For those that were published, the compared the published versions and results with the versions and results in the FDA documentation. There were 74 FDA-registered trials. Of those, 38 trials showed postive effects of antidepressant drugs, 36 showed negative effects of antidepressant drugs. This is a 51% - 49% split of results: about half of the FDA-registered trials showed a positive effect of antidepressant drugs, the other half showed a negative effect of antidepressant drugs. 37 of the 38 trials that showed positive effects were published in the literature, basically all of them except 1 trial. However, of the 36 trials that showed negative or no effects of antidepressant drugs, only 3 were published that included those exact findings. 22 of the 36 trials were not published at all, and the remaining 11 trials were published with the outcomes reframed in a more positive way. In the literature, in the end, we ended up with a total of 51 trials being published, 48 of which showed positive effects of antidepressant drugs on depression. That is 94% of this sample of papers. Only 6% (3 papers) showed a negative effect of antidepressant drugs on depression. As a consequence, those of us searching the literature for these papers, will get a distorted via of the effectiveness of antidepressants. Therefore, the authors were able to show a substantial publication bias in the antidepressant trial literature.

More than just publication bias

(a) All conducted studies
(b) **Study publication bias**: non-publication of an entire study
(c) **Outcome reporting bias**: non-publication of negative outcomes within a published article or switching the status of (non-significant) primary and (significant) secondary outcomes
(d) **Spin**: authors conclude that the treatment is effective despite non-significant results on the primary outcome
(e) **Citation bias**: Studies with positive results receive more citations than negative studies

De Vries, Y. A., Roest, A. M., de Jonge, P., Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine, 48*(15), 2453–2455. https://doi.org/10.1017/S0033291718001873

Presenter Notes: This paper by de Vries and colleagues (2018) showed that there might be more than "just" publication bias at play. What we are looking at here in this graph is the following:

a) The authors assembled a cohort of 105 randomised controlled trials (RCTs) of antidepressant treatments for depression, drawn from the Food and Drug Administration (FDA) registration/trial-submission database (74 trials from an earlier study plus 31 newer ones). They found that of these 105 trials: 53 (50%) trials were showed statistically significant positive effects of antidepressant treatments and 52 (50%) trials did not show a statistically positive effect of antidepressant treatment.

b) They found that all of the trials with positive results, so a totlal of 53 trials, we then published. However, as you can see here, only about half of the trials with negative results were published. This represents the classical study publication bias we discussed previously.

c) Of the trials that were published, something interesting happened to the ones with negative results. Upon publishing, some trials were published in a way that made them appear positive by omitting unfavourable outcomes, or by switching the designation of primary versus secondary outcomes.

d) You can see now how the number of published trials with negative outcomes are slowly decreasing. The next reduction of negative published trials comes from what we call spin. For some trials, even when primary outcomes are non-significant, authors may write the abstract in a positive way. In other words, authors interpret statistically negative results in a more positive light than supported by the data.

e) Finally, we also have citation bias, which is simply the phenomenon that studies with positive results get cited more, as you can see here by these different circles. The ones that come out very strongly are the ones that get cited more often, and you can see that all of those are the trials with positive results. In contrast, the ones with "spin results" or negative results, get cited much less often. In the study's cohort, after accounting for publication and outcome-reporting biases, the remaining published positive studies tended to have higher visibility (more citations) than the published negative ones. The authors also note that positive trials were published in journals with higher median impact factor than negative ones (for antidepressants: 3.5 vs 2.4), which likely contributes to higher citations and visibility.

Instructor Notes: This is a complex figure, take your time with walking learners through this step-by-step.

Confirmation Bias

a.k.a. "Thought so already" bias



Presenter Notes: We are humans. And as humans, we often fall victims to our own inherent biases, sometimes or quite often without us noticing. Sometimes we cannot help but to feel like we just unlocked evidence or proof for something that we suspected all along. And it can be that it is true, or just that we stopped looking for more evidence and arguments as soon as we found the evidence that corresponded to our way of thinking. This is called the **confirmation bias**. Confirmation bias refers to the tendency to look for, interpret, or remember information in a way that confirms your pre-existing beliefs or expectations, while ignoring or downplaying evidence that contradicts them. In research, this can lead to selective attention to data, skewed analyses, or overconfident conclusions. For example; imagine a researcher believes that a new study technique helps students perform better. While analyzing results, they might focus on the small group of students who improved and overlook those whose performance stayed the same or worsened. Even unintentionally, this bias can make the findings seem stronger or more conclusive than they really are.

Hindsight Bias

a.k.a. "I knew it" bias



- The **tendency to believe**, after seeing the results of a study, that the **outcome was obvious** or predictable all along, even if it was not.

Presenter Notes: Another bias we are unconsciously struggling with in the research process is this **hindsight bias**. It is the tendency to believe, after seeing the results of a study, that the outcome was obvious or predictable all along, even if it was not. For example: Imagine a study finds that people who exercise regularly have lower stress levels. After seeing the results, a researcher might think, "Of course, I knew exercise would reduce stress!" Even though before the study, this connection was not at all certain. Hindsight bias can make results seem more predictable than they really were, which can lead researchers or readers to overestimate the certainty of findings or ignore alternative explanations.

Your turn! Match the bias to the description

Bias	Description
1. Confirmation bias	**A.** Non-publication of negative outcomes within a paper, or switching non-significant primary outcomes with significant secondary ones.
2. Spin	**B.** Studies with positive results receive more citations than negative studies.
3. Study publication bias	**C.** After learning the outcome, believing "I knew it all along."
4. Hindsight bias	**D.** Non-publication of an entire study (e.g., trials with null results never submitted).
5. Outcome reporting bias	**E.** Tendency to seek or interpret information in ways that confirm existing beliefs.
6. Citation bias	**F.** Authors conclude the treatment is effective despite non-significant primary outcomes.

Work in pairs. Take 5 minutes for this exercise.

Presenter Notes: Time for another exercises, To consolidate the different biases a little, let's try and match the biases to their respective descriptions. Work in pairs for 5 minutes. Then we discuss in the group.

Instructor Notes: Solutions: 1E, 2F, 3D, 4C, 5A, 6B. Give about 5 minutes for this exercise and 2 minutes for discussing the correct results.

Pre-break survey

Trust levels

Based on what you learnt so far, how would you *now* rate your trust in published scientific findings on a scale from 1 - 5? (1 = not trusting any of the findings, 2 = trusting only some findings, 3 = trusting about half of the findings, 4 = trusting the majority of the findings, 5 = trusting all findings)

a. 1

b. 2

c. 3

d. 4

e. 5

Presenter Notes: Coming back to the question from the beginning, what is your current trust in research with this information from the slides?

Term: Replicability

What does *replicability* in research mean?

a. Obtaining the same results using the original dataset and code

b. Obtaining consistent results when a new study collects new data using the same methods

c. Publishing results in more than one journal

d. Repeating the statistical analysis multiple times

Instructor Notes: The correct answer is b.

Term: Publication bias

What is *publication bias* in research?

a. The tendency for journals to publish studies only from well-known researchers

b. The requirement that all published studies must be peer-reviewed

c. The practice of publishing the same study in multiple journals

d. The tendency for studies with positive results to be published more often than studies with non-significant or negative results

Instructor Notes: The correct answer is d.

Break! 15 minutes

Post-break survey discussion

What do we see in the results?

Instructor Notes: Compare between the answers given before the session, particularly for trust levels.

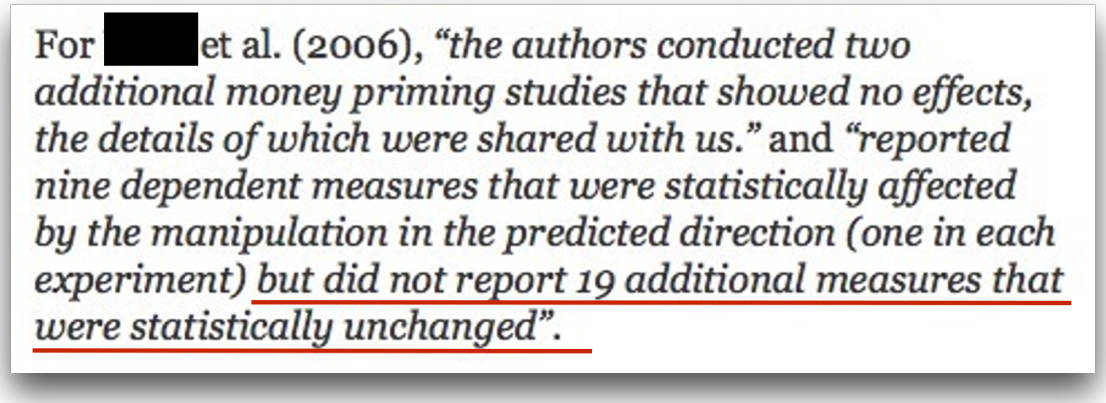
---

p-hacking and other problems ...

> ⚠ Warning
>
> Do **not** try the following at home.

"Hack" 1: Add *lots* of outcome variables

- For **two outcome variables**: False positive rate increases from 5% to **9.5%**
- For **five outcome variables** (and one-sided testing): False positive rate increases from 5% to **41%**



Presenter Notes: Let us say you designed an experiment, generated a hypothesis and collected some data. Now you are analysising your data and are running a statistical test. When you run a statistical test, you're checking whether your data could have happened just by random chance. If the p-value, which is the probability of getting your result if there's really no effect, is less than 0.05, you say the result is "statistically significant." That means there's less than a 5% chance that the result happened by luck, assuming the null hypothesis is true. With a p-value below 0.05, you are saying, "I'm fairly confident this effect is real. There's only about a 5% or a 1 in 20 chance that I incorrectly conclude that an effect or a difference between two conditions exists". In other words, it is the probability that I find a false positive.

However, when you run multiple tests, each one carries its own 5% chance of producing a false positive, and these chances add up. As a result, the overall likelihood of finding at least one false positive across all tests, known as the family-wise error rate, becomes much higher than 5%. For example, if you test two outcome variables, each at the 5% significance level, the combined chance that at least one result is a false positive increases to about 9.5%.

If you test five outcome variables, that overall false positive rate can rise dramatically, up to around 41% for one-sided tests. In simple terms, the more outcomes you test, the greater your risk of finding a "significant" result purely by chance. So your chance of a false positive results is extremely high. I would therefore recommend including as many outcome variables as possible, you are sure to find a significant results for at least one of them.

Instructor Notes: Additional information for students: To reduce the false-positive rate, researchers often apply corrections, such as the Bonferroni adjustment, to keep the overall false positive rate close to the intended 5%.

"Hack" 2: Run as many comparisons as possible

- Run as many different comparisons on **different outcomes, subgroups, time windows** etc. as you can
- **Only** report the comparisons that produced a statistically significant result

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The Chrysalis Effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376–399. https://doi.org/10.1177/0149206314527133

Presenter Notes: When you have more outcome variables, this often also means that you can run more statistical tests. Each time you try another independent test at = 0.05 you have a 5% chance of a Type I error for that test. If you perform many uncorrected tests, the chance that at least one will be (falsely) significant rises quickly. For example, doing 20 independent tests at = 0.05 gives a probability of at least one false positive of$1 - (1 - 0.05)^2 = 1 - 0.95^2$ 0.6415 (64%). Therefore, the trick here is to just run as many comparisons as possible on different groups, different time windows, different subsections of

your data, you will be guaranteed to find significant results. Then, of course, only report those comparisons that produced the significant result. Just run many more statistical tests and report only the ones you like. Easy, right?

"Hack" 2: Run as many comparisons as possible

Example: Subgroup Analyses

Research question: *Do aggressive primes trigger aggressive behavior?*

> "A study in Turner, Layton, and Simons (1975) collects a larger sample of men and women driving vehicles of all years. The design was a 2 (Rifle: present, absent) × 2 (Bumper Sticker:"Vengeance", absent) design with 200 subjects."

presumably, no effect … (yet! Do not give up so easily)

Cited from Joe Hilgard's excellent blog post on the Weapon Priming Effect; Turner et al. (1975). Naturalistic studies of aggressive behavior: aggressive stimuli, victim visibility, and horn honking. JPSP, 31(6), 1098–1107.

Presenter Notes: In thr 1960ties, there was a study by Berkowitz and LePage who recruited 100 male university students to answer the question whether exposure to an aggressive stimulus, such as a weapon, make an already angry person behave more aggressively? Participants first received either one electric shock or seven electric shocks, supposedly from another student. Receiving seven shocks was intended to make participants more angry or aroused. The participant was then given an opportunity to administer electric shocks to that same person. Depending on the condition, there was either: a rifle and revolver near the shock apparatus, badminton racquets, or nothing nearby. The number of shocks the participant subsequently administered served as the measure of aggressive behavior.

A subsequent study from Turner, Layton, and Simons in 1975. The asked the question whether aggressive cues in the environment would make drives behave more aggressively. Their sample was 200 participants, including both men and women and drivers of different vehicle ages. The researchers used a 2 × 2 design. One factor was whether a rifle was present or absent in the vehicle. The other was whether the vehicle displayed a "Vengeance" bumper sticker or no bumper sticker. The researchers used horn honking as their measure of aggression.

A participant was stopped in traffic behind a car that had ither the rifle visible or not or the vengeance sticker visible or not. The car did not immediately move when the light went green. The researchers then recorded whether and how quickly the driver honked. They found no effect. Neither the rifle nor the sticker were linked to honking. But, not so fast!

"Hack" 2: Run as many comparisons as possible

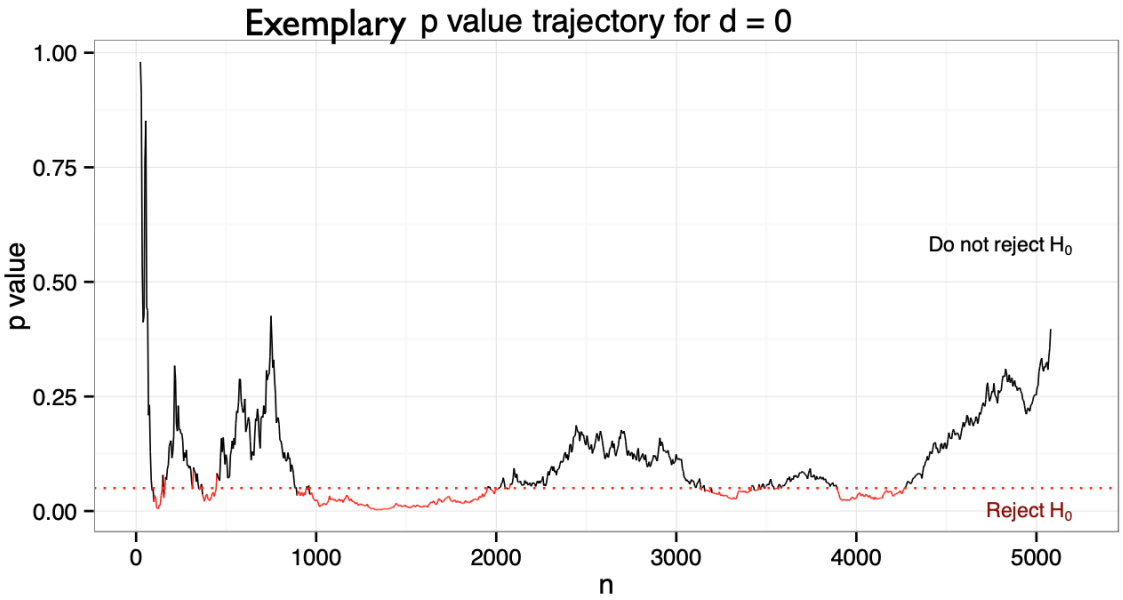
Example: Subgroup Analyses

> "They divide this further by driver's sex and by a median split on vehicle year. They find that the Rifle/Vengeance condition increased honking relative to the other three, **but only among newer-vehicle male drivers**, F(1, 129) = 4.03, p = .047. But then they report that the Rifle/Vengeance condition decreased honking among older-vehicle male drivers, F(1, 129) = 5.23, p = .024! No results were found among female drivers."

Cited from Joe Hilgard's excellent blog post on the Weapon Priming Effect; Turner et al. (1975). Naturalistic studies of aggressive behavior: aggressive stimuli, victim visibility, and horn honking. JPSP, 31(6), 1098–1107.

"Hack" 3: Optional stopping

- Scenario: You collect a sample of n=60 and obtained a p-value of .07.
- "Trending towards significance"! **What is your intuition how to proceed?**



Presenter Notes: Another p-hacking strategy is called peeking or **optional stopping**. In other words, you start doing your analysis already before all data have been collected and perform multiple interim analyses "just to have a look". With every interim analysis, you are, of course, increasing your chances of finding a false positive, or, what you would consider a "true" significant effect that is, unfortunately not true at all.

You do not have to take my word for it that this practice can increase the false positive rates, but there have been some people who looked at this. Their work demonstrated that when researchers test interim results multiple times and stop data collection once a significant result appears, the false positive rate does not stay at 5%. Instead, this rate increases exponentially with each additional interim analysis. Even a modest number of repeated tests can double or triple the chance of declaring a false positive. This inflation occurs because each interim look introduces another opportunity for a random fluctuation to cross the significance threshold. In other words, you can move from a 5% chance of incorrectly concluding that there is an effect to an 11% chance with only one additional analyses.

So, the moral of the story here is that you need to just do as many interim analyses as possible before you finish your data collection, and I guarantee you will obtain a significant results at some point. When you do, of course "there is no point in sticking to your original plan of minimum sample size, so to not waste time, you should also stop data collection immediately".

Instructor Notes: Explicitly highlight the sarcastic aspect of this slide, since there are large cultural differences in whether and how sarcasm is perceived as a linguistic cue and processed. Reframe this slide if you have the impression that the audience will not benefit from sarcasm.

"Hack" 3: Optional stopping

- Optional stopping: Collect an initial sample, analyze the results, add participants if the results are not significant
- But under H_0, the p-value is a random walk between 0 and 1, fluctuating forever.
- Eventually, it will "dip" below the critical threshold
- The false positive rate depends on the number of optional sample size increases:
  - One analysis: $= 5\%$
  - Two analyses: $= 11\%$
  - But with enough increases can be pushed to 100%!

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General), 132*, 235–244.

"Hack" 4: Drop participants you do not like

- Selectively **exclude data/ outliers** after seeing the results until the results are satisfying (*a.k.a. until significance has been reached*)

Presenter Notes: Another common form of p-hacking happens when researchers selectively remove outliers or participants from their dataset to make their results look more significant. In simple terms, an outlier is a data point that looks very different from the rest. This is maybe a participant who scored unusually high or low compared to the rest of the participants. Sometimes, removing outliers is justified, for example, if there was a clear measurement error. But in order to obtain significant results, you can also perform a first analysis, evaluate your p-values and think about excluding those participants that are just really, really biasing your data. Of course, this practice increases the chance of getting a "statistically significant" result by luck rather than truth, but at least, you have a statistically significant result. Some work by Simmons et al. showed that flexible data analysis decisions, including dropping outliers, could make almost any hypothesis appear statistically significant.

Instructor Notes: More information about Simmons et al. (2011): According to Simmons, Nelson, and Simonsohn (2011) in their paper "False-Positive Psychology" published in Psychological Science, even small and seemingly reasonable choices about including certain data from selected participants can dramatically increase the likelihood of finding a false positive result. In their experiments, they showed that flexible data analysis decisions, including dropping outliers, could make almost any hypothesis appear statistically significant.

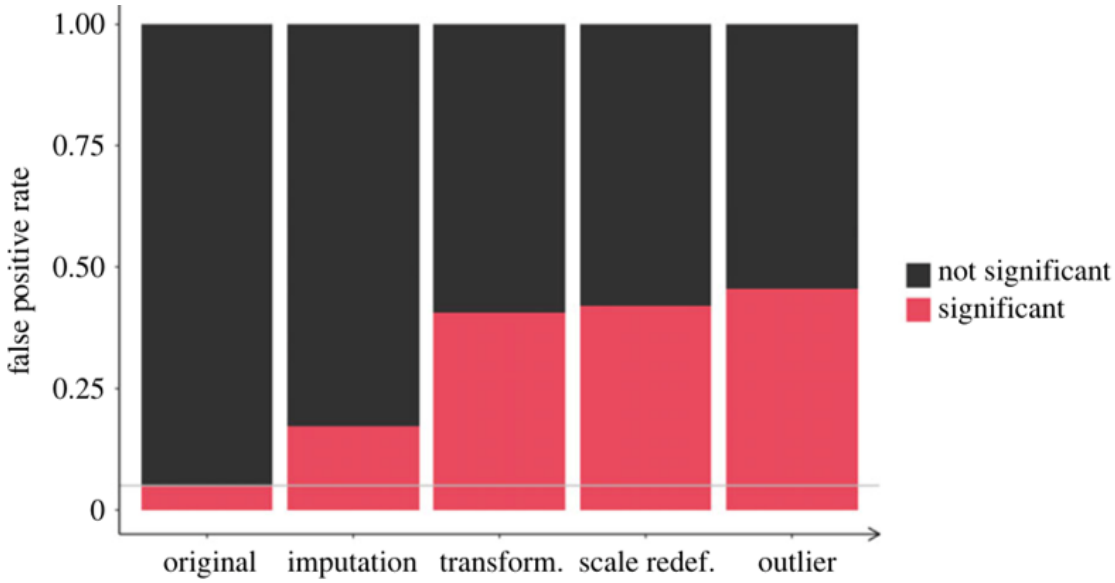
"Hack" 5: HARK-ing

- **H**ypothesizing **A**fter **R**esults are **K**nown = presenting an exploratory finding to match a hypothesis that was created only after analysing the results
- a.k.a. "The Texas sharpshooter fallacy"

Presenter Notes: Let's talk about another so-called "hack". Note that once again, what I will say next is meant purely in a sarcastic way. This tip to obtain positive results is probably one of my favorites. What you do is the following: you look at your results first, and then come up with a hypothesis afterward. Just forget about the original hypothesis you had and that you tested for. For example: imagine a study testing several relationships between variables. Only one of them turns out significant. Instead of reporting that this was an unexpected result, the researcher rewrites the introduction to make it look like that was their original prediction.

Of course, this practice undermines the credibility of scientific findings because it blurs the line between prediction and explanation. It makes results seem more confirmatory than they really are and increases the risk of false theories being accepted. But, what is important is that at the end you have a positive result and a good story that you can tell to the journal, so I am sure that you will get accepted.

Combining these "hacks"



- Combining some of these hacks (*a.k.a. questionable research practices*) can **raise false positive rates** from 5% to > 50%!
- With *p*-hacking, the logic of the p-value is **corrupted** and "*renders the reported p-values essentially uninterpretable.*"

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science, 10*(2), 220346. Wasserstein et al. (2016). The ASA's statement on p-values: context, process, and purpose. American Statistician.

Presenter Notes: Even seemingly minor practices, e.g., selectively reporting results, stopping data collection once a significant result appears, or testing multiple hypotheses without proper correction, can have a huge effect. When these practices are combined, they can inflate false positive rates dramatically, taking them from the expected 5% under proper statistical testing to over 50%. This means that more than half of the reported significant findings could be false alarms.

As a result, the logic of the p-value itself becomes corrupted. P-values are supposed to indicate the probability of observing the data if the null hypothesis is true. When questionable practices are used, this interpretation no longer holds. In other words, it renders reported p-values essentially uninterpretable, because they no longer reflect the true probability of the observed effect under rigorous, unbiased conditions.

A shocking report from an experienced researcher

A note to remember

> ⚠ Warning
>
> The so-called "hacks" on the past few slides represent **questionable research practices**. Do **not** try at home.

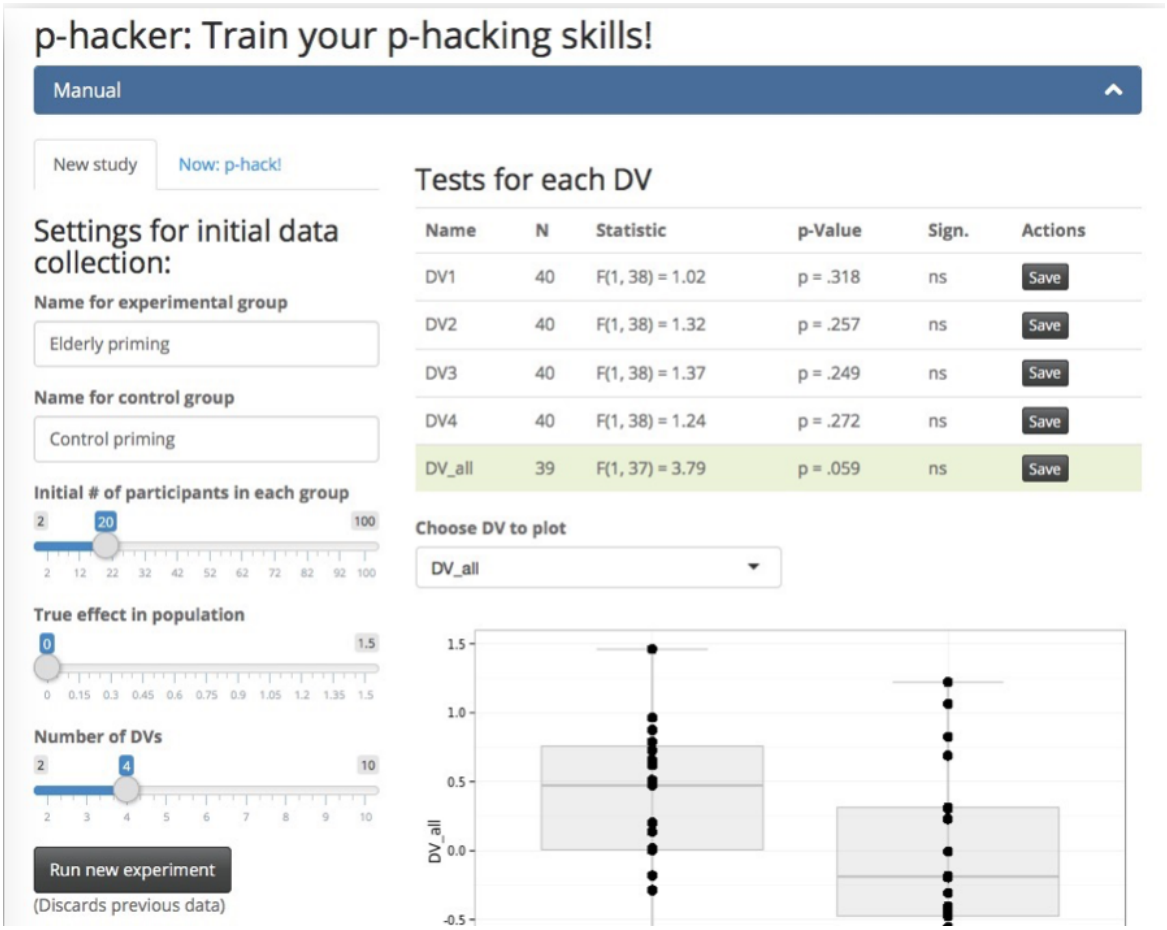
Presenter Notes: I want to be clear before we move on: the way the past slides were slide is deliberately sarcastic. It is meant to show what can happen if researchers engage in questionable research practices, like p-hacking, but it's **not a recommendation**.

Instructor Notes: Stress this point again to potential cultural differences in terms of processing sarcasm.

Practice *p*-hacking yourself

...with the P-Hacker app!

http://shinyapps.org/apps/p-hacker/



Your turn! Identifying p-hacking

Scenario: You are reviewing a study examining whether drinking white tea improves short-term memory, where the researchers report:

- Hypothesis: *White tea improves memory test scores.*
- Sample size: 28 participants per group (tea-drinkers vs. water-only-drinkers)
- Results:
  - Effect was "stronger in women" ($p = 0.049$)
  - Effect "even stronger when excluding two outliers" ($p = 0.044$)
  - No effect in men ($p = 0.31$)
  - Reaction time difference significant ($p = 0.046$)
- Conclusion: *The results show that white tea reliably improves cognitive performance.*

Which potential p-hacking strategies are at play here?

Work in pairs. Take 5 minutes for this exercise.

Presenter Notes: Let us get to work! I now have a task for you to reflect on different biases and/ or questionable research practices. Imagine that you are reviewing a study examining whether drinking white tea improves short-term memory. In the report, you read the information provided here. Which potential p-hacking strategies can you identify? Work in pairs, take 5 minutes for this task.

Instructor notes: Plan about 5 minutes for the discussion of this exercise.

Exercise - solutions

Scenario: You are reviewing a study examining whether drinking white tea improves short-term memory, where the researchers report:

- Hypothesis: *White tea improves memory test scores.*
- Sample size: 28 participants per group (tea-drinkers vs. water-only-drinkers)
- Results:
  - Effect was "stronger in **women**" ($p = \mathbf{0.049}$)
  - Effect "even stronger when excluding two outliers" ($p = \mathbf{0.044}$)
  - No effect in **men** ($p = 0.31$)
  - **Reaction time** difference significant ($p = \mathbf{0.046}$)
- Conclusion: The results show that white tea reliably improves **cognitive performance**.

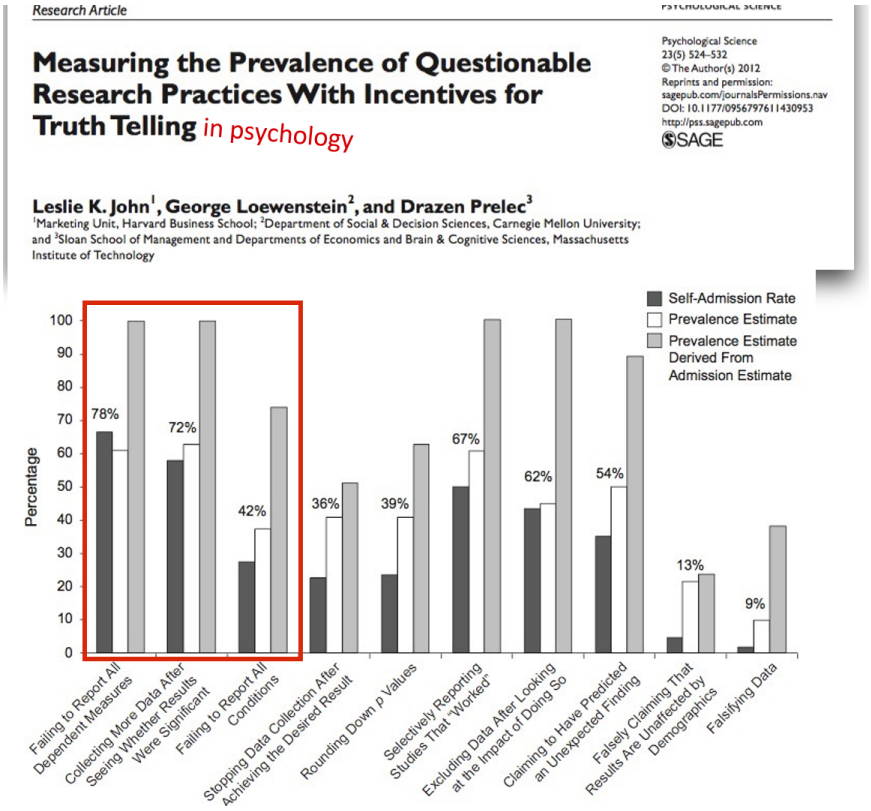
Instructor Notes: Possible answers can include:

- Large number of uncorrected comparisons with selective reporting
- The conclusion variable does not match the original hypothesis
- Research question too vague and non-directional given the analysis
- Results report comparisons that are not specified in the research question
- Many p values bunching just below 0.05
- Lack of pre registration or analysis plan when claims are confirmatory
- Outcome variables added that were not specified in the hypothesis: e.g., reaction time

---

Does this *actually* happen?

Come on. Surely not..?



John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science, 23*(5), 524-532.

Presenter Notes: At this point you might be thinking well ok, we heard that p-hacking can be accidental. Surely then that means it does not happen that often? Surely you can count the cases of this happening on one hand. John and colleagues surveyed over 2,000 psychologists about their involvement in questionable research practices.

In this survey, they asked respondents if they had ever failed to report all outcome variables or they had ever collected more data after seeing that the collected data did not result in the significant results they wanted. The dark bar you see here are self-admissions rates, which for these two practices in particular are above 70%, which is extremely high.

In their paper, they also used some estimation methods I won't go into detail but just see that, well, over 70% are **telling** us that they engage in these practices, what would this percentage be in real life? This is what is shown in the light grey bar, which for these two practices specifically estimates the true percentage of engaging in these two practices at almost 100%, meaning that everyone has done this at some point in their research career.

Instructor Notes: Quoting the relevant part of the paper: "The impact of truth-telling incentives on self-admissions of questionable research practices was positive, and this impact was greater for practices that respondents judged to be less defensible. Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm."

.. and across fields?

- Survey among 6,813 academic researchers in The Netherlands: **Self-reported prevalence of fabrication and falsification in the last 3 years**

		Disciplinary field			
QRP	**Description (In the last three years.)**	**Life and medical sciences**	**Social and behavioural sciences**	**Natural and engineering sciences**	**Arts and humanities**
Fabrication	Making up of data or results				
Falsification	Manipulating research materials, data or results				
Any FF	Fabrication and/or Falsification				

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE, 17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Presenter Notes: So far, we focused on questionable research practices. Of course, there is also the next level, namely fabrication and falsification, which both constitute outright scientific misconduct. The authors conducted a large-scale cross-sectional survey (the National Survey on Research Integrity) targeting academic researchers across all disciplines and ranks in the Netherlands. Its aim was to estimate the prevalence of fabrication and falsification and a range of questionable research practices (QRPs) over the preceding three-year period, and to examine factors potentially associated with higher or lower engagement in these behaviors. Over 6,800 researchers completed the survey. What are your guesses here on the percentages they found?

Instructor Notes: Collect some guesses before moving on to showing the solution.

.. and across fields?

- Survey among 6,813 academic researchers in The Netherlands: **Self-reported prevalence of fabrication and falsification in the last 3 years**

		Disciplinary field			
QRP	**Description (In the last three years.)**	**Life and medical sciences**	**Social and behavioural sciences**	**Natural and engineering sciences**	**Arts and humanities**
Fabrication	Making up of data or results	5.5 (3.2, 7.7)	4.8 (2.2, 7.5)	2.5 (0, 5.5)	0.7 (0, 5.1)
Falsification	Manipulating research materials, data or results	4.9 (2.7, 7.2)	2.0 (0, 4.6)	5.3 (2.2, 8.4)	6.1 (1.4, 10.9)
Any FF	Fabrication and/or Falsification	10.4 (7.1, 13.7)	5.7 (1.8, 9.5)	7.6 (3.1, 12.1)	8.4 (1.6, 15.3)

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE, 17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Presenter Notes: The key findings from their study were quite shocking. The prevalence of **fabrication**, a.k.a. making up data, stands at about 4.3% (with a 95% confidence interval of 2.9 to 5.7%), while **falsification**, a.k.a. outright manipulating data, occurs at roughly 4.2% (95% CI 2.8 to 5.6%). These rates vary across disciplines: for example, medical fields tend to report higher rates, whereas arts and humanities report lower. Psychology shows a self-reported fabrication rate of nearly 5% and falsification around 2%, indicating it is somewhat in the middle range.

When it comes to questionable research practices, which include 11 types of behaviors assessed, the prevalence ranges widely. The least frequent QRPs occur in about 0.6% of respondents, while the most common QRP is reported by as many as 17.5% (95% CI 16.4 to 18.7%). Notably, over half of the respondents (51.3%) admitted to frequently engaging in at least one QRP in the past three years.

Further analysis reveals that junior researchers, such as PhD candidates, and male researchers have higher odds of frequently reporting at least one QRP. On the other hand, individuals who strongly subscribe to scientific norms, a.k.a. those that value honest research, and those who believe that peer reviewers are likely to detect misconduct show lower odds of engaging in misconduct or QRPs. Conversely, researchers who experience higher publication pressure are more likely to frequently engage in QRPs, with a 22% increase in odds (OR ~1.22, 95% CI 1.14 to 1.30).

Instructor Notes: Ask for the thoughts of your learners.

(Un)Intentional?

- Intentional?
  - "Evil researcher" who only cares about his/her career and not at all about truth-seeking?
  - „*We urge the social science community to redefine p-hacking as a series of deceptive research practices rather than ones that are merely questionable.*" (Craig et al., 2020)
- Unintentional?
  - **Lack of education/knowledge**?
  - Wrong/uncritical **standards of the field**?
  - **Pushed** by supervisors, reviewers, or editors? http://bulliedintobadscience.org/
  - **Simply being human?**
- Distorting effects on the published record are probably comparable, but the ethical evaluations differs strongly.

Presenter Notes: Summarising from the previous slides I showed, we now know that questionable research practices and scientific misconduct are extremely prevalent and frequent. The question here is: why? Are we doing this **intentionally**, for example to publish more papers and to advance our careers? Because we don't care about the truth and about producing good quality, reliable research?

Or maybe this is all **unintentional**? Perhaps we did not fully know that some of things we learnt from our PIs or supervisors are considered bad practice? Maybe we are being convinced by supervisors? Maybe there are just unclear standards of what is considered "good quality research" in the field? Or maybe we are also just humans? We already talked about inherent biases, but perhaps we also just make mistakes or are unaware of our own weaknesses?

(Honest) errors and mistakes

Human errors and honest mistakes

"90% Excel-Gate"

Presenter Notes: In their 2010 paper Growth in a Time of Debt, economists Carmen Reinhart and Kenneth Rogoff claimed that when government debt exceeds 90% of GDP,

economic growth falls sharply. Their findings strongly influenced debates about austerity policies. However, in 2013, Thomas Herndon, Michael Ash, and Robert Pollin from the University of Massachusetts–Amherst re-examined the data and found serious flaws: an Excel coding error that excluded several countries. In essence, what had happened was that when using a formula, the authors had made a mistake when selecting the range of cells where to apply the formula too. In this, they accidentally omitted some countries from the calculations. The conclusion was that for advanced economies, when the government gross debt exceeds about 90% of GDP, annual real GDP growth falls sharply. This conclusion informed policies in the respective countries.

Instructor Notes: Additional details about the error: The omitted rows related to countries with high debt and growth (for example Belgium) that should have been part of the calculation. Because of this omission, the reported average for the >90% debt category was around –0.1% growth. Once corrected, the average growth rate for high-debt countries rose from –0.1% to about +2.2%, showing no clear "90% tipping point." The original authors later acknowledged the spreadsheet error and issued a correction.

Human errors and honest mistakes

"90% Excel-Gate" - Lessons learned

> **!** Important
>
> The most important point of the story: The original authors **shared their raw data**, which made it possible to **correct** the honest mistake!

Presenter Notes: The case became a landmark example of how data errors, selective reporting, and methodological choices can mislead economic policy, and of why transparency and replication are essential in research. Only because the data were available, the mistake could be found and corrected!

Statistical errors

A Decade of Replications: Lessons from the Quarterly Journal of Political Science

Posted on December 9, 2014 by nicholaseubank

- Reproducible analysis **code and open data required** at submission - "in-house checking" in review process
- 54% of all submissions had results in the paper that **did not match** the computed results from the code
  - Wrong signs, wrong labeling of regression coefficients, errors in sample sizes, wrong descriptive stats

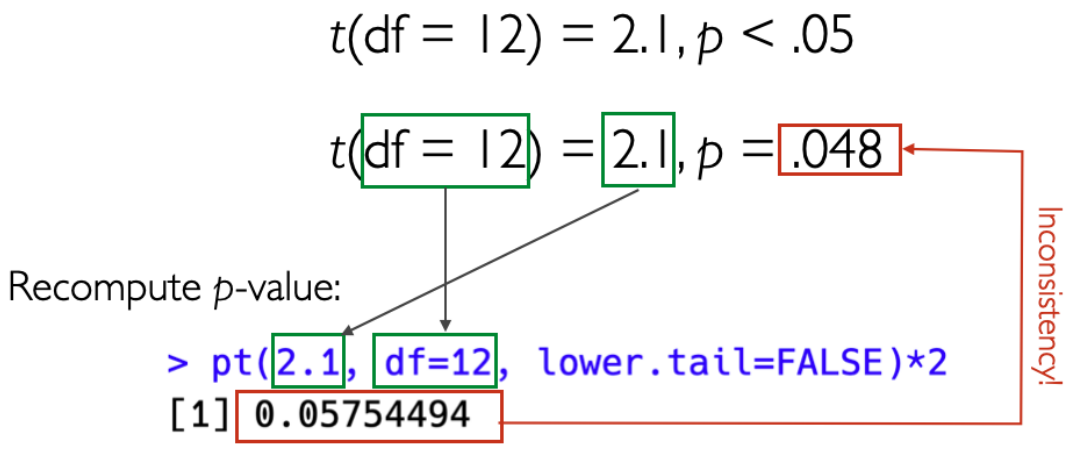
Eubank, N. (2016). Lessons from a decade of replications at the quarterly journal of political science. *PS: Political Science & Politics, 49*(2), 273-276 https://doi.org/10.1017/S1049096516000196

Presenter Notes: This is an interesting paper which looked at code and data packages that were required at submission to the Quarterly Journal of Political Science. They took the analysis code that and the data that were submitted, reran the analysis and checked whether the results matched what was described in the paper.

It finds that of the 24 empirical papers subjected to in-house replication review since September 2012, only four packages did not require any modifications. Most troubling, 14 packages (58%) had results in the paper that differed from those generated by the author's own code. The authors identified a number of mistakes in the analysis scripts that were submitted and in the results that were published, for example that the wrong signs used for reporting, coefficients were labelled the wrong way, descriptive statistics were imprecise or incorrect.

Based on these experiences, this article presents a set of guidelines for authors and journals for improving the reliability and usability of replication packages.

Statistical inconsistencies



Statistical inconsistencies

statcheck

- Does the statistical conclusion change? „strong/gross inconsistency“
- Psychology:
  - 16,695 scanned papers with **statcheck tool** (Nuijten et al., 2015)
  - 50% of papers contain statistical inconsistencies
  - 13% contained **strong errors** (i.e., where the statistical conclusion changes).
- Economics
  - 3,677 scanned papers with DORIS tool (Bruns et al., 2023)
  - 36% of papers contain statistical inconsistencies
  - 15% contained strong errors (i.e., where the statistical conclusion changes).
- Other fields (e.g. Sociology):
  - Not even (automatically) checkable, due to unstandardized reporting practices.
- … and these are only the detected errors of one type (stat. inconsistency)!

Nuijten et al. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods, 48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2. Bruns et al.(2023). Statistical reporting errors in economics [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/mbx62

Presenter Notes: Researchers have also examined how often published psychology papers contain statistical errors. Using the *Statcheck* tool (Nuijten et al., 2015), over 16,000 papers were automatically scanned to compare reported test statistics with their p-values. The results were worrying: about 50% of papers contained at least one statistical inconsistency, and around 13% had serious errors, meaning the reported statistical conclusion (for example, whether a result was “significant”) would actually change if calculated correctly.

More recent work by Crüwell et al. (2022) found that the numerical results in fewer than 30% of papers can be fully reproduced, even when all the information needed is available.

I want to point out here that these are only errors of the statistical type, other types of errors were not examined here. I think you see the picture that I am painting here and I am hoping to get across that these are some serious challenges to research.

Reproducibility success

- Crüwell et al. (2023), Psychology: Numerical results of **less than 30% of papers** can be **reproduced** (Crüwell et al., 2023)
- Krähmer et al. (2026), Sociology:
  - Only ***3% of authors** share analysis **code proactively**
  - Upon request, 34% shared their code.
    - Of these, 51% were perfectly reproducible.
  - If non-shared cases are classified as non-reproducible, 18% of results reproduced.

Krähmer, D., Schächtele, L., & Auspurg, K. (2026). Crüwell et al. (2023).

Presenter Notes: Crüwell and colleagues found that fewer than 30% of psychology papers they examined could have their numerical results reproduced. A recent study in sociology by Krähmer et al. paints a similar picture. Only 3% of authors proactively shared their analysis code. When researchers contacted authors, 34% shared their code, and only about half of those cases were perfectly reproducible. If we treat cases where code was not shared as non-reproducible, only 18% of results could be reproduced.

What does this all mean?

Bias + (maybe unintentional) *p*-hacking + human (honest) mistakes = untrustworthy research findings?

> **i** Note
>
> - Published findings across fields to be viewed with **caution**?
> - "We know" –> **"We *think* we know"**?
> - More research to **verify** existing "truths"?

Presenter Notes: Taking all of this together, where does this leave us? We have several elements which, as we have explored in the last few slides, contribute to non-robust research. We have different types of biases, accidential or not so accidental p-hacking, and we have the human component of just making mistakes from time to time in the process. What conclusions can we draw about research, then?

Perhaps there should be more caution around published papers across fields and we should not view findings as the ultimate truth, just because they were published? Perhaps we should also shift from "we know" to "we think we know" pending further investigations and perhaps pending replication? And finally, perhaps our efforts should also go into checking these so-called truths we think exist in the academic world and perform more replication studies?

Instructor Notes: This is a critical reflection moment, give enough time between this slide and the next slide for learners to digest and integrate what they learned.

A romanticized idea of research?

Presenter Notes: In other words, perhaps we have a romanticized view of research? Typically we think of research as something along the following lines: The smartest heads in the world immerse themselves into a research topic for years. In that process, they become the experts, nobody knows more about that topic.The boundaries of knowledge have been pushed forward. When the researchers are confident in their findings, they publish them in the best scientific journals, with the highest standards of quality, rigor, and integrity.

Instructor Notes: Encourage a mental shift by contrasting how we used to think about research and the current status quo discussed over the past slides. Prepare the "stage" for a new approach to research and guide your learners towards the question of "but then how can we do research differently"?

Going even further: The perpetuating role of AI

Emsley, R. (2023). ChatGPT: These are not hallucinations - they're fabrications and falsifications. *Schizophrenia, 9*(1), 52. Lautrup, A. D., Hyrup, T., Schneider-Kamp, A., Dahl, M., Lindholt, J. S., & Schneider-Kamp, P. (2023). Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. *Open Heart, 10*(2). Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A., & Maes, P. (2025). People Overtrust AI-Generated Medical Advice despite Low Accuracy. *NEJM AI, 2*(6), AIoa2300015.

Presenter Notes: Thinking about how mistakes happen, how accidental p-hacking happens, how there is pressure to publish positive results, how we are, to some extent, victims of our own biases without noticing, I think you get an image of the low credibility of research that is currently out there. Now we are also facing a new challenge in the wake of artificial intelligence. We know that ChatGPT and other large language models are routinely used by people to, for example, obtain medical advice. Given what we heard before, how accurate do you believe this information provided by the chatbots is? Papers demonstrate by now that the accuracy is low, yet people turn to ChatGPT for medical advice. We also know about so-called hallucinations of ChatGPT. Some researchers view this as the "modern form of falsification and fabrication", just like we discussed in the context of scientific misconduct. But what if the research feeding into this advice that chatbots provide and the "papers" they are suggesting to you for your essay are already of low quality? Is AI then not simply distributing non-credible and flawed research, without the warning message that this might be the case?

Instructor Notes: Bonus slide depending on the interest of your learners.

Thesis:

Our current incentives foster questionable research practices, which decrease the truth value of our shared knowledge.

What is good for the individual careers of researchers leads to a collective fiasko.

Researchers who do it right (i.e., high power, no QRPs, transparency) have a clear competitive disadvantage.

Anti-Thesis:

Society pays for us that we generate valid and robust knowledge.

Our incentives should be chosen in a way that they foster good science.

Researchers who do it right should be supported and promoted.

Now what?



This image was taken from the Geograph project collection. The copyright on this image is owned by Chris Martin and is licensed for reuse under the Creative Commons Attribution-ShareAlike 2.0 license.

Presenter Notes: At this point, you might think that research is doomed. That perhaps humanity is doomed along with it. Despite everything we talked about today, this could not be further from the truth. There are different approaches to research, which mitigate many of these challenges we discussed today. These approaches have the aim to raise scientific standards and the quality of research, so that society and researchers can trust the published work.

A new way of doing research

Open Research *a.k.a.* **A scientific framework for the 21. century**

Presenter Notes: What I mean with this alternative approach is Open Research, which is a framework for how to do research in the 21st century. We will have the entire following session to dive deeper into what Open Research even means or how it can possibly make research more credible. The focus of the next session will all be about how Open Research can change the game.

Instructor Notes: Make the shift from describing the problem, threats and challenges of research to a more solution-focused framework of reference.

To be continued ...

Presenter Notes: What exactly Open Research entails will be the focus of the following session. **Instructor Notes**: Emphasize that Open Research will be discussed in great detail in the upcoming session and that today's focus were on understanding the current problems in research.

Reflection activity

One-minute paper: Imagine you would have to explain the current challenges in research you heard about today to a friend. Write down what you would say to them.

Presenter Notes: Wrapping things up: Imagine you would have to explain the current challenges in research you heard about today to a friend. Write down what you would say to them.

Instructor Notes: Ask for 2-3 volunteers willing to share their papers in class. Alternatively, have your learners either hand you in the pieces of paper they wrote this on, or ask your learners to send you their one-minute-papers by email.

Take-home message

What are you taking away from today?

> Remember: There are solutions!
>
> Research is **not "doomed" - on the contrary**. More on this in the *next* session!

References

Craig, R., Cox, A., Tourish, D., & Thorpe, A. (2020). Using retracted journal articles in psychology to understand research misconduct in the social sciences: What is to be done? *Research Policy*, *49*(4), 103930. https://doi.org/10.1016/j.respol.2020.103930

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a badge? A computational reproducibility investigation of the Open Data Badge Policy in one issue of Psychological Science. *Psychological Science*, *34*(4), 512–522. https://doi.org/10.1177/09567976221140828

Schmidt, R., Curry, S., & Hatch, A. (2021). Creating SPACE to evolve academic assessment. *eLife*, *10*, e70929. https://doi.org/10.7554/eLife.70929

Contribution statement

Creator: Von Grebmer zu Wolfsthurn, Sarah (0000-0002-6413-3895)

Reviewer: Schönbrodt, Felix (0000-0002-8282-3910)

Consultant: Ihle, Malika (0000-0002-3242-5981)

License and citation statement

This work was originally created by Felix Schönbrodt under a Creative Commons Attribution 4.0 International License (CC-BY 4.0). This current work by Sarah von Grebmer zu Wolfsthurn, Malika Ihle and Felix Schönbrodt is licensed under a Creative Commons Attribution 4.0 SA International License (CC-BY-SA 4.0). Code snippets are dedicated to the public domain and licenced under a Creative Commons Universal Licence (CC0 1.0).

Please cite these materials as: Von Grebmer zu Wolfsthurn, S., Ihle, M., & Schönbrodt, F. (2026). *Threats to credible research: Where are we at?* [Quarto slides]. LMU Open Science Center. https://lmu-osc.github.io/teaching-material-for-novices/open-research-principles/reproducibility-replicability/

Presenter Notes: The Creative Commons Attribution 4.0 International SA License, or CC BY-SA 4.0, allows others to copy, share, and adapt a work in any medium, including for commercial purposes. These permissions are broad and cannot be withdrawn as long as the license terms are followed. The main requirement is attribution: users must give appropriate credit to the original creator, provide a link to the license, and clearly indicate whether any changes were made, without implying endorsement by the original author. In addition, the ShareAlike condition means that if someone modifies or builds upon the work, the resulting material must be distributed under the same CC BY-SA 4.0 license, or a compatible one. Finally, users are not allowed to apply legal or technical restrictions, such as DRM, that would prevent others from exercising these same rights.

Code snippets are dedicated to the public domain and licensed under a CC0 1.0 Creative Commons Universal License. You may use, modify, distribute, and sell the code snippets for any purpose, without permission or attribution. The code snippets are provided “as is”, without warranty of any kind.

Additional content

Additional exercises: Replicability and reproducibility

Decide whether each scenario in the following slides is an example of **reproducibility** or **replicability**.

Scenario 1

A computational neuroscientist reruns a published fMRI analysis using the original dataset and Python scripts to verify the reported brain activation patterns.

Instructor Notes: Correct answer: Reproducibility

Scenario 2

An environmental scientist repeats a field experiment on soil nutrient levels using the same sampling protocol at a different site.

Instructor Notes: Correct answer: Replicability

Scenario 3

A linguist reanalyzes a corpus of historical texts using the same annotation guidelines and code to verify reported patterns of syntactic structures.

Instructor Notes: Correct answer: Reproducibility

Scenario 4

A psychology lab replicates a social behavior experiment using new participants from a different cultural background.

Instructor Notes: Correct answer: Replicability

What is (academic) "trust" in research?

- Trust is **essential for effective collaboration** among researchers (includes co-authorship, peer review, data sharing, replication, teaching, mentoring, etc.). We operate under the assumption of good faith.
- Scientists reading published research trust that the **work was conducted as described**, that **all relevant methodological details** are **disclosed**, and that the data have **not been fabricated or falsified**.

Resnik, D. B. (2011). Scientific research and the public trust. *Science and engineering ethics, 17*(3), 399-409.

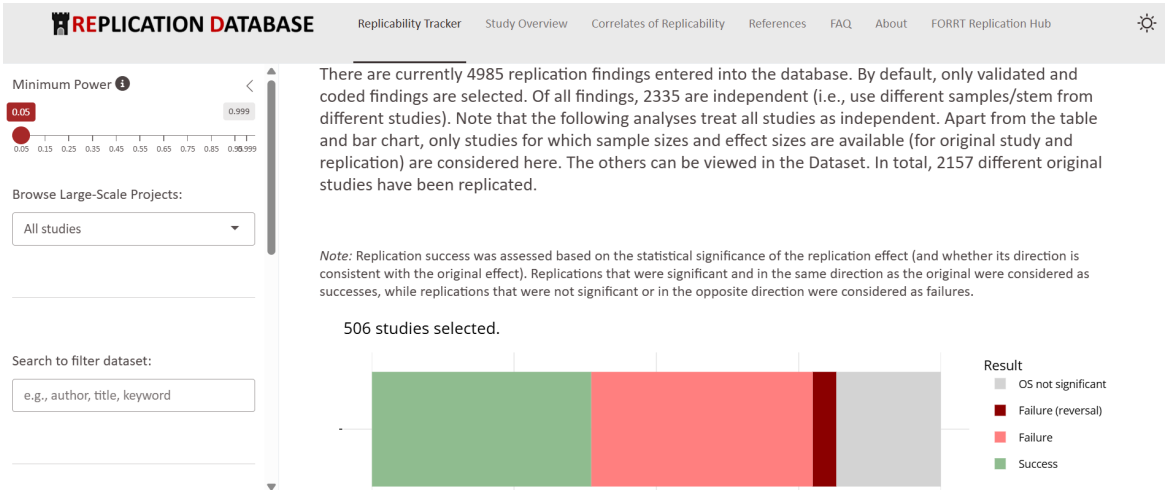
Presenter Notes: When discussing the importance of trust in research by society, Resink highlights that trust is absolutely essential for effective collaboration in science. Researchers rely on one another in so many ways, e.g., co-authoring papers, reviewing each other's work, sharing data, attempting replications, and even in day-to-day activities like teaching and mentoring. None of these collaborative practices would function smoothly without a basic level of confidence that everyone is acting with integrity.

Likewise, when scientists read published research, they trust that the study was carried out as the authors describe. They expect that all the important methodological details are reported transparently, so the work can be evaluated or repeated. And they assume that the data are genuine, or in other words that nothing has been fabricated, falsified, or manipulated; intentionally or unintentionally. Without this trust in the accuracy and honesty of the research record, the entire scientific research process becomes unstable and no longer credible.

Instructor Notes: Emphasis that trust in research operates at multiple levels and that it may not be a yes/no issue but more of a multidimensional concept along different continuum. Trust is not automatic, and can be gradually lost, as we have for example seen during the COVID-19 pandemic or the Tuskegee Syphilis Study in the United States. In the latter, researchers did not ask participants for consent in the study and withheld treatment for participants. The case went to court, and study participants received a court settlement in 1974.

Psychology: The Replication Database by FORRT

- Collects replication results across different psychological fields
- Provide detailed overview of the original findings and the replication outcomes



FORRT Replication Database (https://forrt-replications.shinyapps.io/fred_explorer/)

Why should we trust researchers?



Snake oil salesman



Researcher

Uses a method which produces anecdotal evidence and false positive results in the desired direction.	Uses a sound scientific method which, in the long run, arguably arrives at an increasingly correct account of the world.
Has a predetermined attitude about „what works“.	Changes his/her mind in the light of new evidence.
Hides data and methods; nobody can check or reconstruct what has been done.	Is transparent about the primary data, methods, and analyses.
Commercial interests/conflicts of interest.	No commercial interests / COIs, or discloses them.
„Believe me! I have the greatest snake oil.“	„Nullius in verba“ - motto of the Royal Society, world's oldest scientific society: „Don't trust my words“

Your current understanding of key terms (6 min)

Think - Pair - Share

> Note - this is not a test!
>
> These terms probably have not been covered in your courses yet. We just want to see what your current understanding of these terms is.