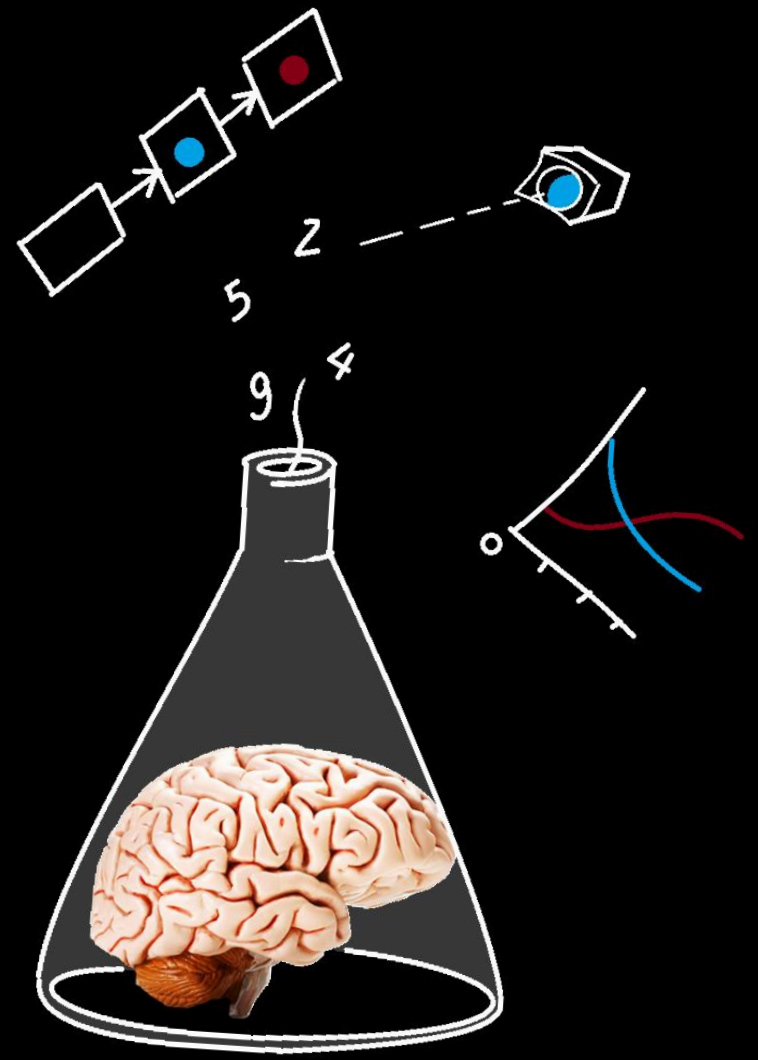
11-09

From theory to prediction to experiment

+

OpenSesame workshop

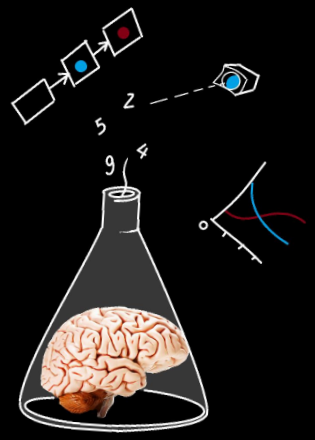


Why talk about theory?

Methods and techniques aren't our only instruments; theories are instruments too!

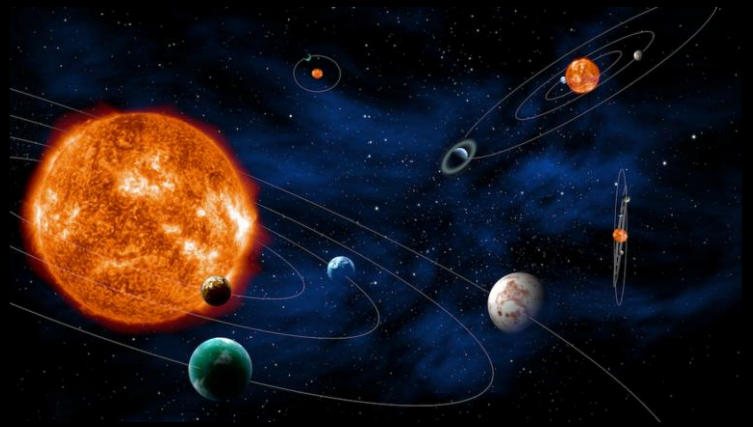
Methods and techniques dictate *how* we investigate things…

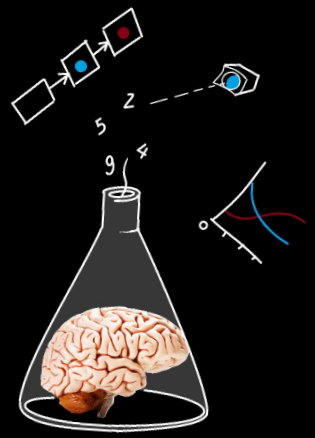
Theory dictates *what* we investigate and *why*



Do we always need theory?

No. Purely descriptive / observational research can be useful too.

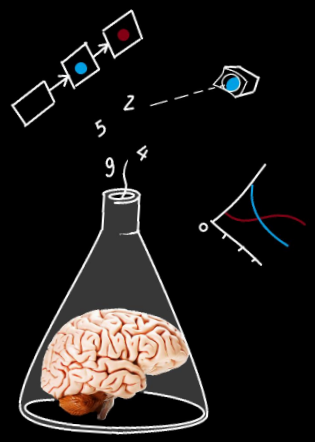




Do we always need theory?

Ultimately science is not just about knowing every thing, but about discovering the laws that connect all things

→ Not just *what*, but also *why*
Laws are captured in theory

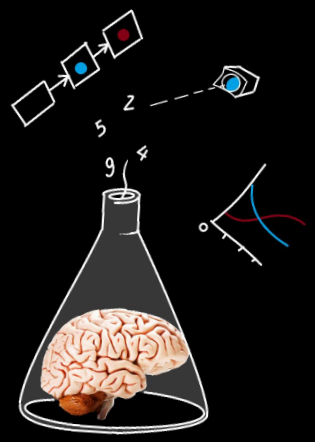


What is a good theory?

Explanatory scope

Parsimony

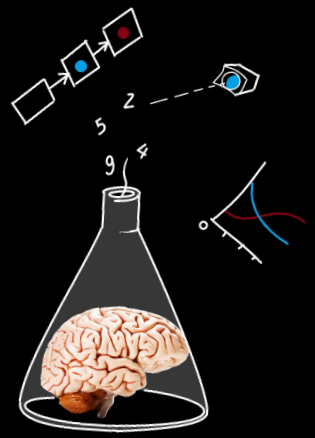
A theory must be *falsifiable*
If a theory cannot be tested,
we have zero knowledge about its plausibility



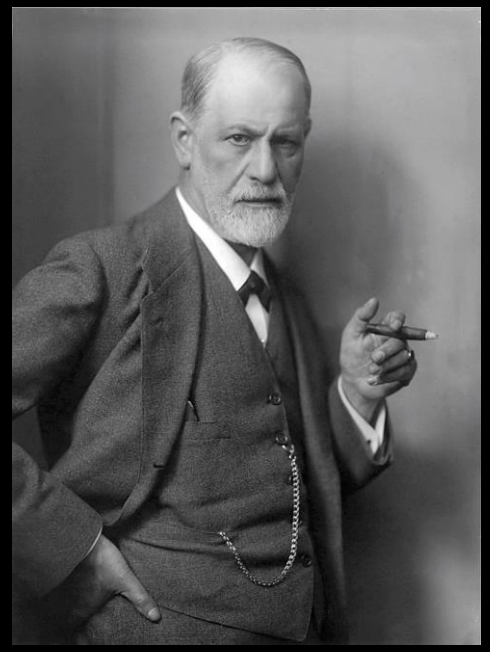
A theory must be *falsifiable*

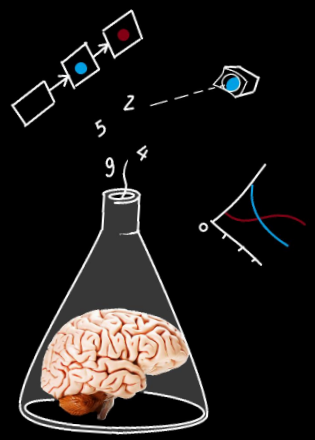
Example: serpent churches?

"These signs will accompany those who have believed: in My name they will cast out demons, they will speak with new tongues; they will pick up serpents, and if they drink any deadly poison, it will not hurt them; they will lay hands on the sick, and they will recover," (Mark 16:17-18).



Sigmund Freud: Psycho-analysis





Testing a theory in cognitive psychology

"A causes B"
"B relies on A"
"The influence of A on B relies on C"

The cognitive psychologist's challenge:
How can we manipulate A and measure B?

Testing a theory in cognitive psychology

"Expectations influence perception"

The cognitive psychologist's challenge:
How can we manipulate 'expectations' and measure 'perception'?

Testing a theory in cognitive psychology

"Expectations influence perception"

We need experimental conditions
e.g., one with 'high expectation', the other with 'low expectation'

We need to devise a behavioral measure of perception
e.g., an indication by a participant whether a visual object is blue or yellow

Testing a theory in cognitive psychology

We have to make everything quantifiable
→ we must be able to translate everything into numbers

'psychonomics'

The time it takes to respond in high vs. low expectation condition

The proportion of 'blue' responses in high vs. low expectation condition

Do we take other factors into account?

"But wait, older people are more sensitive to yellow than to blue"

"But wait, when the sky outside is blue, this may influence responses"

No problem: within-subjects design

Do we take other factors into account?

Within-subjects design: each subject is tested in all conditions
Between-subjects design: different subjects are tested in each condition

We typically always prefer a within-subjects design over a between-subjects design, as it allows us to ignore factors that potentially have an influence

The replication crisis in psychology

Psychonomic Bulletin & Review (2019) 26:1596–1618
https://doi.org/10.3758/s13423-019-01645-2

THEORETICAL REVIEW

Addressing the theory crisis in psychology

Klaus Oberauer [1] · Stephan Lewandowsky [2,3]

Published online: 12 September 2019

Abstract
A worrying number of psychological findings are not replicable. Diagnoses of the causes of this "replication crisis," and recommendations to address it, have nearly exclusively focused on methods of data collection, analysis, and reporting. We argue that a further cause of poor replicability is the often weak logical link between theories and their empirical tests. We propose a distinction between discovery-oriented and theory-testing research. In discovery-oriented research, theories do not strongly imply hypotheses by which they can be tested, but rather define a search space for the discovery of effects that would support them. Failures to find these effects do not question the theory. This endeavor necessarily engenders a high risk of Type I errors—that is, publication of findings that will not replicate. Theory-testing research, by contrast, relies on theories that strongly imply hypotheses, such that disconfirmation of the hypothesis provides evidence against the theory. Theory-testing research engenders a smaller risk of Type I errors. A strong link between theories and hypotheses is best achieved by formalizing theories as computational models. We critically revisit recommendations for addressing the "replication crisis," including the proposal to distinguish exploratory from confirmatory research, and the preregistration of hypotheses and analysis plans.

Keywords Replication · Scientific inference · Hypothesis testing · Computational modeling · Preregistration

Psychology has a problem. Over the past decade it has become clear that many findings, among them some deemed well established, are not replicable (Marsman et al., 2017; Science Collaboration, 2015). N... generation are sufficient to ... our ...

ANNUAL REVIEWS

Annual Review of Psychology

Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis

Patrick E. Shrout[1] and Joseph L. Rodgers[2]

[1]Department of Psychology, New York University, New York, New York 10003; email: pat.shrout@nyu.edu

[2]Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, Tennessee 37203; email: joseph.l.rodgers@vanderbilt.edu

Annu. Rev. Psychol. 2018. 69:487–510

Keywords

aps PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 632–638
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691612463078
http://pps.sagepub.com

SAGE

An Agenda for Purely Confirmatory Research

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit
University of Amsterdam, The Netherlands

Abstract
The veracity of substantive research claims hinges on the way experimental data are collected and analyzed. In this article, we discuss an uncomfortable fact that threatens the core of psychology's academic enterprise: almost without exception, psychologists do not commit themselves to a method of data analysis before they see the actual data. It then becomes tempting to fine tune the analysis to the data in order to obtain a desired result—a procedure that invalidates the interpretation of the common statistical tests. The extent of the fine tuning varies widely across experiments and experimenters but is almost impossible for reviewers and readers to gauge. To remedy the situation, we propose that researchers preregister their studies and indicate in advance the analyses they intend to conduct. Only these analyses deserve the label "confirmatory," and only for these analyses are the common statistical tests valid. Other analyses can be carried out but these should be labeled "exploratory." We illustrate our proposal with a confirmatory replication attempt of a study on extrasensory perception.

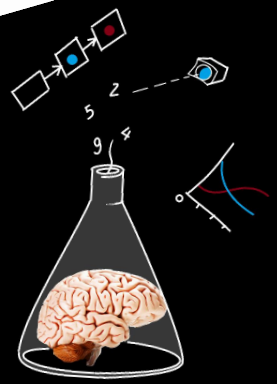
Keywords
confirmatory experiments, wonky statistics, ESP, Bayesian hypothesis test

You cannot find your starting hypothesis in your final results. It makes the stats go all wonky.

—Ben Goldacre (2009, p. 221)

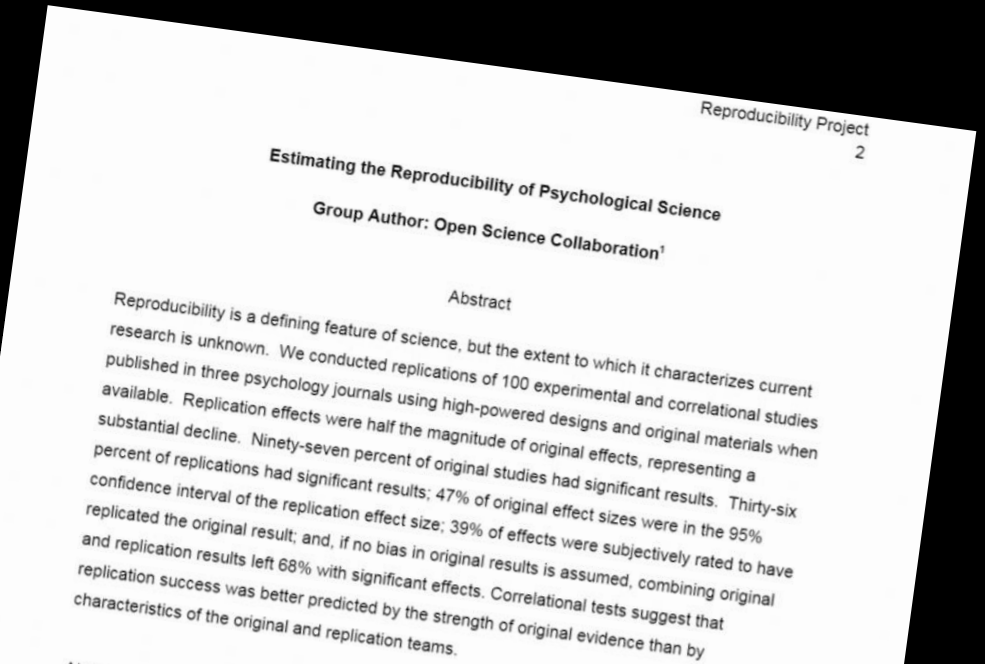
Psychology is a challenging discipline. Empirical data are noisy, formal theory is scarce, and the processes of interest (e.g., attention, jealousy, loss aversion) cannot be observed directly. Nevertheless, psychologists have managed to generate many key insights about human cognition and behavior. For instance, research has shown that people tend to seek confirmation rather than disconfirmation of their beliefs—a phenomenon known as *confirmation bias* (Nickerson, 1998). Confirmation bias operates in at least three ways. First, ambiguous information is readily interpreted to be consistent with one's prior beliefs; second, people tend to search for information that confirms rather than disconfirms their preferred hypothesis; third, people more easily remember information that supports their position. We also know that people fall prey to hindsight bias, the tendency to judge an event as more predictable after it has occurred (Roese & Vohs, 2012). In light of these and other biases[1] it would be naive to believe that, without special protective measures, the scientific research process is somehow exempt from the systematic imperfections of the human mind. For example, one indication that bias influences the research process is that researchers seek to confirm, not falsify, their main hypothesis (Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). The impact of bias is exacerbated in an environment that puts a premium on output quantity. When academic survival depends on how many papers one publishes, researchers are attracted to methods and procedures that maximize the probability of publication (Bakker, van Dijk, & Wicherts, 2012; John, Loewenstein, & Prelec, 2012; Neuroskeptic, 2012; Nosek, Spies, & Motyl, 2012). It should be noted that such behavior is ecologically rational in the sense that it maximizes the proximal goals of the researcher. However, when each researcher acts this way in an entirely understandable attempt at academic self-preservation, the cumulative effect on the field as a whole can be catastrophic. The primary concern is that many published results may simply be false, as they have been obtained partly by dubious or inappropriate methods of observation, analysis, and reporting (Jasny, Chin, Chong, & Vignieri, 2011; Sarewitz, 2012).

Corresponding Author:
Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychological Methods, Weesperplein 4, 1018 XA Amsterdam, The Netherlands
E-mail: EJ.Wagenmakers@gmail.com



The replication crisis in psychology

Approximately 40% of published studies could be replicated… might 60% of the science out there be erroneous?



Reproducibility Project
2

Estimating the Reproducibility of Psychological Science

Group Author: Open Science Collaboration[†]

Abstract

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had significant results. Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, combining original and replication results left 68% with significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[4,16], Brian A. Nosek[6,7,16*], Thomas Pfeiffer[8,16], Adam Altmejd[2], Nick Buttrick[7,9], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,9], Emma Heikensten[2], Lily Hummer[6], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

Being able to replicate scientific findings is crucial for scientific progress[1–5]. We replicate 21 systematically selected experimental studies in the social sciences published in *Nature* and *Science* between 2010 and 2015[6–26]. The replications follow analysis plans reviewed by the original authors and pre-registered prior to the replications. The replications are high powered, with sample sizes on average about five times higher than in the original study. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. Replicability varies between 12 (57%) and 14 (67%) studies for complementary replicability indicators. Consistent with these results, the estimated true-positive rate is 67% in a Bayesian analysis. The relative effect size of true positives is estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility. Furthermore, we find that peer beliefs of replicability are strongly related to replicability, suggesting that the research community could predict which results would replicate and that failures to replicate were not the result of chance alone.

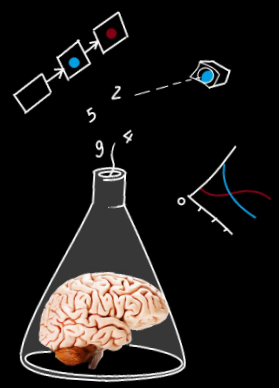
The replication crisis in psychology

Various causes

- publication bias

"file drawer problem"

- Bad practices – intentional and unintentional!
- Overall pressure to produce 'sexy' results



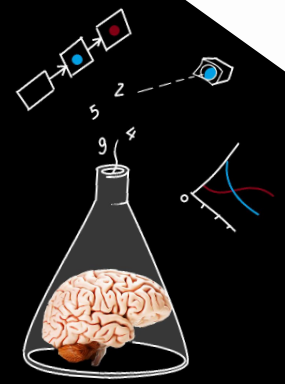
The replication crisis in psychology

How can we have more confidence in our re

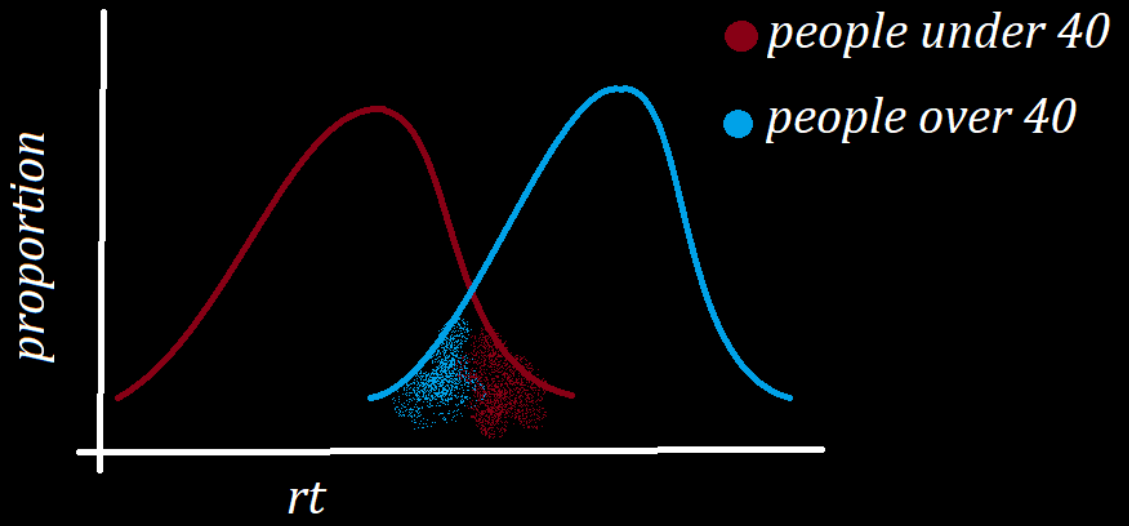
→ Have abundant statistical power

→ Replicate our own experiments

→ Engage in theory-driven research; preferably attempting to *falsify* a theory rather than to confirm it

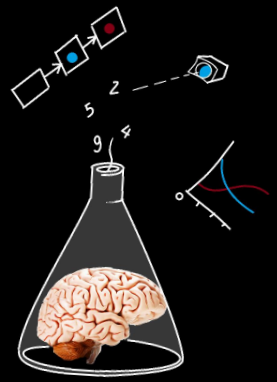


The replication crisis in psychology



Statistical power
→ the chance that an effect is established, given that the hypothesis is true

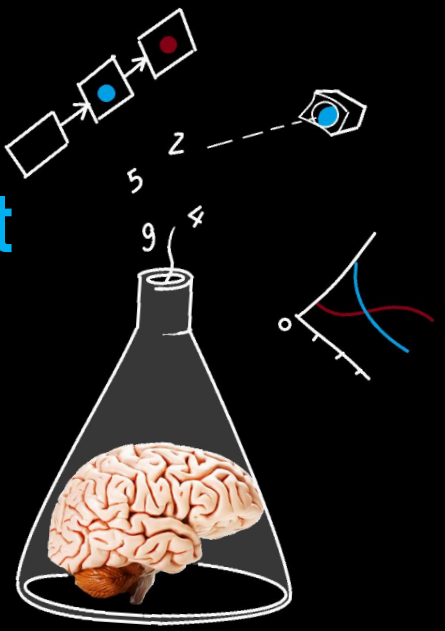
Brysbaert & Stevens (2018): 1,600 measurements per condition (e.g., 25 subjects, 64 trials per condition)



Testing theories without experiments

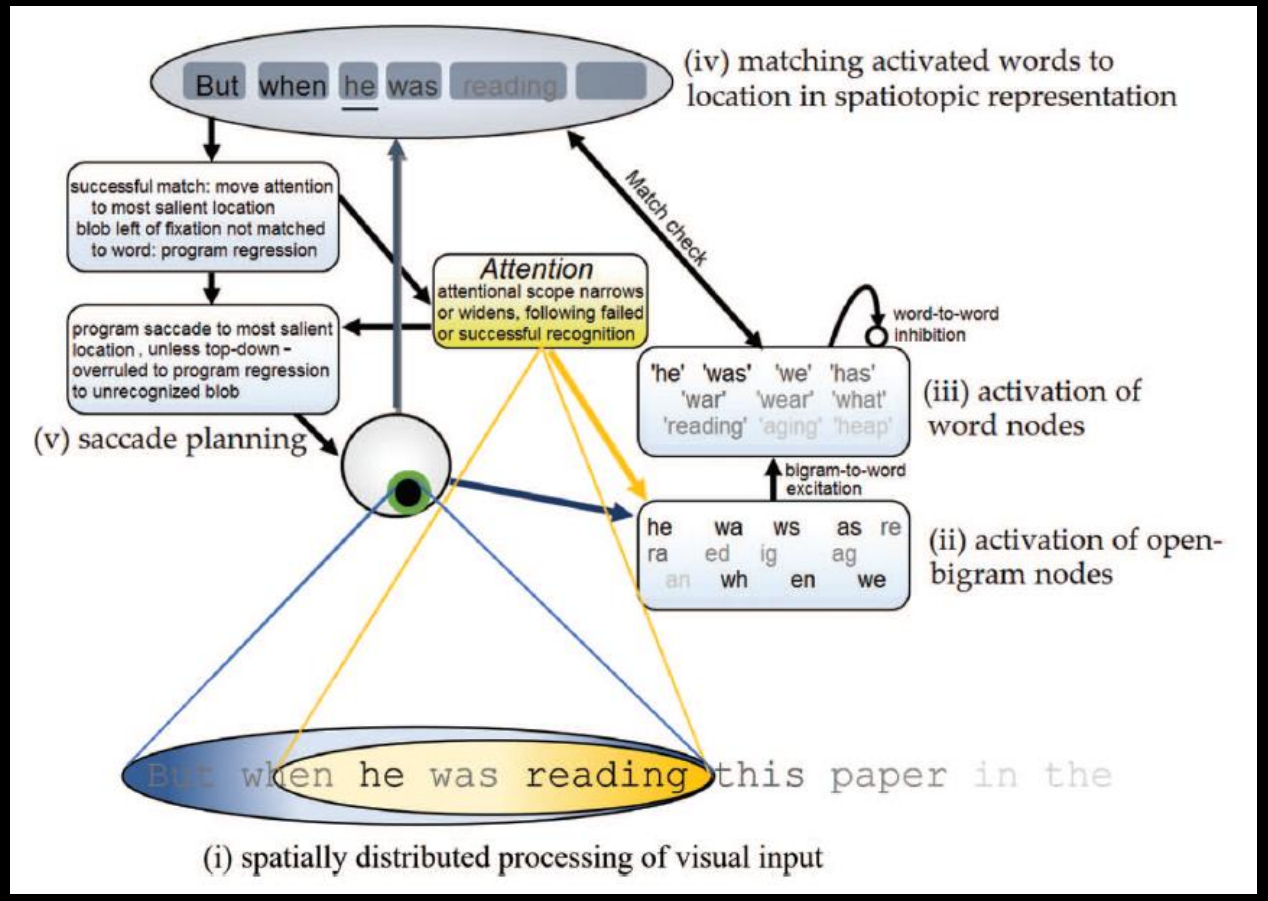
→ Computational models

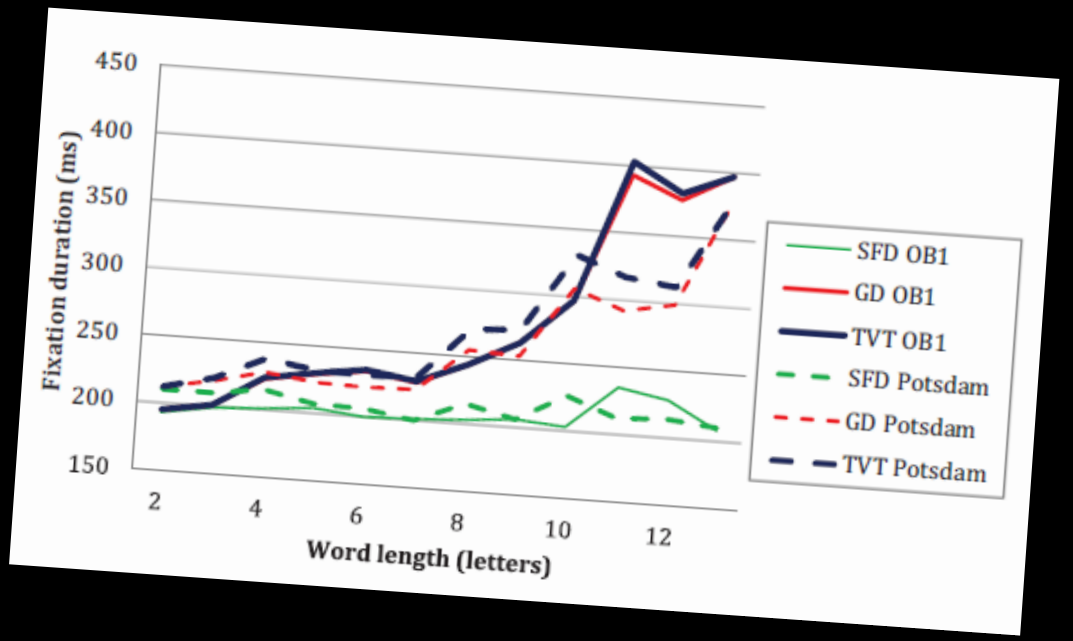
We can test the extent to which a computer model can accurately simulate behavior. The more accurately it simulates behavior, the more support we have for each of the model's assumptions.



Testing theories without experiments

→ Computational models





Summary

- Theories are instruments, just like methods and techniques. Methods: *how*. Theory: *what* & *why*
- Without theories we can still make valuable observations. But theories are the glue that should hold everything together.
- The 3 characteristics of a good theory: Explanatory scope, Parsimony, and Falsifiability
- Don't hold venomous snakes.
- H: A → B. Our mission: devise ways to manipulate A, and to measure *(quantify)* B.
- When possible, stick to a within-subjects design

Summary

- For a future without replication crisis: Have abundant statistical power, replicate our own experiments, aim to falsify theories
- Approx. 1,600 measurements per condition should do.
- Testing theory without experiment: computational model. Model outputs behavior, behavior is compared to human behavior; the more resemblance, the more support for model assumptions.