Carl von Ossietzky
Universität
Oldenburg

meta rep

Multiverse Analysis: Theory and Conceptualisation

Cassie Short, Daniel Kristanto, Micha Burkhardt, Andrea Hildebrandt

Pre-Conference-Workshop, PuG 2025

17.-18.06.2025

The schedule for the afternoon

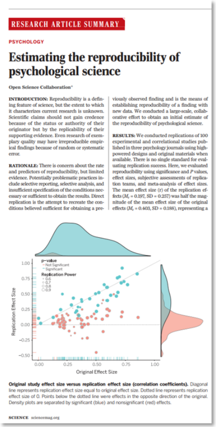
- 13:30-15:00: Multiverse Analysis Theory and Conceptualisation
- 15:00-15:30: Coffee break
- 15:30-17:00: Multiverse Analysis in Neuroimaging
- 17:00-18:00: Conceptualisation activity

Content

Why Multiverse Analysis

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

The replication problem



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

It can be proven that most claimed research findings are false.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let *R* be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. *R* is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that *c* relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1-\beta)R/(R - \beta R + \alpha)$. A research finding is thus

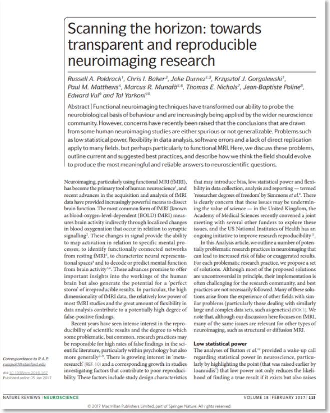
Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124



Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1,3], Krzysztof J. Gorgolewski[1], Paul M. Matthews[4], Marcus R. Munafò[5,6], Thomas E. Nichols[7], Jean-Baptiste Poline[8], Edward Vul[9] and Tal Yarkoni[10]

Abstract | Functional neuroimaging techniques have transformed our ability to probe the neurobiological basis of behaviour and are increasingly being applied by the wider neuroscience community. However, concerns have recently been raised that the conclusions that are drawn from some human neuroimaging studies are either spurious or not generalizable. Problems such as low statistical power, flexibility in data analysis, software errors and a lack of direct replication apply to many fields, but perhaps particularly to functional MRI. Here, we discuss these problems, outline current and suggested best practices, and describe how we think the field should evolve to produce the most meaningful and reliable answers to neuroscientific questions.

Neuroimaging, particularly using functional MRI (fMRI), has become the primary tool of human neuroscience[1], and recent advances in the acquisition and analysis of fMRI data have provided increasingly powerful means to dissect brain function. The most common form of fMRI (known as blood-oxygen-level-dependent (BOLD) fMRI) measures brain activity indirectly through localized changes in blood oxygenation that occur in relation to synaptic signalling[2]. These changes in signal provide the ability to map activation in relation to specific mental processes, to identify functionally connected networks from resting fMRI[3], to characterize neural representational spaces[4] and to decode or predict mental function from brain activity[5,6]. These advances promise to offer important insights into the workings of the human brain but also generate the potential for a 'perfect storm' of irreproducible results. In particular, the high dimensionality of fMRI data, the relatively low power of most fMRI studies and the great amount of flexibility in data analysis contribute to a potentially high degree of false-positive findings.

Recent years have seen intense interest in the reproducibility of scientific results and the degree to which some problematic, but common, research practices may be responsible for high rates of false findings in the scientific literature, particularly within psychology but also more generally[7–9]. There is growing interest in 'meta-research' (REF. 10) and a corresponding growth in studies investigating factors that contribute to poor reproducibility. These factors include study design characteristics that may introduce bias, low statistical power and flexibility in data collection, analysis and reporting — termed 'researcher degrees of freedom' by Simmons *et al.*[8]. There is clearly concern that these issues may be undermining the value of science — in the United Kingdom, the Academy of Medical Sciences recently convened a joint meeting with several other funders to explore these issues, and the US National Institutes of Health has an ongoing initiative to improve research reproducibility[11].

In this Analysis article, we outline a number of potentially problematic research practices in neuroimaging that can lead to increased risk of false or exaggerated results. For each problematic research practice, we propose a set of solutions. Although most of the proposed solutions are uncontroversial in principle, their implementation is often challenging for the research community, and best practices are not necessarily followed. Many of these solutions arise from the experience of other fields with similar problems (particularly those dealing with similarly large and complex data sets, such as genetics) (BOX 1). We note that, although our discussion here focuses on fMRI, many of the same issues are relevant for other types of neuroimaging, such as structural or diffusion MRI.

Low statistical power

The analyses of Button *et al.*[12] provided a wake-up call regarding statistical power in neuroscience, particularly by highlighting the point (that was raised earlier by Ioannidis[7]) that low power not only reduces the likelihood of finding a true result if it exists but also raises

Correspondence to R.A.P.
russpold@stanford.edu

doi:10.1038/nrn.2016.167
Published online 05 Jan 2017

Three main reasons for the replication problem

- Fragmented theoretical landscape and bold hypotheses
- Modest sample size resulting in low statistical power
- Undisclosed flexibility in data analyses

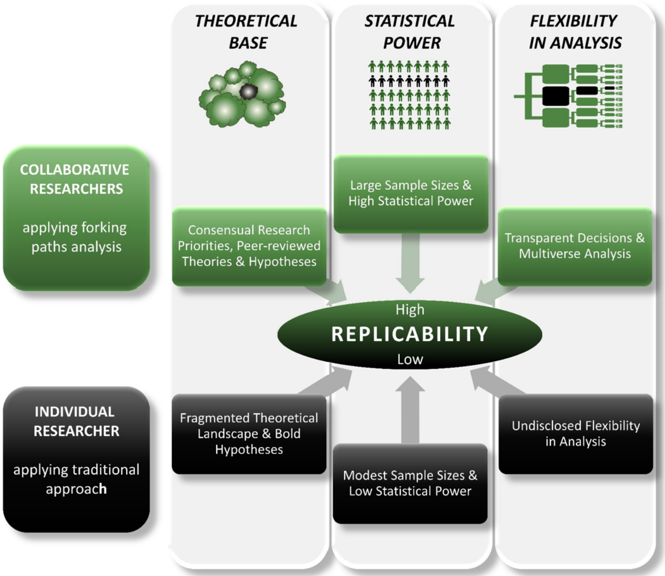
e.g., Paul, Short et al. (2022)

Three main reasons for the replication problem

- Fragmented theoretical landscape and bold hypotheses
- Modest sample size resulting in low statistical power
- Undisclosed flexibility in data analyses

e.g., Paul, Short et al. (2022)

Collaborative forking path analyses (cFPA, Wacker, 2017)



THEORETICAL BASE

STATISTICAL POWER

FLEXIBILITY IN ANALYSIS

COLLABORATIVE RESEARCHERS

applying forking paths analysis

Consensual Research Priorities, Peer-reviewed Theories & Hypotheses

Large Sample Sizes & High Statistical Power

Transparent Decisions & Multiverse Analysis

High

REPLICABILITY

Low

INDIVIDUAL RESEARCHER

applying traditional approach

Fragmented Theoretical Landscape & Bold Hypotheses

Modest Sample Sizes & Low Statistical Power

Undisclosed Flexibility in Analysis

Paul, Short et al. (2022)

Content

1. Theoretical input
    - The replication problem
    - The five sources of uncertainty in empirical research
    - Traditional empirical research approach
    - On the multiplicity of analysis strategies
    - Multiverse analysis
    - Principled multiverse analysis
    - Systematic multiverse analysis construction
    - Summary and conclusions

Introduction - The goal of quantitative empirical research

- The aim of most empirical research is to test the (causal) relationship between an outcome (Y) and explanatory variables (X_1, X_2,..., X_n)
- Estimate a function to describe bi- or multivariate relationships

Introduction - The goal of quantitative empirical research

- The aim of most empirical research is to test the (causal) relationship between an outcome (Y) and explanatory variables (X_1, X_2,..., X_n)
- Estimate a function to describe bi- or multivariate relationships
- In many cases in psychology we model multiple outcomes simultaneously, but for simplicity in this workshop we will only talk about models with a single outcome variable.

Introduction - The goal of quantitative empirical research

- The aim of most empirical research is to test the (causal) relationship between an outcome (Y) and explanatory variables (X_1, X_2,..., X_n)
- Estimate a function to describe bi- or multivariate relationships

Mathematical function for relating an outcome variable to explanatory variable(s) and confounder(s)

$$Y = f(X_{1-n}, Z_{1-n}) \tag{1}$$

Sources of uncertainty in empirical research - Introduction

- The aim of most empirical research is to test the (causal) relationship between an outcome (Y) and explanatory variables (X_1, X_2,..., X_n)
- Estimate a function to describe bi- or multivariate relationships

Mathematical function for relating an outcome variable to explanatory variable(s) and confounder(s)

$$Y = f(X_{1-n}, Z_{1-n}) \tag{2}$$

- Let's discuss an example

Example

- Personality psychology
  - Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N) ?

Example

- Personality psychology
  - Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N) ?

Mathematical function

$$ES = f(O, C, E, A, N) \tag{3}$$

Example

- Personality psychology
  - Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N) ?

- The variables ES, O, C, E, A, N have to be operationalised
  1. Is ES a binary success variable or a quantitative variable?
  2. Are the variables O, C, E, A, N observed or latent?
- Are there confounders to be considered? For example motivation.
- The functional form of the association and the statistical method to estimate the model parameters need be specified, and these decisions will depend on the type of variables, on sample size, on sample characteristics, etc.

The five sources of uncertainty in empirical research - see Hoffmann et al. (2021)

- Measurement uncertainty arises from randomness in the operationalisation or measurement of input and output variables
- Data processing uncertainty arises from the multiplicity of choices in selecting the data to be analysed and in defining, cleaning and transforming the input and output variables
- Model uncertainty arises from the multiplicity of choices in the specification of the model structure to describe the phenomenon of interest
- Method uncertainty arises from the multiplicity of potential decisions in the choice of statistical methods to estimate model parameters
- Sampling uncertainty arises from randomness in the sampling from the population of interest

The five sources of uncertainty in empirical research - see Hoffmann et al. (2021)

- Measurement uncertainty arises from randomness in the operationalisation or measurement of input and output variables
- Data processing uncertainty arises from the multiplicity of choices in selecting the data to be analysed and in defining, cleaning and transforming the input and output variables
- Model uncertainty arises from the multiplicity of choices in the specification of the model structure to describe the phenomenon of interest
- Method uncertainty arises from the multiplicity of potential decisions in the choice of statistical methods to estimate model parameters
- Sampling uncertainty arises from randomness in the sampling from the population of interest

Note that in frequentist statistics, sampling uncertainty is the only one routinely taken into account, e.g., in null hypothesis significance testing.

This workshop

...is concerned with

- measurement uncertainty
- data processing uncertainty
- model uncertainty
- method uncertainty

Back to our example

- Personality psychology
  - Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N) ?

If ES is taken as quantitative and the personality traits are operationalised and quantified as sum scores on the NEO-FFI questionnaire, the sample is representative of the population and the predictors are all normally distributed...

- Which model could be estimated to address the above question?
- How could the model parameters be estimated?

Back to our example

- Personality psychology
  - Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N) ?

If ES is taken as quantitative and the personality traits are operationalised and quantified as sum scores on the NEO-FFI questionnaire, the sample is representative of the population and the predictors are all normally distributed:

Mathematical function (option 1)

$$ES = b_0 + b_1 \cdot O + b_2 \cdot C + b_3 \cdot E + b_4 \cdot A + b_5 \cdot N + \epsilon \tag{4}$$

The model parameters can be estimated with the Ordinary Least Squares (OLS) method or the Maximum Likelihood Estimator (MLE).

Back to our example

- Personality psychology
  - Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N) ?

If ES is taken as binary and the personality traits are operationalised and quantified as sum scores on the NEO-FFI questionnaire, the sample is representative of the population and the predictors are all normally distributed:

Mathematical function (option 2)

$$P(ES = 1|O, C, E, A, N) = \frac{e^{b_0 + b_1 \cdot O + b_2 \cdot C + b_3 \cdot E + b_4 \cdot A + b_5 \cdot N}}{1 + e^{b_0 + b_1 \cdot O + b_2 \cdot C + b_3 \cdot E + b_4 \cdot A + b_5 \cdot N}} \tag{5}$$

The model parameters can be estimated with the Maximum Likelihood Estimator (MLE) or alternatives.

Content

Theoretical input

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

Traditional empirical research approach

Researcher A

1. Research question: Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N)?
2. Measurement: Use years of education as outcome and five NEO-FFI scale sum scores as predictors
3. Data processing: Exclude observations with missing data, exclude observations with $M \pm 3SD$ identified as outliers
4. Model: Apply an additive multiple regression model with the link function for the Gaussian distribution
5. Estimation method: Estimate the parameters by using the OLS method
6. Sampling: Null hypothesis statistical significance testing at $\alpha = 0.05$

Traditional empirical research approach

Researcher B

1. Research question: Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N)?
2. Measurement: Use university degree (yes vs. no) as outcome and five NEO-FFI principal component scores as predictors
3. Data processing: Use mean imputation replace missing data on the predictors, exclude observations with $M \pm 2.5SD$ identified as outliers
4. Model: Apply a logistic regression model with the link function for the Binomial distribution
5. Estimation method: Estimate the parameters by using the MLE
6. Sampling: Null hypothesis statistical significance testing at $\alpha = 0.01$

Content

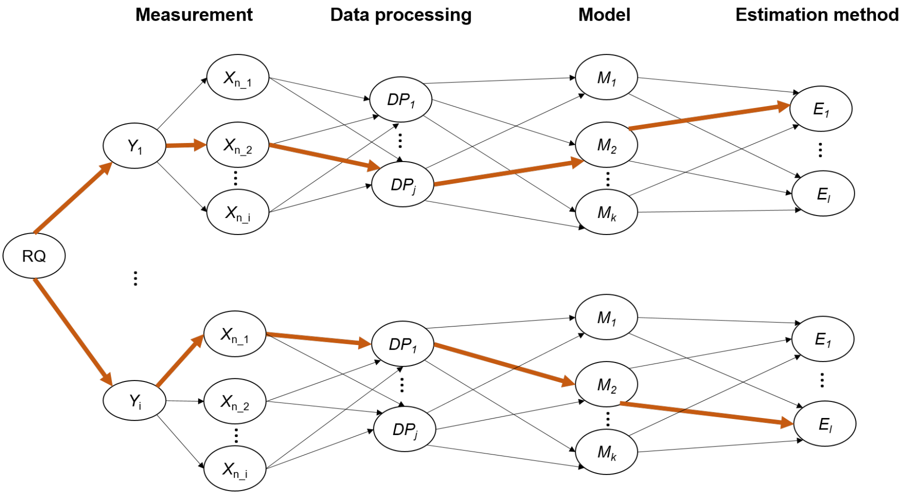
Theoretical input

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

On the multiplicity of analysis strategies (Gelman & Loken, 2014)

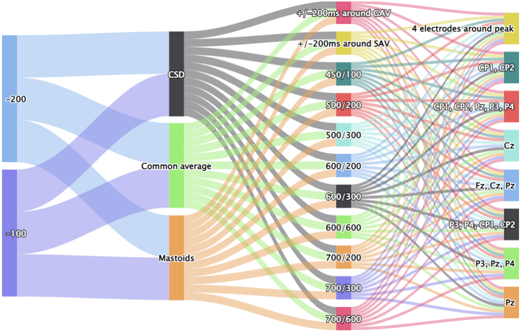
- The combination of many possible choices throughout the research process results in a theoretical multiverse of statistical outcomes, known as the *Garden of Forking Paths*.

The multiplicity of analysis strategies



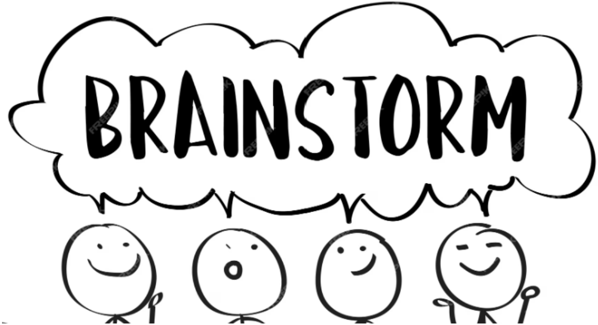
Sankey Diagram

- The combination of many possible choices throughout the research process results in a theoretical multiverse of statistical outcomes, known as the *Garden of Forking Paths*.



Magnitude of the multiverse = Cartesian product of all methodological decisions: $2 \cdot 3 \cdot 11 \cdot 8 = 528$ forking paths.

How to address the multiplicity of analyses strategies?



How to address the multiplicity of analyses strategies? - Hoffmann et al., 2021

- Reduce uncertainty
  - integrate existing knowledge
  - improve measurements
  - formulate more precise theories
  - increase sample size

How to address the multiplicity of analyses strategies? - Hoffmann et al., 2021

1. Accept uncertainty
   - aim for multiple lines of evidence
   - conduct replication studies
   - acknowledge constraints on generalizability
   - conduct meta-analysis

How to address the multiplicity of analyses strategies? - Hoffmann et al., 2021

1. Integrate uncertainty
    - Bayesian model averaging
    - Bayesian deep learning
    - Probabilistic sensitivity analysis

How to address the multiplicity of analyses strategies? - Hoffmann et al., 2021

- Report uncertainty
  - sensitivity analysis
  - robustness analysis
  - multiverse analysis / vibration of effects / specification curves

How to address the multiplicity of analysis strategies? - Our focus

1. Reduce uncertainty
2. Accept uncertainty
3. Integrate uncertainty
4. Report uncertainty

Content

Theoretical input

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

Multiverse analysis

Alternative terms for same or very similar approaches: "Vibration of effects" in epidemiology, "specification curve analysis" in psychology, "measure of robustness to misspecification" in economics, "multimodel analysis" and "computational robustness analysis" in sociology

- Primary goal of multiverse analysis is to enhance research transparency and uncover various sources of uncertainties

Multiverse analysis

- Primary goal of multiverse analysis is to enhance research transparency and uncover various sources of uncertainties
- Multiverse analysis investigates the effect of arbitrary decisions at the level of data processing and statistical modeling given the raw data file, research questions, and hypotheses

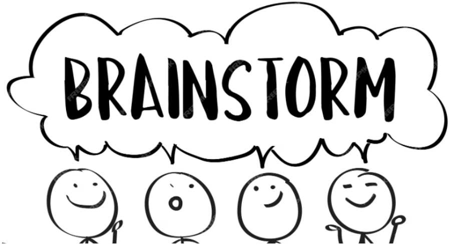
Multiverse analysis

- Primary goal of multiverse analysis is to enhance research transparency and uncover various sources of uncertainties
- Multiverse analysis investigates the effect of arbitrary decisions at the level of data processing and statistical modeling given the raw data file, research questions, and hypotheses
- Steps
  1. Identify decision nodes and arbitrary alternatives concerning preparation of the data and the statistical approach to construct all reasonable combinations of decisions
  2. Perform statistical analysis across all combinations specified in the previous step to obtain a set of results for each combination
  3. Examine variability in results by graphical representations
  4. Potentially: Conduct joint inference across the multiverse of findings

Back to our example and identify decision nodes

Research question: Is there an association between educational success (ES) and the big five personality traits (O, C, E, A, N)?

- In two groups



Content

Theoretical input

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

Principled multiverse analysis

- Garbage in - garbage out
- Del Giudice & Gangestad (2021) - "The multiverse is a dangerous place"

Principled multiverse analysis

- Del Giudice & Gangestad (2021) - "The multiverse is a dangerous place"
  - "In principle, multiverse-style analyses can be highly instructive. At the same time, analyses that explore multiverse spaces that are not homogeneous can produce misleading results and interpretations, lead scholars to dismiss the robustness of theoretically important findings that do exist, and discourage them from following fruitful avenues of research. This can hinder scientific progress just as much as the proliferation of false, unreplicable findings does."
  - "The main danger of multiverse-style methods lies in their potential for combinatorial explosion. Just a few decisions incorrectly treated as arbitrary can quickly explode the size of the multiverse, drowning reasonable effect estimates in a sea of unjustified alternatives."

Principled multiverse analysis

- Del Giudice & Gangestad (2021) - "The multiverse is a dangerous place"
  - "In principle, multiverse-style analyses can be highly instructive. At the same time, analyses that explore multiverse spaces that are not homogeneous can produce misleading results and interpretations, lead scholars to dismiss the robustness of theoretically important findings that do exist, and discourage them from following fruitful avenues of research. This can hinder scientific progress just as much as the proliferation of false, unreplicable findings does."
  - "The main danger of multiverse-style methods lies in their potential for combinatorial explosion. Just a few decisions incorrectly treated as arbitrary can quickly explode the size of the multiverse, drowning reasonable effect estimates in a sea of unjustified alternatives."

A principled multiverse analysis is necessary!

Principled multiverse analysis (Del Giudice & Gangestad, 2021)

- Multiverse analysis in a principled way is based on the a priori assessment of the equivalence of alternatives at each decision node

Principled multiverse analysis (Del Giudice & Gangestad, 2021)

- Multiverse analysis in a principled way is based on the a priori assessment of the equivalence of alternatives at each decision node
- Type E decisions - principled equivalence: Specifications are equivalent and effectively arbitrary, e.g., alternatives have comparable validity, examine the same effect, or estimate the effect with comparable precision
- Type N decisions - principled equivalence: Specification are nonequivalent, i.e., some of the alternatives are more justified than others as a means of estimating the effect of interest

Principled multiverse analysis (Del Giudice & Gangestad, 2021)

- Multiverse analysis in a principled way is based on the a priori assessment of the equivalence of alternatives at each decision node
- Type E decisions - principled equivalence: Specifications are equivalent and effectively arbitrary, e.g., alternatives have comparable validity, examine the same effect, or estimate the effect with comparable precision
- Type N decisions - principled equivalence: Specification are nonequivalent, i.e., some of the alternatives are more justified than others as a means of estimating the effect of interest
- Type E decisions should be selected for multiverse analysis exploring robustness, whereas Type N decisions should not be selected for multiverse analysis

Equivalence assessment of alternatives

- The assessment of the equivalence of alternatives at each decision node can be based on three kinds of nonequivalence

Equivalence assessment of alternatives

- The assessment of the equivalence of alternatives at each decision node can be based on three kinds of nonequivalence
    1. Measurement Nonequivalence: Alternative measurement choices yield systematic differences in validity and reliability
        - ★ Example?

Equivalence assessment of alternatives

- The assessment of the equivalence of alternatives at each decision node can be based on three kinds of nonequivalence
  1. Measurement Nonequivalence: Alternative measurement choices yield systematic differences in validity and reliability
  2. Effect Nonequivalence: Alternative specification investigates a different effect not in line with the effect of interest
     - ★ Example?

Equivalence assessment of alternatives

- The assessment of the equivalence of alternatives at each decision node can be based on three kinds of nonequivalence
  1. Measurement Nonequivalence: Alternative measurement choices yield systematic differences in validity and reliability
  2. Effect Nonequivalence: Alternative specification investigates a different effect not in line with the effect of interest
  3. Power/Precision Nonequivalence: Alternative specification results in a lower precision of estimating an effect and lower statistical power to detect an effect
     - ★ Example?

Content

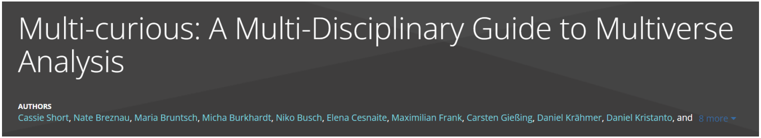
Theoretical input

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

Systematic multiverse analysis construction

- **Community guidelines** for transparent, comprehensive and systematic multiverse analysis construction
  - interdisciplinary guidance on key procedural considerations
- **Systematic Multiverse Analysis Registration Tool (SMART)**
  - increased transparency, systematicity and reproducibility
  - reduced uncertainty and potential for QRPs

Community Guidelines

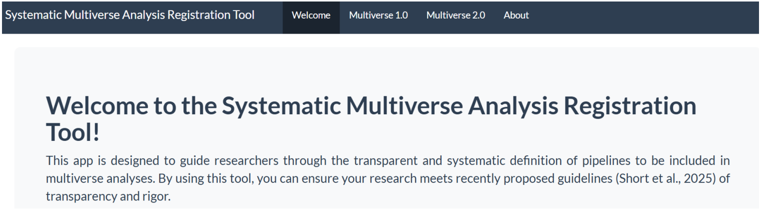


Multi-curious: A Multi-Disciplinary Guide to Multiverse Analysis

AUTHORS

Cassie Short, Nate Breznau, Maria Bruntsch, Micha Burkhardt, Niko Busch, Elena Cesnaite, Maximilian Frank, Carsten Gießing, Daniel Krähmer, Daniel Kristanto, and 8 more

SMART App



Welcome to the Systematic Multiverse Analysis Registration Tool!

This app is designed to guide researchers through the transparent and systematic definition of pipelines to be included in multiverse analyses. By using this tool, you can ensure your research meets recently proposed guidelines (Short et al., 2025) of transparency and rigor.

Content

Theoretical input

- The replication problem
- The five sources of uncertainty in empirical research
- Traditional empirical research approach
- On the multiplicity of analysis strategies
- Multiverse analysis
- Principled multiverse analysis
- Systematic multiverse analysis construction
- Summary and conclusions

Summary

- Multiverse analysis is an open science practice for dealing with uncertainties in empirical research
- There are multiple sources of uncertainty (measurement, data processing, model, method, and sampling), but only sampling uncertainty is routinely taken into account
- Multiplicity of analysis strategies resulting in a multiverse of results can be explored using multiverse analysis, where decision nodes and arbitrary alternatives are identified
- Importantly, a priori assessment of the equivalence of alternatives at each decision node is needed to only include Type E (equivalence) decisions in the multiverse analysis
- Note that multiverse analysis aims at assessing robustness to arbitrary changes in data analysis, but is not intended to assess substantive robustness

Uncertainty in research has in most cases an epistemic source, resulting from a lack of knowledge (Hoffmann et al., 2021). Multiverse analysis helps to increase knowledge and reduce uncertainty in the longer term.

https://news.uchicago.edu/

The schedule for the afternoon

- 13:30-15:00: Multiverse Analysis Theory and Conceptualisation
- 15:00-15:30: Coffee break
- 15:30-17:00: Multiverse Analysis in Neuroimaging
- 17:00-18:00: Conceptualisation activity

The schedule for the afternoon

- 13:30-15:00: Multiverse Analysis Theory and Conceptualisation
- 15:00-15:30: Coffee break
- 15:30-17:00: Multiverse Analysis in Neuroimaging
- 17:00-18:00: Conceptualisation activity

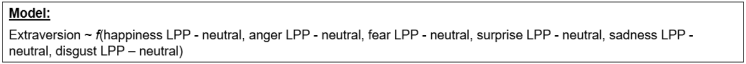
Conceptualisation Activity

- EEG Multiverse Analysis
  - Construct a multiverse analysis for one of your EEG projects, or
  - Construct a multiverse analysis for our example (next slides)
- fMRI Multiverse Analysis
  - Construct a multiverse analysis for one of your fMRI projects, or
  - Construct a multiverse analysis for our example (next slides)

We will discuss your garden of forking paths together.

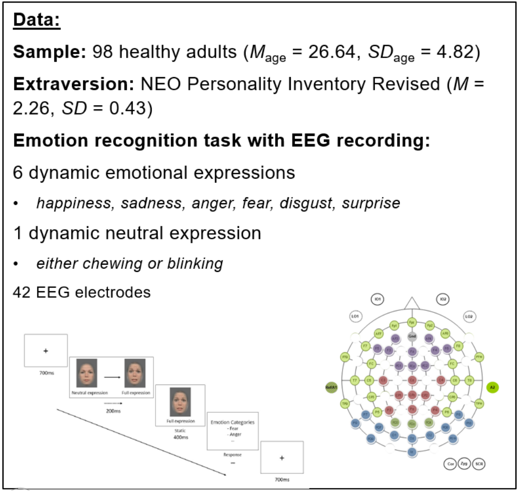
Conceptualisation Activity

EEG multiverse anlaysis example



Model:

Extraversion ~ *f*(happiness LPP - neutral, anger LPP - neutral, fear LPP - neutral, surprise LPP - neutral, sadness LPP - neutral, disgust LPP – neutral)



Data:

Sample: 98 healthy adults (M_{age} = 26.64, SD_{age} = 4.82)

Extraversion: NEO Personality Inventory Revised (M = 2.26, SD = 0.43)

Emotion recognition task with EEG recording:

6 dynamic emotional expressions

- *happiness, sadness, anger, fear, disgust, surprise*

1 dynamic neutral expression

- *either chewing or blinking*

42 EEG electrodes

+
700ms
Neutral expression
Full expression
200ms
Full expression
Static
400ms
Emotion Categories
- Fear
- Anger
...
Response
+
700ms
IO1
IO2
LO1
LO2
Fpz
Fp1
Fp2
AF7
AF8
Gnd
AF3
AF4
F7
F5
F3
Fz
F4
F6
F8
FT9
FT10
FC
FC1
FCz
FC2
FC
Ref/A1
A2
T7
C5
C3
Cz
C4
C6
T8
CP5
CP1
CPz
CP2
CP6
TP9
TP10
P7
P5
P3
Pz
P4
P6
P8
P9
P10
PO7
PO3
POz
PO4
PO8
PO9
PO10
O1
Oz
O2
Iz
Cor
Zyg
SCR

References

Botvinik-Nezer et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature, 582(7810)*:84-88

Del Giudice, M., Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4*.

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist, 102*, 460-465.

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science, 8*, 4.

Paul, K., Short, C. A., Beauducel, A., Per Carsten H., Härpfer, K., Hennig, J., Hewig, J., Hildebrandt, A., Kührt, C. Mueller, E.M., Osinsky, R. Porth, E., Riesel, A., Rodrigues, J., Scheffel, C., Stahl, J., Strobel, A., & Wacker, J. (2022). The Methodology and Dataset of the CoScience EEG-Personality Project – A Large-Scale, Multi-Laboratory Project Grounded in Cooperative Forking Paths Analysis. *Personality Science, 3*, e7177.

Short, C. A., Breznau, N., Bruntsch, M., Burkhardt, M., Busch, N., Cesnaite, E., Frank, M., Gießing, C., Krähmer, D., Kristanto, D., Lonsdorf, T., Neuendorf, C., Nguyen, H., Rausch, M., Schmalz, X., Schneck, A., Tabakci, C., Hildebrandt, A. (2025). Multi-curious: A Multi-Disciplinary Guide to Multiverse Analysis. *MetaArXiV*.

Short, C. A., Inceler, Y. C., Frank, M., Hildebrandt, A. (2025). The Systematic Multiverse Analysis Registration Tool (SMART) for Defining Multiverse Analyses. *MetaArXiV*.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behavior, 4*, 1208-1214.

Wacker, J. (2017). Increasing the reproducibility of science through close cooperation and forking path analysis. *Frontiers in Psychology, 8*, 1332.

Don’t get lost in the Garden of Forking Paths

Think carefully about the analytical choices you can make!

Thank you for attending this workshop!

andrea.hildebrant@uol.de
cassie.short@uol.de
daniel.kristanto@uol.de
micha.burkhardt@uol.de

Day 2: EEG Multiverse Practical

Model:

Extraversion ~ *f*(happiness LPP - neutral, anger LPP - neutral, fear LPP - neutral, surprise LPP - neutral, sadness LPP - neutral, disgust LPP – neutral)

Data:

Sample: 98 healthy adults (M_{age} = 26.64, SD_{age} = 4.82)

Extraversion: NEO Personality Inventory Revised (M = 2.26, SD = 0.43)

Emotion recognition task with EEG recording:

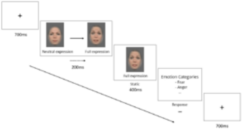
6 dynamic emotional expressions

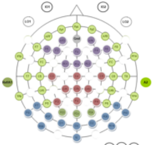
- *happiness, sadness, anger, fear, disgust, surprise*

1 dynamic neutral expression

- *either chewing or blinking*

42 EEG electrodes





Multiverse Use Case:

Baseline correction (2)

- -100m
- -200ms

Reference (2)

- Common average reference
- Linked mastoids

Time Window for LPP quantification (4)

- 400 – 600ms
- 500 – 700ms
- 450 – 750ms
- +/- 200 ms around subject average peak

Electrode cluster for LPP quantification (4)

- P3, P4, CP1, CP2
- P3, Pz, P4
- CP1, CP2
- Pz

Cartesian product: 2 * 2 * 4 * 4 = **64 pipelines**