# Beyond *p*-values:

# Rethinking Statistical Frameworks for Addressing the Replication Crisis [1]

Fernando Marmolejo-Ramos [#*]; Jose D. Perezgonzalez [$]; Raydonal Ospina [^]; Freddy Hernandez-Barajas [+]; Mauricio Castillo [//]; Rafael Izbicki [ᚠ]; Rafael B. Stern [■]; & Julian Tejada [∂]

[#] College of Education, Psychology, and Social Work; Flinders University, Adelaide 5042, SA, Australia. Email: fernando.marmolejoramos@flinders.edu.au *: corresponding author.

[$] Massey Business School, Massey University, Palmerston North 4442, New Zealand. Email: j.d.perezgonzalez@massey.ac.nz

[^] Departamento de Estatística, LInCa, Universidade Federal da Bahia, Cidade Universitária, Salvador/BA, 40170–110, Brazil. Email: raydonalmartinez@ufba.br

[+] Departamento de Estadística; Universidad Nacional de Colombia, sede Medellín; Colombia. Email: fhernanb@unal.edu.co

[//] Center for Basic Research in Psychology (CIBPsi), Faculty of Psychology, Universidad de la República; Montevideo 11200, Uruguay. Email: castillomega@gmail.com

[ᚠ] Department of Statistics, Federal University of São Carlos (UFSCar); Brazil. Email: rafaelizbicki@gmail.com

[■] Department of Statistics, University of São Paulo; Brazil. Email: rbstern@gmail.com

[∂] Department of Psychology, Federal University of Sergipe, Brazil. Email: jtejada@academico.ufs.br

[1] Contributed chapter for handbook "Research Handbook on the Replication Crisis" by David Trafimow (to appear in 2025 or thereabouts).

- *CRediT author statement = FM-R*: Conceptualization, Investigation, Methodology, Visualization, Writing- Original draft, Writing- Reviewing and Editing, Supervision. *J.D.P, R.O., F.H-B., M. C., R. I., R.S.,* and *J.T.:* Visualization, Writing- Reviewing and Editing, Investigation, Methodology.
- *Conflict of Interests* = The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- *Data availability statement* = all data and R codes are available within this manuscript
- *Acknowledgements* = none to disclose

**Abstract**

The replication crisis across scientific disciplines has prompted critical examination of statistical practices underpinning empirical research. This chapter analyzes how significance testing contributes to replication challenges through the lens of the TASI (Theoretical, Auxiliary, Statistical, Inferential) model. We examine three distinct contexts—inferential knowledge, region of acceptance testing, and statistical learning—to identify limitations in conventional significance testing and propose alternative frameworks. We argue that most problems with significance testing stem from researchers' tendencies to confirm rather than falsify hypotheses, regardless of statistical approach. We introduce REACT (Region of Acceptance Testing) as a stronger alternative to *p*-values, offering a structured decision-making process that integrates effect sizes and confidence intervals while explicitly recognizing when data is insufficient for definitive conclusions. Additionally, we propose Generalized Additive Models for Location, Scale, and Shape (GAMLSS) as a comprehensive statistical learning framework that transcends traditional hypothesis testing, focusing instead on descriptive, explanatory, and predictive modeling. The chapter ends with a concise exploration of the connection between data science and artificial intelligence.

*Keywords* = TASI model; REACT methodology; hypothesis testing; replication crisis; statistical learning.

## Introduction

The replication crisis in psychology describes the widespread inability to reproduce many published research findings in subsequent studies (Open Science Collaboration, 2015). This issue is not confined to psychology—it has emerged across multiple disciplines. For instance, biomedical research has faced similar challenges, with a substantial proportion of studies proving irreproducible (Baker, 2016). Likewise, replication failures have been documented in economics (Hensel, 2021) and neuroscience (Barch & Yarkoni, 2013), where several influential findings have not held up under repeated testing.

The replication crisis has been attributed to various factors, such as publication bias, small sample sizes, WEIRD samples, P-Hacking, HARCKing and poor statistical practices (Ioannidis, 2005; Nosek et al., 2012; Open science collaboration, 2015). As a result, many scientists across various fields have been working to address the replication crisis by implementing more rigorous and transparent research practices, such as pre-registration of studies, open data and materials, and replication studies. This chapter examines how significance testing contributes to the replication crisis, using the TASI model (Trafimow, 2019) as an analytical framework. The TASI model breaks down scientific research assumptions into four categories: Theoretical (from underlying theories), Auxiliary (connecting theory to observations), Statistical (about data analysis), and Inferential (about sampling and populations). We examine hypothesis testing through this model in three contexts: inferential knowledge, region of acceptance testing, and statistical learning. The chapter begins with an overview of significance testing, questionable research practices, replicability issues, generalizability, and the role of artificial intelligence (AI) in open science. It concludes with practical recommendations based on this analysis.

### *A recap on what significance testing is*

Significance testing is a statistical method used to determine whether a property assumed to be present in a population is compatible with what is observed in a sample (Kaplan, 2004). In other words, a significance test is conducted to determine whether the data observed in the sample is consistent with the hypothesis that is being made about a population. Significance tests are widely used in many scientific fields to assess the evidence against the null hypothesis by calculating the probability of obtaining the observed result, or a more extreme result, if the null hypothesis were true (Hernandez-Sampieri et al., 2014; Trafimow, 2019).

This probability is represented by the *p*-value, which is compared to a predetermined threshold, usually .05 (5%; although a lower threshold of .005 has also been proposed by Benjamin et al., 2018) in Social sciences, to determine whether the result is statistically significant or not. If the *p*-value is less than the predetermined threshold, it is typically concluded that the observed result is statistically significant, and the null hypothesis is rejected in favour of the alternative hypothesis. On the other hand, if the *p*-value is greater than the threshold, it is concluded that the observed result is not statistically significant, and the null hypothesis cannot be rejected (Wasserstein & Lazar, 2016).

In the following section, we will examine how questionable research practices (QRPs) and poor statistical practices in significance testing impact the replicability and generalizability of research findings.

### *Questionable research practices impact on Transparency, Replicability and Generalizability*.

Transparency, replicability, and generalizability are closely tied to the process of validating scientific knowledge because they are essential characteristics that ensure the

verifiability and reliability of scientific claims (Bunge, 2018). As will be developed further, these three methodological criteria are embedded in the execution and publication of a scientific study and are interrelated in such a way that difficulties in one can directly or indirectly impact the others.

In recent years, growing concerns about research integrity have emerged, particularly in light of several high-profile cases of scientific misconduct, fraudulent research, and Questionable research practices (QRP) (Catanzaro, 2023; John et al., 2012; Meho & Akl, 2025; O'Grady, 2024; Open science collaboration, 2015; Yarkoni, 2022). The former intentionally aims to circumvent the proper design or development of scientific research, which leads to a discussion about the ethics of scientific practice and the functioning of the scientific publishing system; whereas QRP are not necessarily carried out with that intention. On the other hand, QRP may occur due to omissions in the research design or constraints during the research process, such as limited access to a statistically appropriate sample or interruptions in the data collection process (John et al., 2012). Some of the key issues that contribute to these problems include a lack of transparency in data processing, weak experimental designs, low statistical power and poor statistical practice (Open science collaboration, 2015; Poldrack et al., 2017).

Some of the QRP are related to not following the typical process of designing a research and its implementation process, which constitutes a lack of transparency and could impact the replicability or generalizability of results. The QRP described here can be addressed by following the methodological steps outlined in the research design. However, the second and third QRP described could be addressed through statistical strategies. The first one, Hypothesizing After the Results are Known (HARKing) refers to the practice of researchers developing hypotheses after they have already analysed their data (Kerr, 1998). This practice increases the risk of false positives (Type error I) because hypotheses are not

based on a priori knowledge, but rather on the observed data. Another consequence is a reduction in the replicability of research findings because the hypothesis is not formulated based on previous studies (Kerr, 1998; Rubin, 2022).

In this way, Sasaki and Yamada (2023) introduced a QRP defined as Sample-size Planning After the Results are Known (SPARKing). This might be related to concerns about the study's sample size and that is justified by researchers in a posteriori fashion. As a result, the experimental design is less transparent and the difficulty observed in relation to the robustness of the results obtained is overcome with a justification created for that purpose. To determine sample size in a non-arbitrary manner, statistical power analysis software, such as G*Power (Faul et al., 2007) or simulation studies (Kumle & Draschkow, 2021) can be employed. The next section will further explore good practices in conducting simulation studies.

Another widely reported QRP is the P-hacking, which refers to the selective reporting or manipulation of statistical analyses to achieve statistically significant results (Head et al., 2015). This can include running multiple analyses until a significant result is obtained, or selectively excluding data points or variables to achieve a desired outcome (Andrade, 2021). In the same line as previously reported QRP, this practice decreases the transparency of the experimental design and can produce false positives due to looking for patterns or relationships that are statistically significant, without any a priori hypotheses or a clear theoretical rationale for testing a specific hypothesis. However, Multiverse analysis (Steegen et al., 2016) serves as an effective methodological strategy to mitigate p-hacking by standardizing criteria for data exclusion, variable transformation, and group selection. This approach prevents biased data selection driven by *p*-value attainment through the systematic visualization of all possible data processing decisions and their impact on outcomes. By doing so, it precludes the concealment of variability in findings, enabling researchers to

evaluate the robustness of results and detect whether outcomes shift depending on analytical decisions.

These QRPs, along with the concepts explored in this chapter, must be contextualized within the frameworks of statistical significance testing and studies grounded in inductive reasoning, where conclusions are derived from observations supported by inferential statistical methods (Bunge, 2018; Hernandez-Sampieri et al., 2014). It is critical to emphasize that certain research designs incorporate flexible components—such as sample selection or hypothesis formulation—which may be adjusted or omitted during the study's execution. This consideration is not trivial, as it also necessitates critical reflection on methodological aspects prior to design—for instance, the balance between internal and ecological validity—to delineate the boundaries of the knowledge generated and ensure the generalizability of inferences.

In this context, one promising approach to promoting methodological rigor and addressing some of the challenges that lead some researchers to conduct QRPs, particularly in areas like sample size determination and evaluating analytical choices, lies in the application of simulation studies, which will be explored in the subsequent section.

### *Transparency and Replicability in Simulation Studies*

Simulation studies involve creating a model or simulation of a system, and then using that model to generate data that can be analyzed using statistical methods. Simulation studies are conducted to ascertain an appropriate design for a particular investigation. As previously mentioned, simulation studies can be employed for sample size planning or to assess the robustness of statistical methods under varying conditions. Additionally, these studies are valuable for comparing different statistical analysis, enabling researchers to identify the most

appropriate approach for their specific research context (Friedrich & Friede, 2023; Morris et al., 2019).

Transparency and replicability are just as important in simulation studies for estimating sample sizes, resample data from existing datasets, or serve as standalone studie to determine the properties of a statistical procedure. To be considered valid, it must be possible for other researchers to replicate the study using the same model and methods. This requires careful documentation of the simulation model and parameters, as well as clear reporting of the statistical methods used to analyze the data (Luijken et al., 2024). Simulation models are often highly specific to a particular system or set of conditions, and it can be difficult to generalize the findings to other contexts or populations. To address this issue, researchers can use sensitivity analyses to examine how changes in the model affect the results. They can also collaborate with other independent researcher groups to compare the results and produce more generalizable findings (Pawel et al., 2023).

### *Open Science and Collaboration as a Framework for Enhancing Transparency in Psychological Studies and AI Research*

Contemporary scientific practice relies on collaboration, a necessity driven not only by the inherently collective nature of scientific inquiry but also by the multistage complexity of modern research. High-quality investigations often require specialized expertise at distinct phases of the process: statisticians ensure methodological rigor, computational scientists design algorithms, psychometricians validate measurement tools, and domain experts contextualize findings. This interdependence underscores the need for coordinated efforts across disciplines to address the multifaceted demands of rigorous research.

This is critical even in research contexts that do not involve the collection of human data, such as computational simulation studies or the development of large-scale AI models,

where reproducibility remains foundational to validate models and their results. In these domains, collaborative efforts are indispensable for addressing persistent challenges in computational reproducibility and technological innovation, ensuring that models and algorithms can be independently verified, generalized across contexts, and iteratively improved by the broader scientific community (Gundersen et al., 2018; Stodden et al., 2014). Moreover, such collaboration inherently promotes transparency and replicability in IA research, as it necessitates the open sharing of research data, training datasets, code, methodologies, and findings with both the scientific community and society at large (Haibe-Kains et al., 2020; Prieto, 2022; Toribio-Flórez et al., 2021).

In this way, open science is an approach to research that emphasizes transparency, collaboration, and accessibility to scientific knowledge. In addition to traditional research methods, Open science also involves using new technologies and tools, such as open data repositories (i.e. Open Science Framework; Foster & Deardorff, 2017), collaborative research platforms (i.e. Overleaf), platforms to store and share code (i.e. GitHub), repositories of trained machine learning (ML) models (i.e. Hugging face, TensorFlow Hub) and pre-registration of study protocols (Pownall et al., 2021). Consequently, these technological infrastructures require workflows that facilitate collaborative efforts. While advanced tools have emerged to support joint efforts, their effective implementation depends on parallel workflows that standardize processes, streamline communication, and embed collaboration into the research design itself. Such frameworks ensure that teams can efficiently leverage technological resources while adhering to shared protocols for transparency and reproducibility.

As AI models become more complex and powerful, it is important to ensure that they are transparent, explainable, and ethical (Haibe-Kains et al., 2020; Kapoor & Narayanan, 2023). Open science involves sharing data, software, and research materials to help ensure AI

systems are developed in a responsible manner, and that their potential risks and benefits are fully understood. For example, it would allow the identification and evaluation of errors such as variable repetition; the presence of low-quality data (i.e., images) within the dataset that could negatively influence the integrity and accuracy of the obtained results (Roberts et al., 2021). Also, it is necessary to consider the possibility that the material used contains personal information, sensitive data, or explicit content, which could have privacy and ethical implications.

Besides, transparency in AI systems has gained new dimensions with recent technological shifts (Larsson & Heintz, 2020). First, the emergence of reasoning models enables it to articulate its decision-making steps in natural language during problem-solving. This innovation allows users to trace the model's logical pathways (e.g., "I will first analyze the premise, then compare historical precedents..."), overcoming the "black box" reasoning nature of previous Large Language Models (LLM) and fostering trust through explainable intermediate steps that are useful to the common user. Secondly, DeepSeek's disruptive release of a free public API redefines openness standards in LLM (Liu et al., 2024). By providing unrestricted access to its architecture, DeepSeek disrupts the industry’s status quo, advancing a shift toward open science principles in AI research and development. This dual transparency creates unprecedented accountability: while procedural explanations empower end-users to audit AI reasoning, open API availability lets researchers scrutinize model behavior at scale and develop or adapt bespoke systems for specific applications (Bommasani et al., 2021, Sapkota et al., 2025).

***Many-Lab Approaches: Enhancing Generalizability Through Collaborative Research***

One of the challenges in generalizing the results of studies involving human data lies in the potential influence of cultural, economic, educational, and political variables on the

treatment variable. In this way, the concept of WEIRD societies (Western, Educated, Industrialized, Rich, and Democratic) has become increasingly important in recent years, as researchers have recognized that many research studies are based on samples that are not representative of the global population (Alves et al., 2022; Henrich et al., 2010). This can lead to difficulties in the generalization of the findings in fields where the variables to be investigated depend on sociocultural variables, for example in moral psychology or education studies (Alves et al., 2022; Yarkoni, 2022). These studies lack representativeness of national populations due to non-probabilistic sampling frameworks or reliance on convenience samples drawn from university student populations. In contexts marked by significant socioeconomic disparities or stratified access to education—common in low- and middle-income countries—such methodological limitations critically undermine external validity, rendering findings inapplicable to broader demographic groups (Henrich et al., 2010).

One way to address this issue is through the use of Big Team Science, multi-lab/many-lab approaches. These ones involve conducting large-scale, collaborative studies across multiple labs and locations, using diverse samples from different populations and cultures. This can help to ensure that research findings are more generalizable, robust and applicable to a wider range of populations (Baumgartner et al., 2023; Yarkoni, 2022).

Many-lab approaches also help to increase the transparency and replicability of research, as multiple labs can independently test the same hypotheses and methods. This can help to identify inconsistencies or errors in the research, and improve the overall quality of the findings (i.e.: Coles et al., 2022; Morey et al., 2021; Wagenmakers et al., 2016). To successfully implement Big Team Science and multilab approaches, it is imperative to establish clear governance frameworks and robust collaboration agreements from the outset, ensuring that leadership roles, decision-making processes, and conflict resolution mechanisms are explicitly defined. As previously discussed, the Open Science Framework,

encompassing both technological infrastructure and conceptual workflows, provides a scaffold for collaborative research by standardizing shared protocols. Moreover, fostering an inclusive, interdisciplinary team through targeted recruitment and equitable authorship policies (Allen et al., 2014), exemplified by structured systems like the Contributor Roles Taxonomy (CRediT), not only integrates diverse perspectives but also reinforces methodological rigor and ultimately enables the production of robust, generalizable scientific insights.

In the next sections, we refer to the TASI model (Trafimow, 2019) as a framework to categorize and evaluate the different types of assumptions researchers make in testing theoretical predictions to ensure the validity and reliability of their findings. The TASI model stands for Theoretical, Auxiliary, Statistical, and Inferential assumptions. This model helps in understanding and evaluating the robustness of research findings by considering the following four categories of assumptions: *Theoretical Assumptions*: assumptions related to the underlying theory being tested; *Auxiliary Assumptions*: assumptions necessary to bridge the gap between theoretical concepts and actual observations; *Statistical Assumptions*: assumptions related to the statistical methods and models used in the analysis; and *Inferential Assumptions*: assumptions needed to make statistical inferences about populations from samples (see Figure 1).

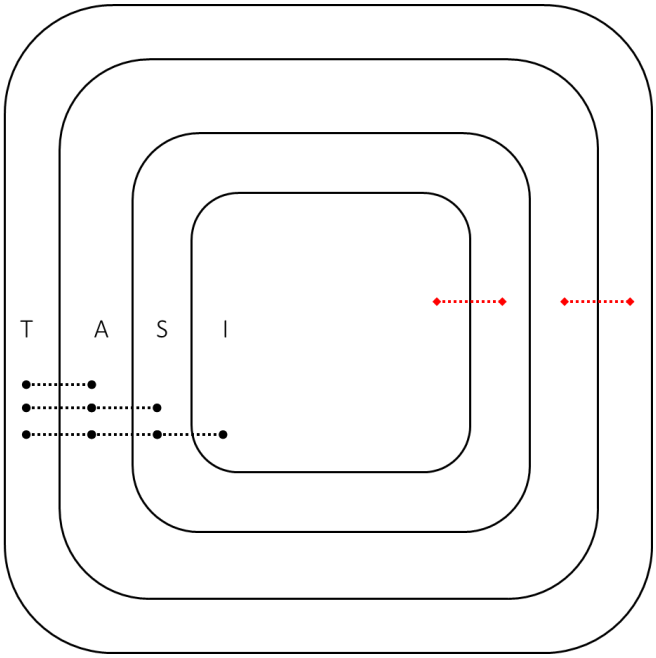


*Figure 1*. Representation of the components in the TASI model and their relatedness. T = theoretical assumptions, A = auxiliary assumptions, S = statistical assumptions, I = inferential assumptions. The black round dotted lines with oval arrows represent three variants of the TASI model; TA, TAS, and the full TASI model. The red round dotted lines with diamond arrows indicate potential ad-hoc strong bonds (i.e. tight bonds between T and A and S and I).

The simplest version of the TASI model is the TA model (see Figure 1). This model only includes theoretical assumptions and auxiliary assumptions and is used when researchers focus on connecting theory to observable phenomena. For example, studying working memory processing using cognitive load theory (*T*) plus assumptions about brain activity patterns and behavioral responses (*A*) allows researchers to connect abstract theoretical concepts about memory and neural processing to concrete, observable measurements without requiring complex statistical or inferential assumptions. The TAS model adds statistical assumptions to the TA model and is used when researchers need to analyze data but do not need to make population inferences. However, it includes choices about what statistics to use and how to analyze data. Although more complex than TA, it is simpler than a full TASI model. For example, studying working memory processing using cognitive load theory (*T*),

plus assumptions about brain activity patterns and behavioral responses (*A*), plus statistical assumptions about how to analyze brain activation patterns, including choices about averaging BOLD signals across brain regions, normalizing response times, and using parametric tests for comparing conditions (*S*) allows researchers to analyze relationships between memory load and brain activation using appropriate statistical techniques, while avoiding the additional complexity of population-level inferential assumptions. The full TASI model, as described by Trafimow (2019), is challenging because it necessitates that all inferential assumptions about populations and sampling be correct for *p*-values to be required. This full model may never be perfectly correct in real research, as it is rare for all assumptions to be entirely accurate.

As previously noted, the TA model explicitly links the *T* and *A* components within the TASI framework. Additionally, there is an ad hoc connection between the *S* and *I* components, given that both appear to share methodological ties by dealing with the quantitative analysis of research data. Statistical assumptions guide how data should be analyzed and determine which statistical tools are appropriate, while inferential assumptions determine how these analyses can be utilized to draw broader conclusions. These components are frequently considered together when planning data analysis strategies, and changes in statistical approaches often necessitate corresponding adjustments in inferential assumptions (see Figure 1). For example, in statistical learning, the connection between statistical and inferential components becomes evident when considering a ML model using cross-validation for performance estimation (cross-validation typically works by splitting the data into *k* subsets (folds), training the model on *k*-1 folds, testing it on the remaining fold, repeating this process so each fold serves as the test set once, and averaging the results). The *S* assumptions involve the stability of data distribution across training and test sets, the consistency of feature relationships across different data splits, and the appropriateness of

chosen performance metrics and validation methods. These *S* assumptions are inherently linked to their *I* counterparts: assumptions about how representative the training data is of the target population, how well the model will generalize to unseen data, and whether test set results truly indicate real-world performance. This interconnection manifests clearly in practice - if researchers' statistical assumption about data distribution stability is violated, their inferential assumption about generalization becomes compromised. Similarly, their choice of cross-validation method directly shapes what they can infer about model generalization, while their selection of performance metrics determines what conclusions they can draw about real-world applications. How researchers handle feature relationships statistically impacts their ability to make meaningful predictions about new data, demonstrating how statistical choices and inferential conclusions are tightly coupled in statistical modelling (see Figure 2).

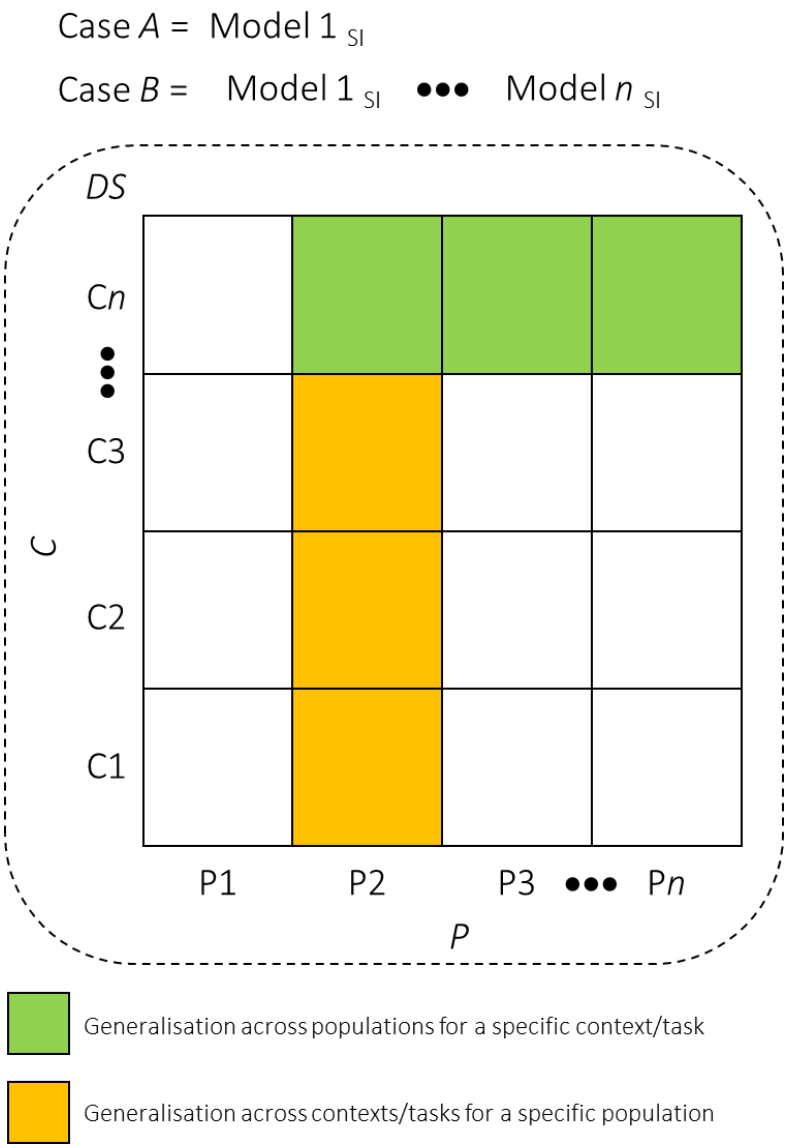


*Figure 2.* Relationship between statistical and inferential assumptions considering target populations and research contexts. DS = design space or space of potential experiments requiring a combination of target populations (P) and contexts or tasks (C). In case *A*, generalizations across populations (green) or contexts (orange) result from a specific SI model. In case *B*, generalizations across populations (green) or contexts (orange) result from more than one SI model.

Next, the TASI model is revisited in the context of the inferential knowledge (including significance testing, severity testing, and Bayesian approaches), region of acceptance testing, and statistical learning.

## Substantiating Inferential Knowledge

Today's replication crisis in science can be considered a testing crisis: We fail to reproduce past results ('reproducibility crisis'), we fail to replicate current results ('replication crisis'), and, thus, we have little confidence that this state of affairs will change in future ('crisis of confidence', Fanelli, 2018; Frias-Navarro et al., 2020). Significance testing is a prominent scapegoat for such crises, underlying the many problems pinpointed above, such as poor statistical practices, P-Hacking, HARKing and publication bias. In this section, however, we shall argue that it may all boil down to an underlying problem with inferential soundness and not with testing itself.

Inferential knowledge aims to achieve solid reasoning, and solid reasoning rests on logical structures from which a conclusion is formally derived from its premises. For a conclusion to be logical, it needs to be deduced from a valid logical structure with sound premises (e.g., Grimes, 2019). The correct inference naturally derives from ensuring sound premises and following the correct procedure, such as *Modus Ponens* or *Modus Tollens*.

The main problem we have is that our knowledge is not given but discoverable, not based on deducting truths but on learning from experience in an inductive manner. Popper (1962) provided the necessary bridge to move from induction to deduction by setting our theories as hypotheses (a.k.a., provisional truths), to be falsified via *Modus Tollens*. The refutation attempt ought to be a severe one, as the dependability of the conclusion rests on the soundness of the theory and on valid and reliable methods of observation and experimentation (e.g., Popper, 1962; Mayo, 2018).

### *Sound theories in the Inferential tree*

Trafimow's (2019) TASI heuristic can help us structure the inferential tree, from theory to conclusions, by mapping it onto *Modus Tollens*. *Theoretical* and *Auxiliary assumptions* (TA)

comprise the first premise of the syllogism, with theory (or particular hypotheses, HA) acting as the *Antecedent* of the premise (see Figure 3). The theory serves to set the expected outcomes (O) to be observed, thus setting the *Consequent* for the premise too. In a nutshell, TA links a theory (or particular hypothesis) with the theoretical results expected from such theory. The primary concern for this first premise is its soundness. Not just the theory ought to be sound (auxiliary assumptions play a key role here), but the link between theoretical *Antecedent* and expected *Consequent* needs to be sound too. If the theory is true, then certain outcomes shall be expected. If we encounter outcomes that contradict those expectations, we shall deduce the theory is false.

| If TA \| HA | Then, O |
|---|---|
| | Not O |
| Thus, no TA \| HA | |

*Figure 3.* The *Modus Tollens* inferential process: Theoretical and auxiliary assumptions (TA) help make the main premise sound, and the conclusion reliable.

***Reliable methods for testing***

The second premise of *Modus Tollens* is where the testing of the first premise occurs. Here is where we find the *Statistical* and *Inferential assumptions* (SI) of the TASI heuristic, as procedures, data collection, and data analyses also need to be sound (i.e., valid, and reliable– see Figure 4).

| If TA \| HA | Then, O |
|---|---|

| | Not O (given SI) |
|---|---|
| Thus, no TA \| HA | |

*Figure 4.* The *Modus Tollens* inferential process: Statistical and inferential assumptions (SI) help make the minor premise sound, and determine its reliability when denying the consequent of the main premise

The main problem with the second premise is that it can proceed in two ways. The correct path is towards falsifying the first premise (e.g., HA: All swans are white. O: A black swan is observed. Therefore, Conclusion: The hypothesis is false). The incorrect path is towards confirming the first premise (e.g., O: Another white swan is observed. Conclusion: the hypothesis is proved, or corroborated). The former path is the formal deduction allowed by *Modus Tollens*, thus a logical conclusion. The latter path is a formal fallacy known as *Affirming the Consequent*, thus leading to a formally illogical conclusion.

Tests of significance were born at a time when there were no stringent protocols for experimentation nor any other statistical technology down the *Modus Tollens* path. Before those tests, both descriptive statistics and Bayesian technology were used to confirm hypotheses, thus rendering the path into a fallacious conclusion. Fisher (1925, 1935; also Neyman, 1967) was the first statistician providing a basis for severity testing via experimental control and *p*-values, as a tool for falsifying hypotheses. The *p*-value has been so much confused since then that it may seem to offer little value today, but it was created when there was nothing else–contemplating alternative hypotheses and statistical power would need of Neyman and Pearson's constructs (e.g., 1928, 1942), effect sizes and sample size calculations for power would need of Cohen's contributions (e.g., 1988), and confidence intervals would also be derived from Neyman's proposal (1935).

The primary concern with *p*-values and, thus, significance tests, is that they are misused not to falsify hypotheses but to confirm them, thus going down the path of the confirmation fallacy. It is this bias towards confirming hypotheses that seems to underlie the myriad of breaches of statistical and inferential assumptions (SI) described earlier in the chapter. Therefore, most problems with significance testing are born out of attempts by researchers–intentionally or unwittingly–to support or confirm their own theories and hypotheses. Furthermore, such problems are not exclusive to significance tests. It will not really matter if we substitute effect sizes for *p*-values, for example. As long as we are intended to confirm rather than falsify, then the misuse of statistical and methodological assumptions are not going to disappear.

***Refutation, Severity, In/Out Model Inferences***

Among approaches exhorting the *Modus Tollens* inferential tree, we have Popper's falsificationism (1962), which mostly focuses on the first premise of the *Modus Tollens* syllogism. Popper's falsificationism offers a viable solution to the problem of induction: We can learn by proposing provisional hypotheses and submitting them to stringent tests for potential refutation. Science advances by proposing novel hypotheses to test, improving weak theories by trimming off failing assumptions, and rejecting bad theories that fail to pass those tests. The most recognisable motto of Popper is that a scientific theory is demarcated from a pseudoscientific one insofar the former is open to be tested to fail, and genuine attempts are made to do so.

Fisher's (1954), and Neyman-Pearson's (1928) approaches on research methods and statistics provided a workable environment for developing the second premise of the *Modus Tollens* syllogism. Pearson and, especially, Neyman, were most concerned with mathematics than

with philosophy, yet provided enduring tools for statistics beyond null hypotheses and *p*-values, such as alternative hypotheses, Type I and Type II errors, confidence intervals, and–mostly popularised by Cohen years later (e.g., 1962)–sample size, and power.

Mayo's approach (e.g., 2018) follows Popper's philosophy more closely, albeit it mostly focuses on the second premise of the *Modus Tollens* syllogism. The main take of Mayo's philosophy is that testing differs in their quality, from no severity (BENT, bad evidence, no test), to weak severity, to strong severity. Plus, not just the one test is enough: Multiple severe tests in agreement are needed to better substantiate whatever statistical inference is derived from them. Working with the disjunctive-hypotheses strategy inherited from Fisher's and Neyman-Pearson's approaches, Mayo seeks a reliable method to support a hypothesis via falsifying its opposite: We have evidence for a hypothesis just to the extent it survives a stringent scrutiny with severe tests designed to falsify it. (Mayo's corroboration argument based on a principle of coincidence among tests, however, is eerily close to flip-flopping towards the fallacy of *affirming the consequent*–also Schurz, 2019, similarly criticising Popper on his corroboration argument. However, it is technically possible because the disjunctive-hypotheses strategy used by frequentists aligns with that Sherlock Holmes's intuition (Doyle, 1927) that when you have eliminated all which is impossible–a.k.a., the null hypothesis–then whatever remains, however improbable,–a.k.a., the untested alternative–must be the truth.

Taleb's (2007) work may be substantiating another worthy approach, also targeting the second premise of the *Modus Tollens* syllogism. Most inferential attempts have been on rejecting those hypotheses that fail to pass null hypothesis tests (whether severely or not). Taleb (2001), most concerned with not being fooled by our own inferences, provides a path to also contemplate the hypotheses that are rejected. Namely, what is not rejected remains in the corpus of current knowledge and, thus, in our awareness, so they will not surprise us. And

yet, it is what is rejected (the highly improbable yet highly consequential) that could come back to bite us in the end–something particularly relevant in volatile fields such as economics or safety management.

Finally, the Bayesian approach (e.g., Kruschke, 2015) is worth considering too. Although Bayesian philosophy is eminently enumerative and confirmatory, thus down the path of the formal fallacy of *affirming the consequent*, there is no reason why it cannot be flipped to play a falsificationist role. Indeed, rather than focusing on the most supported hypothesis as the most credible, we can test hypotheses severely and discard the least supported ones as least credible. One way of doing so is to use priors to make Bayesian tests more severe (e.g., instead of assuming priors of 50/50 when we know nothing about the hypotheses, we could 'bias' the priors 95/5 in favour of whatever 'main hypothesis' we are using). If posterior probabilities still rule against the most supported hypothesis, such results serve to falsify it in the same manner than a frequentist approach may do, and severely so.

The advantage of Bayesian methods lies in their ability to move beyond the singular perspective of frequentist statistics. For instance, in Jeffrey's approach (1961), analyzing the same data under both normal and Cauchy distributions is typical. However, comprehensive Bayesian methods enable the simultaneous consideration of various models, as well as model-free Markov Chain Monte Carlo simulations (e.g., Metropolis et al., 1953). A second benefit is that Bayesian approaches are more interested in the hypotheses or model themselves [(p(H|D)] and not just on the location of the data in a single distribution [p(D|H)]. Thus, using a Bayesian approach in a falsificationist manner is not only possible, but also accords well with the logical path of *Modus Tollens*, thus, rendering formally logical conclusions.

In sum, lack of reproducibility and lack of replication are not negative in themselves, as they signal a failure of theories and hypotheses to hold. The problem is when reproducibility

and replication cannot be trusted because theories become impervious to being falsified (thus, becoming unsound, even pseudo-scientific), and methods are biased towards confirming those theories (thus, becoming invalid and unreliable). Good inferential work benefits from two complementary mindsets: (a) serious attempts at putting our theories and hypotheses to severe tests that may falsify them, and (b) an overall eagerness not to be fooled by our own biases towards confirmation. The two mindsets need not be restricted to just methods and statistics, but should permeate theories and hypotheses as well. Poor attempts should be considered BENT, opening us to be fooled by our own decisions and procedures. Weak probing already works in the right direction and gives us good leeway to learn from our research work (e.g., screening out auxiliary assumptions whose role seems to be that of saving our theories, pilot studies, exploratory data analysis…). Strong probing is the cornerstone of highly sophisticated inferential work (e.g., open science, reproducibility, replication, meta-analysis…). Tests of significance are appropriate tools in the inferential toolbox. However, they are so much misunderstood and misapplied that today's researcher might be less fooled if she were to substitute alternative tools that came afterwards: confidence intervals (the other side of a test of significance, but centred on the sample data rather than a theoretical null hypothesis, e.g., Perezgonzalez, 2015), effect sizes (to probe practical importance, not just statistical significance), replication and meta-analysis (to probe via a principle of coincidence), etc. Bayesian approaches, if done with severity in mind, can be repurposed to play a falsificationist role and may prove to be quite valuable both for sensible inferences about hypotheses themselves and, more importantly, for substantiating the need for further replication in order to maximise such inferences beyond reasonable doubt.

## REACT: A New Perspective on Hypothesis Testing

Trafimow (2019) argues that *p*-values offer no real benefit to scientific inference. His primary critique, framed within the TASI model, is that *p*-values conflate effect size and sample size, obscuring their contributions to statistical conclusions. Concretely, he argues that because no TASI model is perfectly correct, the usefulness of *p*-values in determining whether a model is "close enough" to reality is questionable. With large sample sizes, any TASI model will eventually be rejected, even if the truth is still in the ballpark of TASI. Similarly, small samples may fail to detect even substantial effects due to high variability. Trafimow thus advocates for reporting effect sizes and sample sizes separately rather than merging them into a single measure like the *p*-value.

In this section, we review REACT (Izbicki et al., in press) and demonstrate that it provides a stronger alternative to *p*-values under the TASI framework. REACT extends equivalence testing (Schuirmann, 1987) and three-way hypothesis testing (Berg, 2004; Esteves et al., 2016; Coscrato et al. 2019), integrating effect sizes and confidence sets into a structured decision-making process. Unlike NHST, which forces binary conclusions of "reject" or "fail to reject" the null hypothesis, REACT introduces an agnostic decision, explicitly recognizing when the available data is insufficient to support a definitive conclusion (for an R implementation visit https://github.com/Monoxido45/REACT).

### *The REACT Method*

Let $\theta$ be a parameter of interest, such as the difference in mean systolic blood pressure reduction between two antihypertensive drugs or the difference in anxiety reduction scores between two therapeutic approaches. REACT operates by defining a null hypothesis not as a single point but as a region of practical equivalence (also called the pragmatic region; Esteves et al., 2019), denoted by $\Theta_0$. For instance, $\Theta_0$ may have the shape

$\Theta_0=\{\theta:|\theta|<\Delta\}$,

where $\Delta$ is a threshold chosen based on domain knowledge, representing the smallest effect size that would be considered practically meaningful. By asking the scientist to specify $\Delta$, REACT invites them to define a plausible and relevant theoretical range for the hypothesis. In the TASI framework, this means that REACT requires the scientist to specify a more plausible theory.

In many settings, it is easier to specify $\theta$ in terms of effect sizes. For example, it could be Cohen's d, which measures the standardized difference between two group means, with $\Delta = 0.2$ often considered a small effect. In a medical application, $\theta$ might also be the distance from the hazard ratio and 1, and $\Delta = 0.1$ could define the smallest increase in risk deemed clinically relevant. In economics, $\theta$ might be a percentage change in income, with $\Delta = 2\%$ setting the threshold for a meaningful policy effect.

Once $\Theta_0$ is defined, the method then consists of the following steps, illustrated in Figure 5:

1. **Construct a Confidence Region:** Given a dataset $\mathbb{D}$, construct a confidence set $C(\mathbb{D})$ for the parameter $\theta$. This set may be frequentist or Bayesian.
2. **Compare to the Equivalence Region:** Compare the confidence set to the pre-specified equivalence range $\Theta_0$, which defines a meaningful difference from the null. One of the following conclusions is drawn:
    - Accept $H_0:\theta\epsilon\Theta_0$ if $C(\mathbb{D})\subseteq\Theta_0$, meaning all plausible values fall within the equivalence region.
    - Reject $H_0:\theta\epsilon\Theta_0$ if $C(\mathbb{D})\cap\Theta_0 = \emptyset$, indicating that all plausible values lie outside the equivalence region.
    - Remain agnostic otherwise, acknowledging that the data does not provide sufficient resolution to make a strong conclusion.

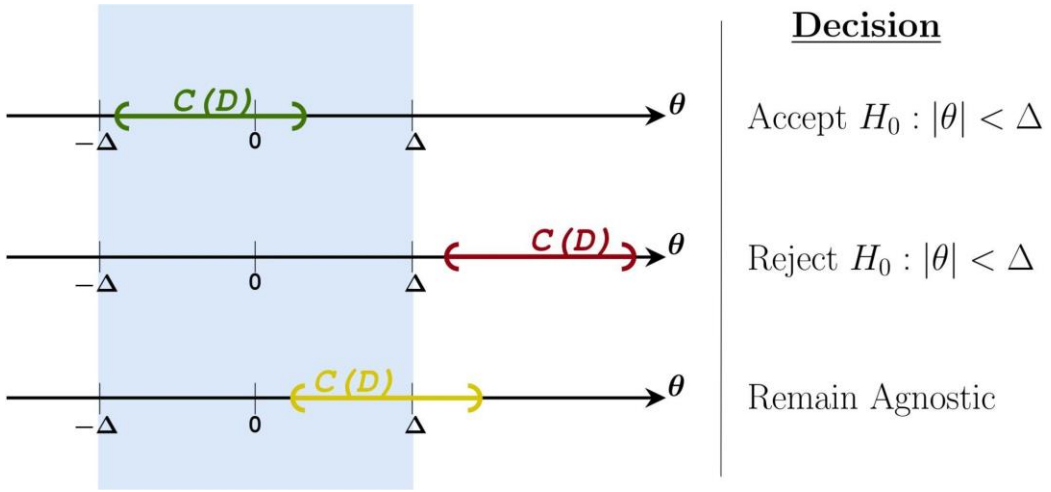


*Figure 5.* Illustration of the REACT procedure. A confidence set for C(𝔻) is build. If the entire set lies within the null hypothesis $H_0$:θϵΘ$_0$, the hypothesis is accepted. If it lies entirely outside, it is rejected. Otherwise, the procedure remains inconclusive.

REACT presents several advantages over traditional NHST methods (see Izbicki et al. 2025 for details):

1. **Identifying Evidence of Absence vs. Absence of Evidence:** REACT explicitly distinguishes between cases where the null hypothesis is strongly supported and cases where the data is inconclusive. This is in contrast to standard hypothesis tests, in which a non-rejection of the null hypothesis can be interpreted as acceptance or failure to reject it (Edwards et al. 1963, Neyman 1976).
2. **Logical Coherence:** Unlike traditional NHST, which can lead to paradoxical conclusions, REACT ensures that conclusions remain logically consistent across multiple hypothesis tests. For instance, concluding that medication A is better than B and that B is better than C, entails the conclusion that A is better than C.
3. **Application to Multiple Hypotheses:** REACT naturally extends to testing multiple null hypotheses simultaneously while controlling family-wise errors. This approach eliminates the need for ad hoc corrections such as Bonferroni adjustments.

4. **Consistency:** REACT prevents automatic rejections in large samples and ensures that small samples do not obscure meaningful effects. When the region of practical equivalence holds, the test will accept it as the sample size grows; otherwise, it will eventually lead to rejection. Additionally, for small samples with wide confidence intervals, the test remains agnostic to the null hypothesis. Thus, this approach directly addresses the limitations of *p*-values highlighted by Trafimow (2019), particularly their restricted informativeness in hypothesis testing.

***Examples illustrating how to REACT***

To illustrate REACT in practice, we apply it to assess whether specific variables significantly impact certain outcomes. We consider three examples: the association between daily red meat consumption and colon cancer (Di Maso et al., 2013), the relationship between sending SMS while driving and the risk of motor vehicle accidents (Kogani et al., 2020), and the link between smoking and lung cancer (Matos et al., 1998).

Let $\theta$ represent the odds ratio between the outcome and the variable of interest. We define three pragmatic hypotheses in the form $\Theta_0=\{\theta:\theta>\Delta\}$. In the first hypothesis, we set $\Delta=1.0$, indicating the presence of at least some effect. For the second hypothesis, we use $\Delta=1.5$, often considered to be a small effect size (Cohen, 1988). For the third hypothesis, we set $\Delta=3.5$, representing a medium effect size. Table 1 summarizes the results, which we describe in the sequence.

*Red Meat Consumption and Esophagus Cancer***.** The estimated odds ratio is 1.46, with a 95% confidence interval of (1.23, 1.72). For $H_0:\theta>1.0$, the entire confidence interval lies above 1, so REACT accepts that red meat consumption has some effect on colon cancer risk. However, for $H_0:\theta>1.5$, the confidence interval includes values below and above 1.5, making

642 REACT agnostic about the presence of a small effect; more data would be needed to reach a
643 conclusion. For $H_0$:θ>3.5, the confidence interval lies entirely below 3.5, so REACT rejects
644 the hypothesis of a medium effect.

645 *Texting While Driving and Motor Vehicle Accidents***.** The estimated odds ratio is 2.3,
646 with a 95% confidence interval of (1.2, 4.4). For $H_0$:θ>1.0, the confidence interval is entirely
647 above 1, so REACT accepts the presence of at least some effect. For $H_0$:θ>1.5, the confidence
648 interval also remains above 1.5, so REACT accepts that texting while driving has at least a
649 small effect on the risk of motor vehicle accidents. However, for $H_0$:θ>3.5, the confidence
650 interval includes values below 3.5, so REACT remains agnostic about whether the effect
651 reaches a medium magnitude, suggesting that more data are necessary to draw a definitive
652 conclusion.

653 *Smoking and Lung Cancer***.** The estimated odds ratio is 8.5, with a 95% confidence
654 interval of (4.3, 16.7). For $H_0$:θ>1.0, the confidence interval lies entirely above 1, so REACT
655 accepts that smoking has some effect on lung cancer risk. Similarly, for $H_0$:θ>1.5 and
656 $H_0$:θ>3.5, the confidence interval remains above the respective thresholds, indicating that
657 REACT accepts that the effect size for smoking on lung cancer is at least medium.

658

659 *Table 1*. REACT results for three problems and hypotheses
660

| | Confidence Interval for Odds Ratio θ | $H_0$:θ:>1.0 (at least some effect) | $H_0$:θ:>1.5 (at least a small effect) | $H_0$:θ:>3.5 (at least a medium effect) |
|---|---|---|---|---|
| Meat vs Esophagus Cancer | (1.23; 1.72) | **Accept** | **Agnostic** | **Reject** |
| Sending SMS vs Accident | (2.3; 4.4) | **Accept** | **Accept** | **Agnostic** |
| Smoking vs Lung Cancer | (4.3; 16.7) | **Accept** | **Accept** | **Accept** |

661

These examples illustrate how REACT provides more nuanced conclusions by incorporating confidence intervals and predefined effect size thresholds. This approach enables researchers to distinguish between different effect magnitudes and the strength of the supporting evidence. In contrast, relying solely on *p*-values would only test the null hypothesis that the odds ratio equals one, without offering insight into the practical relevance of the observed effects. Indeed, in all three datasets, such a test would reject the null hypothesis.

### *Limitations, Assumptions and Conclusions*

REACT assumes the ASI components of TASI are approximately correct. Thus, both rejection and acceptance of the theory are contingent on these components being good approximations. However, contrary to standard point null theories, the ASI assumptions can be good approximations in practice. Moreover, one can use nonparametric methods within REACT, which rely on weaker assumptions (Lassance et al., 2025).

REACT also depends on the choice of the threshold Δ. The value of Δ represents the smallest effect size deemed practically meaningful, and its specification often involves subjective decisions based on domain knowledge. If the chosen Δ is too small, the method might frequently reject hypotheses, while a large Δ could lead to excessive acceptance. Several strategies can be used for selecting Δ, including usage of prior literature, clinical relevance, or heuristic rules like the minimal clinically important difference. To assess the robustness of conclusions to the choice of Δ, researchers can perform a sensitivity analysis by examining the results under a range of plausible thresholds.

In sum, the REACT framework not only mitigates the shortcomings of *p*-values identified within Trafimow's TASI model but also offers a more nuanced perspective on hypothesis testing. By allowing for agnostic conclusions, REACT aligns more closely with the scientific principle of acknowledging uncertainty rather than forcing binary decisions.

Additionally, REACT provides flexibility by accommodating both frequentist and Bayesian approaches when constructing confidence regions. This adaptability ensures that researchers can choose the method best suited to their assumptions, making the inference process more flexible and context-specific.

## Statistical learning instead of hypothesis testing via GAMLSS

The SI sub-model is crucial to the TASI framework as it determines research validity and reliability. *S* assumptions govern statistical methods (distribution, independence, variance), while *I* assumptions address inferential reasoning (generalizability, causality, significance interpretation). These components form an inseparable bond—statistical integrity enables sound inference, and both must be satisfied for valid findings. Statistical violations inevitably compromise inferential quality. Together, they create a fundamental partnership that ensures research robustness and replicability, forming the backbone of scientific inquiry rather than merely serving as technical requirements.

Hypothesis testing serves as the traditional framework within the SI model, where a null hypothesis (assuming no effect or difference) is contrasted with an alternative hypothesis. This process involves assessing the probability of false rejection and results in a binary decision: either rejecting or failing to reject the null hypothesis. In contrast, we advocate for statistical learning as an alternative approach that entirely circumvents hypothesis testing, instead emphasizing descriptive, explanatory, and predictive statistical models. The proposed framework for this approach is Generalised Additive Models for Location, Scale, and Shape (GAMLSS) (Rigby & Stasinopoulos, 2005).

GAMLSS is a statistical learning framework that expands on traditional regression models by modeling all parameters of the response variable's distribution, rather than focusing solely on the mean. It extends linear models (LM) by supporting non-normal

distributions from the exponential family and employing link functions. Additionally, GAMLSS goes beyond generalized linear models (GLM) by integrating non-linear smooth functions of predictors. It further advances generalized additive models (GAM) by modeling all parameters of the distribution (not just location) and accommodating a wider variety of distributions (Stasinopoulos, Rigby, & de Bastiani, 2018) (see Figure 6).

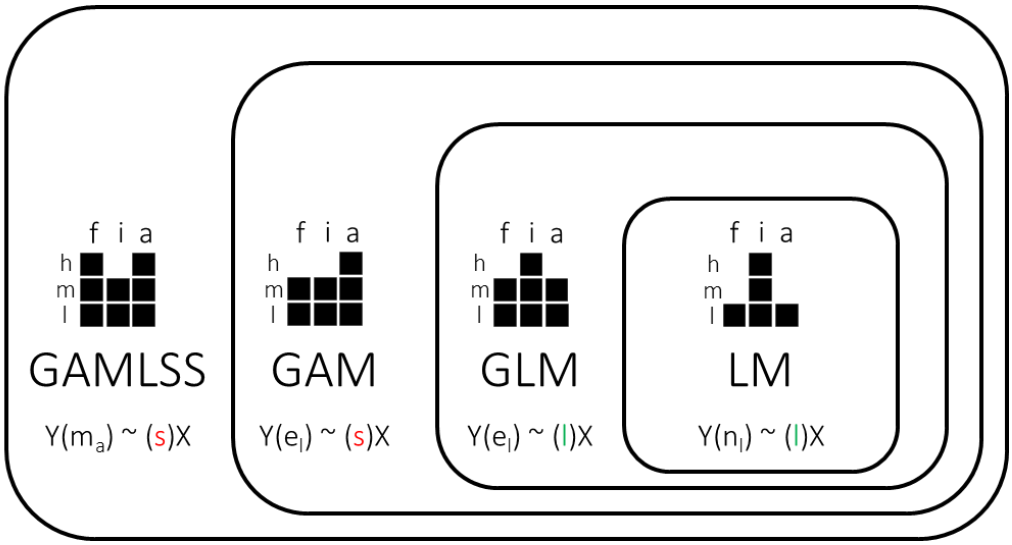


*Figure 6*. Illustration of the relationship among four regression approaches. LM = linear model (a.k.a., ordinary least squares or general linear model); GLM = generalised linear models; GAM = generalised additive models; GAMLSS = generalised additive models for location, scale and shape. In the models 'Y(•) ~ (•)X', Y stands for the dependent (continuous or discrete) variable and X is a design matrix (i.e. set of covariates or explanatory variables). The (fixed-effects) explanatory variables in X are categorical and/or numerical (note that only GAMLSS includes smoothers for both kinds of covariates). Random effects are modellable in all methods but only GAM and GAMLSS can also model random effects via smoothers (e.g. temporal and spatial effects). Modelling of the dependent variable Y(•): n = Normal distribution only; e = family of exponential distributions only; m = several family of distributions (note that m ⊃ e ⊃ n). The subscript by the dependent variable's distribution indicates that while LM, GLM, and GAM can assess the effects of covariates on the dependent variable's location 'l' only (e.g. the mean in the case of LM), GAMLSS can do so for all, 'a', the dependent variable's parameters (i.e. location, scale, skewness, and

kurtosis). Modelling of the covariates in X: s = smoothers, l = linear (note that s ⊃ l). The black squares represent each regression approach's levels of flexibility (f), interpretability (i) and accuracy (a) such that these can be high (h), medium (m) or low (l) (these levels are based on the authors' experience and are inspired in Figure 2.7 in James et al (2021) and Figure 12 in Barredo Arieta et al (2020). Note all these metrics need to be assessed empirically). Symbol '⊃' to be understood as superset.

As a statistical learning framework, GAMLSS enables the understanding, analysis, and extraction of patterns from data to support predictions or inferences. It connects traditional statistical methods (e.g., beta weights in parameters) with modern ML approaches (e.g., using neural networks to model numeric covariates) (Stasinopoulos et al., 2017; Rigby, Stasinopoulos, Heller, & de Bastiani, 2020; Stasinopoulos et al., 2024). Furthermore, as a regression framework, GAMLSS quantifies relationships while accounting for natural variation and random error. When applied correctly, with careful consideration of its underlying assumptions, GAMLSS models can produce reliable statistical inferences. Additionally, GAMLSS emphasizes interpretability by separating data into deterministic and random components, making complex relationships more understandable—ensuring that users can grasp how the model operates internally and how it transforms inputs into outputs.

The following section positions GAMLSS within the broader context of ML and related concepts. Additionally, it shows how GAMLSS can generate descriptive, explanatory, and predictive models, which can serve as a novel substitute for conventional hypothesis testing.

***Description, explanation, and prediction models via GAMLSS***

ML provides the computational tools that enable artificial intelligence (AI) systems to learn from data and improve through experience. These ML algorithms, from simple linear models to complex deep neural networks, derive their theoretical foundations from statistical

principles (Friedrich et al., 2022; Min et al., 2024). AI systems built using these ML techniques rely fundamentally on statistical concepts like probability theory, hypothesis testing, and experimental design to ensure their reliability and validity (Friedrich et al., 2022). The synergy between these fields is particularly evident in areas like explainable AI (XAI), where statistical interpretability methods help make AI decision-making transparent (Min et al., 2024). Rather than separate disciplines, ML, AI and statistics form an interconnected framework - with ML providing the algorithmic implementation, AI delivering the intelligent applications, and statistics supplying the mathematical rigor (Friedrich et al., 2022).

ML is a subfield of AI, and within ML lies statistical learning (SL), a branch grounded in statistics that prioritizes model interpretability and uncertainty quantification (James et al., 2021). GAMLSS exemplify this approach. A specialized method within SL is targeted learning (TL), which focuses on estimating causal or statistical parameters with minimal bias (Coyle et al., 2023; Van der Laan & Starmans, 2014). Notably, recent research suggests that GAMLSS models can be regularized to enhance causal inference (Marmolejo-Ramos et al., 2023). Additionally, ensemble learning (EL) is a versatile technique applicable across SL, TL, and broader ML. Inspired by the "wisdom of crowds," EL combines multiple models to improve predictive accuracy beyond what any single model can achieve (Sagi & Rokach, 2018). In ML, EL is exemplified by random forest, a bootstrap aggregation (bagging) method that combines multiple decision trees to enhance predictive performance (Breiman, 2001b). Similarly, in classical statistical hypothesis testing, Stouffer’s z-score method represents a form of EL by aggregating *p*-values from independent tests (Stouffer et al., 1949; Hoang & Dickhaus, 2022). The type of data used determines whether ML methods are categorized as supervised, semi-supervised, or unsupervised learning algorithms (see Figure 7).

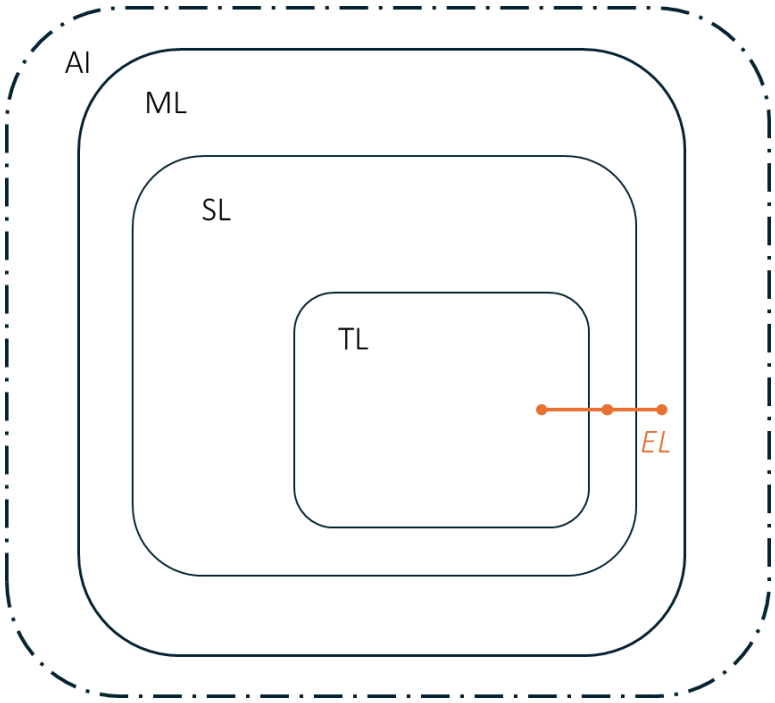


*Figure 7*. Relationship among artificial intelligence (AI), machine learning (ML), statistical learning (SL), targeted learning (TL), and ensemble learning (EL). Supervised learning algorithms work with labeled data (input-output pairs) to perform prediction or classification (as in SL), estimate causal parameters (as in TL), or implement methods like decision trees (in EL). Unsupervised learning algorithms analyze unlabeled data to discover patterns, though this approach is less common in SL (which emphasizes inference over exploration) and TL (which typically requires labeled data for causal inference), but appears in EL through clustering ensembles or consensus clustering (Vega-Pons & Ruiz-Shulcloper, 2011). Semi-supervised learning algorithms utilize small amounts of labeled data alongside larger unlabeled datasets for label propagation or self-training, with specific applications in TL for causal inference with missing data, in EL through ensemble semi-supervised classifiers (e.g., Zhao & Liu, 2021), and in emerging SL approaches such as semi-supervised GAMs (Culp, 2011) and semi-supervised GLMs (Tu et al., 2024).

SL encompasses a variety of methods and data types, but its core objective remains consistent: to build descriptive, explanatory, and predictive models that balance accuracy with interpretability. Descriptive models provide compact summaries of data, serving as a foundation for further analysis without delving into causality or prediction. Explanatory models, in contrast, focus on testing causal hypotheses, relying on theoretical assumptions to

uncover underlying mechanisms while minimizing bias. Predictive models, meanwhile, prioritize forecasting future outcomes, often employing data-driven techniques that favor accuracy over theoretical clarity (Shmueli, 2010).

The choice of approach in SL hinges on how researchers conceptualize the data-generating process (DGP)—the underlying system that produces the observed data (Breiman, 2001a). In some cases, the DGP is treated as stochastic, meaning it follows a probabilistic structure that can be modeled explicitly. Here, the emphasis is on statistical inference, where researchers estimate parameters, test hypotheses, and validate findings through measures like statistical significance. This data-driven approach aligns closely with explanatory modeling, as it seeks to uncover the true mechanisms behind the data while maintaining interpretability. However, when dealing with complex explanatory models where such inherent interpretability is elusive, achieving reliable understanding necessitates an interactive analysis that juxtaposes multiple explanatory methods to counter the misleading potential of any single perspective (Baniecki et al., 2023).

In other cases, the DGP is seen as too complex or unknown—effectively a "black box." Rather than trying to decipher the underlying process, researchers focus on finding algorithms that can reliably predict outcomes (Y) from inputs (X), regardless of theoretical fidelity. This algorithm-driven approach aligns with predictive modeling, where the primary goal is maximizing accuracy, even if it means sacrificing interpretability. Certain class of techniques in machine learning thrive in this domain, with models validated based on their performance on unseen data rather than their adherence to theoretical expectations.

We argue, though, while descriptive models are often overlooked in formal theory-building, they play a crucial role in exploratory analysis and can inform explanatory research by refining measurement tools or suggesting new theoretical directions. However, explanatory models are not without limitations—they face parameterization uncertainty

(ambiguity in model or parameter selection) and description uncertainty (gaps in theoretical understanding), both of which can affect their reliability (Draper, 1995) (see Figure 8).

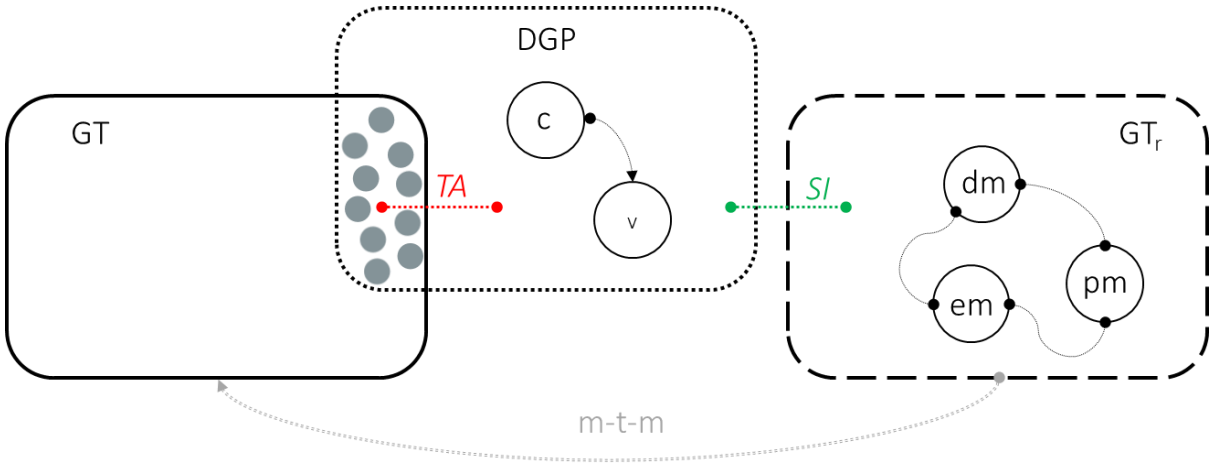


*Figure 8*. Human-data lifecycle. There exists a ground truth (GT), theorized to be generated by an underlying system (the data generation process; DGP). The dotted region denotes the unknown overlap between GT and DGP (i.e. $P(GT \cap DGP)$ is indeterminate). The TA component of the TASI model seeks to link GT and DGP, where the DGP governs the formation of latent constructs (c) and and their manifest variables (v). The SI component of the TASI model operationalizes GT into an empirical representation ($GT_r$) via descriptive (dm), explanatory (em), and predictive (pm) models. However, $GT_r$ is subject to distortion by systematic error (bias) and random error (variance), formalized as $GT_r = GT + bias + error$. Critically, the alignment between $GT_r$ and GT remains unquantified (i.e., $P(GT_r \cap GT)$ is unknown; but see below). All statistical models (i.e. dm, em, and pm) can be used to modify the GT through behavior modification techniques. The "model-then-modify" (m-t-m) strategy uses statistical models ($GT_r$) to actively influence and adjust real-world behaviors (GT), aiming to make those behaviors match the model's descriptions, explanations, or predictions. This establishes a cause-and-effect relationship where the model representation directly impacts the actual system ($GT_r \rightarrow GT$). Thus, $P(GT_r \cap GT)$ is not merely unknown but potentially manipulated through intentional intervention. This introduces a significant complication for the TASI framework, as the statistical model (empirical representation) does not just approximate

the GT but can directly reshape it. This reshaping carries the risk that the actual system (GT) might diverge from the original DGP. While this divergence might misleadingly appear to improve predictive accuracy, it could actually be distorting the underlying reality it seeks to capture. (Note: This concept draws from Shmueli & Tafti, 2023, although their 'predict−then−modify' approach specifically targeted predictive models).

***GAMLSS in action***

*Modelling cortical thickness*. It is well-established that aging is linked to widespread cortical thinning, especially in the frontal and temporal regions, which tend to be more susceptible than other areas (Salat et al., 2004; Piccolo et al., 2016). The relationship between average cortical thickness and age reveals a distinct pattern: a steep decline in MRI-based thickness estimates during childhood and adolescence, followed by a mild, continuous thinning from early adulthood onward, and an accelerated thinning trend in older adulthood (Vidal-Piñeiro et al., 2020). This non-linear association is illustrated in Figure 9.

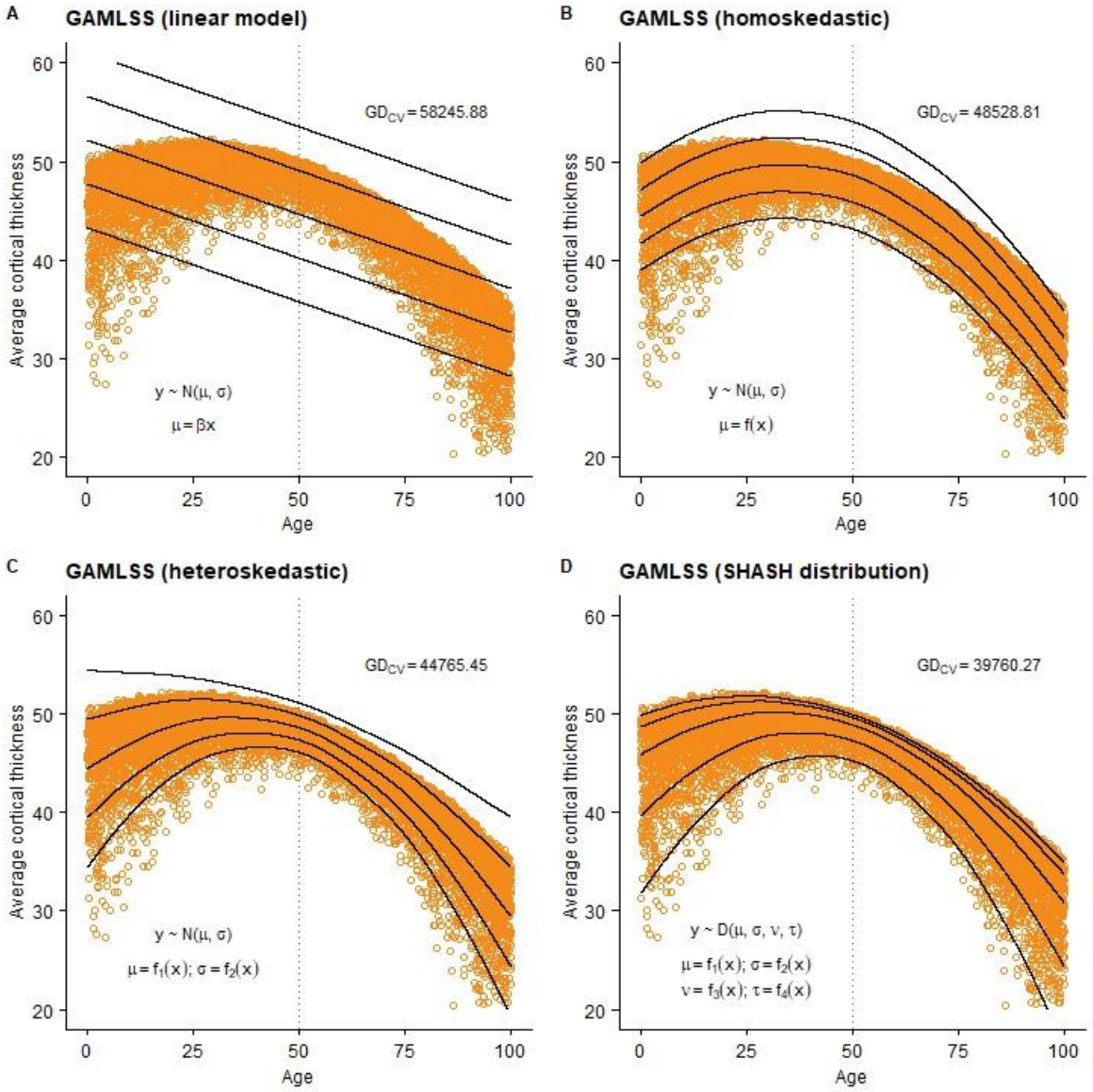


*Figure 9*. Simulated data representing the non-linear negative association between cortical thickness (*y* axis) and age (*x* axis) ($n$=1$e^4$) (adapted from Dinga et al., 2021, Figure 1). A: An LM assuming a normal distribution for the outcome variable, with a linear predictor for the location parameter. B: A GAM assuming a normal distribution for the outcome variable, with a nonlinear predictor for the location parameter. C: A GAM assuming a normal distribution for the outcome variable, with nonlinear predictors for both location and scale parameters. D: A GAMLSS using the third form of the four-parameter Sinh-Arcsinh (SHASH) distribution (Jones & Pewsey, 2009), with nonlinear predictors for location and scale while holding skewness and kurtosis constant. All models were fitted using GAMLSS (see Figure 6). The five curvy solid lines on each plot represent the estimated conditional quantiles of the response variable distribution at each predictor value. From bottom to

top, the lines indicate: q5 curve (5th percentile, lower outer bound), q25 curve (25th percentile, first quartile), q50 curve (median, 50th percentile), q75 curve (75th percentile, third quartile), and q95 curve (95th percentile, upper outer bound). The vertical dotted grey line marks the age of 50. $GD_{CV}$ = Global Deviance CV (Cross-Validation) ($GD_{CV}$ is a metric that reflects a model's predictive ability, as opposed to solely how well it fits the observed data. The model with the lowest $GD_{CV}$ value is generally preferred because it indicates better generalizability, making $GD_{CV}$ especially useful when comparing different models).

Figure 9 shows that both the traditional LM and the GAM approaches do not fit the data well, although the GAMs are an improvement over the LM. The GAMLSS SHASH model provides the best fit overall, both descriptively and predictively, as indicated by the $GD_{CV}$. The five quantile curves, representing specific percentiles of the predicted distribution, can be understood as prediction intervals that illustrate the probable range of individual observations at each predictor value. Thus, for example, the LM predicts that at the age of 50, the average cortical thickness is about 45 ‘mock-meters’, the other three models predict this thickness to be 49 ‘mock-meters’. The cortical thickness predictions for a 50-year-old at the 95th percentile vary across the different models. Both the LM and GAMLSS homoskedastic models predict a thickness of 54 'mock-meters'. This value falls outside the range of the observed data. In contrast, the GAMLSS heteroskedastic and GAMLSS SASHO models predict thicknesses of 51 and 50 'mock-meters', respectively. These values are within the range of the data. The GAMLSS SHASHO model is only a valid explanatory model if age is the only factor associated with cortical thickness and if cortical thickness has been measured with valid metrics given by the most suitable brain imaging tool. Otherwise, the SHASH GAMLSS model is best used as a descriptive and predictive model (the R code for this example can be found at https://cutt.ly/mrfSoIs4. The example above demonstrates how GAMLSS operates in a simple linear regression scenario. For a multiple regression case, and

to explore visual techniques such as worm plots (van Buuren & Fredriks, 2001) and bucket plots (de Bastiani et al., 2022), please visit https://cutt.ly/mrfSy8sm.)

GAMLSS is a highly flexible SL framework that enables the development of diverse models. Beyond its general utility, GAMLSS modeling has had significant real-world impact. For instance, it has been employed in constructing centile estimates for the World Health Organization Child Growth Standards (Borghi et al., 2006), which aim to establish population reference ranges across age and sex to better detect clinically atypical measurements throughout life. Additionally, GAMLSS has proven effective in modeling relationships between gray matter volume and age (Bethlehem et al., 2022). The framework has also been extended in various ways, including: BAMLSS (Umlauf, Klein, & Zeileis, 2018), a Bayesian variant of GAMLSS, NAMLSS (Thielmann et al., 2024) and semi-structured distributional regression (Rügamer, Kolb, & Klein, 2024), which integrate neural networks with GAMLSS, and hybrid approaches combining decision trees with GAMLSS (Constable et al., 2023). It is worth noting that artificial neural networks (ANNs) fundamentally operate as regression models. The estimated parameters (weights and biases) in ANNs function similarly to regression coefficients, mapping input variables (x) to outputs (y)—precisely the goal of regression. In fact, generalized additive models (GAMs), a sub-model in GAMLSS, can be viewed as a special case of ANNs (Cheng & Titterington, 1994). This connection is explicitly leveraged in modern statistical learning tools that model tabular and multimodal data within a regression framework (see Figure 1 in Rügamer et al., 2024).

Regression models are inherently interpretable (as noted by Lehmann, 2008). Since GAMLSS is fundamentally a regression model, it possesses a similar level of interpretability. The primary objective of GAMLSS as an SL approach is to facilitate statistical abductive learning—a process where the data analyst draws logical inferences aimed at finding the simplest and most probable conclusion from multiple (GAMLSS) models.

## Discussion

In this book chapter we argue that the replication crisis in science is fundamentally a testing crisis rooted in the misapplication of statistical frameworks rather than inherent flaws in significance testing itself. Through the TASI framework, we showed that researchers' bias toward confirming rather than falsifying hypotheses undermines the logical foundation of scientific inquiry, regardless of whether *p*-values, effect sizes, or other metrics are employed. Both REACT and GAMLSS offer compelling alternatives that address this fundamental issue by providing more nuanced frameworks for statistical analysis. REACT explicitly acknowledges uncertainty through its agnostic decision option, preventing premature conclusions when data is insufficient, while GAMLSS transcends binary hypothesis testing altogether by focusing on comprehensive modeling of distributions across multiple parameters. These approaches represent a paradigm shift from the reductive nature of significance testing toward statistical frameworks that better accommodate the complexity of real-world phenomena. Below, we offer further reflections on the connections between data science and artificial intelligence

As noted earlier, modern AI models—particularly LLMs—can now explain their reasoning in natural language, reducing the "black box" problem and improving transparency. A growing practice involves describing a statistical problem in plain language (e.g., "Create R code for robust linear regression with spline-modeled covariates") and having the LLM generate functional, well-commented code (see Figure 1 in Marmolejo-Ramos et al., 2023). This approach, sometimes called ‘vibe coding’, streamlines the coding process by translating intuitive prompts into executable scripts. Beyond one-off code generation, specialized platforms (e.g., julius.ai; and ‘data interpreter’, Hong et al., 2024) and ambitious projects (e.g., The Automatic Statistician, Steinruecken et al., 2019; Sakana AI’s AI Scientist, Lu et al., 2024) aim to automate data analysis and even scientific research. While these tools

accelerate processing, they remain just that—tools. The human analyst retains full control, deciding what data to analyze, how to combine datasets, and which statistical methods (AI-suggested or otherwise) to apply. In this way, LLM-powered tools act as collaborators, enhancing—not replacing—human expertise and judgment.

To conclude this chapter, we emphasize that the collaboration between AI-powered data analysis tools and human analysts offers significant benefits for teaching and learning data science. LLM-based statistical tools can play a key role in education by helping to develop course materials (e.g. R shiny apps via vibe coding), deliver personalized learning experiences, and clarify complex statistical concepts (Ellis & Slade, 2023). The ideal outcome of this human-AI partnership is well-informed model and hypothesis testing, where human expertise guides AI-generated insights to ensure rigorous, meaningful analysis.

## References

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, 508(7496), 312–313. https://doi.org/10.1038/508312a

Alves, M. V., Ekuni, R., Hermida, M. J., & Lisboa, J. V. (Eds.). (2022). Cognitive Sciences and Education in Non-WEIRD Populations: A Latin American Perspective. *Springer Nature*. DOI: https://doi.org/10.1007/978-3-031-06908-6

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of Clinical Psychiatry*, *82*(1), 25941. https://doi.org/10.4088/JCP.20f13804

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604). https://doi.org/10.1038/533452a

Baniecki, H., Parzych, D., & Biecek, P. (2023). The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-023-00924-w

Barch, D. M., & Yarkoni, T. (2013). Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research. *Cognitive, Affective & Behavioral Neuroscience*, *13*(4), 687–689. https://doi.org/10.3758/s13415-013-0201-7

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies,

opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Baumgartner, H., Alessandroni, N., Byers-Heinlein, K., Frank, M. C., Hamlin, K., Soderstrom, M., ... & Coles, N. A. (2023). How to build up big team science: A practical guide for large-scale collaborations. *PsyArXiv*. https://doi.org/10.31234/osf.io/j7mt4

Benjamin DJ, Berger JO, Johnson VE et al (2018) Redefine statistical significance. *Nat Hum Behav 2*:6–10 https://doi.org/10.1038/s41562-017-0189-z

Berg, N., 2004: No-decision classification: an alternative to testing for statistical significance. the Journal of socio-Economics, 33, no. 5, 631–65. https://doi.org/10.1016/j.socec.2004.09.036

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., ... Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), 525–533. https://doi.org/10.1038/s41586-022-04554-y

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2108.0725

Borghi, E., de Onis, M., Garza, C., Van den Broeck, J., Frongillo, E.A., Grummer-Strawn, L., Van Buuren, S., Pan, H., Molinari, L., Martorell, R., Onyango, A.W., Martines, J.C., & (2006). Construction of the World Health Organization child growth standards:

selection of methods for attained growth curves. *Statistics in Medicine*, *25*, 247-265. https://doi.org/10.1002/sim.2227

Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199-231. https://doi.org/10.1214/ss/1009213726

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bunge, M. (2018). La ciencia: su método y su filosofía (Vol. 1). Laetoli.

Catanzaro, M. (2023, November 27). *Saudi universities lose highly cited researchers after payment schemes raise ethics concerns. Science.* https://doi.org/10.1126/science.zhs1429

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, *9*(1), 2–54. https://doi.org/10.1214/ss/117701063

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145-153. https://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Psychology Press. https://doi.org/10.4324/9780203771587

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L., ... & Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the many smiles collaboration. *Nature Human Behaviour*, 1-12. https://doi.org/10.1038/s41562-022-01458-9

1041 Constable, P. A., Loh, L., Prem-Senthil, M., & Marmolejo-Ramos, F. (2023). Visual search
1042 and childhood vision impairment: A GAMLSS-oriented multiverse analysis approach.
1043 *Attention, Perception, & Psychophysics, 85*(3), 968–977.
1044 https://doi.org/10.3758/s13414-023-02670-z

1045 Coscrato, V., R. Izbicki, and R. B. Stern, 2019: Agnostic tests can control the type I
1046 and type II errors simultaneously. *Brazilian Journal of Probability and Statistics* 34.2
1047 (2020): 230-250. https://doi.org/10.1214/19-BJPS431

1048 Coyle, J. R., Hejazi, N. S., Malenica, I., Phillips, R. V., Arnold, B. F., Mertens, A., Benjamin-
1049 Chung, J., Cai, W., Dayal, S., Colford, J. M., Hubbard, A. E., & van der Laan, M. J.
1050 (2023). Targeted learning. In *Wiley StatsRef: Statistics Reference Online*. John Wiley
1051 & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat08414

1052 Culp, M. (2011). On Propagated Scoring for Semisupervised Additive Models. *Journal of the*
1053 *American Statistical Association, 106*(493), 248–259.
1054 https://doi.org/10.1198/jasa.2011.tm09316

1055 De Bastiani, F., Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., & Silva, L. A. (2022).
1056 Bucket plot: A visual tool for skewness and kurtosis comparisons. *Brazilian Journal*
1057 *of Probability and Statistics, 36*(3), 421-440. https://doi.org/10.1214/22-BJPS533

1058 Di Maso, M., Talamini, R., Bosetti, C., Montella, M., Zucchetto, A., Libra, M., Negri,
1059 E., Levi, F., La Vecchia, C., Franceschi, S., Serraino, D., & Polesel, J. (2013). Red
1060 meat and cancer risk in a network of case-control studies focusing on cooking
1061 practices. Annals of oncology : official journal of the European Society for Medical
1062 Oncology, 24(12), 3107–3112. https://doi.org/10.1093/annonc/mdt392

Dinga, R., Fraza, C. J., Bayer, J. M. M., Kia, S. M., Beckmann, C. F., & Marquand, A. F. (2021). Normative modeling of neuroimaging data using generalized additive models of location scale and shape. *bioRxiv*. https://doi.org/10.1101/2021.06.14.448106

Doyle, A. C. (1927). *The case-book of Sherlock Holmes*. John Murray.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B, 57*, 45-97. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242. https://doi.org/10.1037/h0044139

Ellis, A. R., & Slade, E. (2023). A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education. *Journal of Statistics and Data Science Education, 31*(2), 128-133. https://doi.org/10.1080/26939169.2023.2223609

Esteves, L. G., Izbicki, R., Stern, J.M., Stern, R.B. (2019). Pragmatic hypotheses in the evolution of science. Entropy. 2019 Sep 11;21(9):883. https://doi.org/10.3390/e21090883

Esteves, L. G., R. Izbicki, J. M. Stern, and R. B. Stern (2016): The logical consistency of simultaneous agnostic hypothesis tests. Entropy, 18, no. 7, 256. https://doi.org/10.3390/e18070256

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *PNAS, 115*, 2628-2631. https://doi.org/10.1073/pnas.1708272114

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191. https://doi.org/10.3758/BF03193146

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association, JMLA, 105*(2), 203–206. https://doi.org/10.5195/jmla.2017.88

Frias-Navarro, D., Pascual-Llobell, J., Pascual-Soler, M., Perezgonzalez, J., & Berrios-Riquelme, J. (2020). Replication crisis or an opportunity to improve scientific production? *European Journal of Education, 55*, 618-631. https://doi.org/10.1111/ejed.12417

Friedrich, S., & Friede, T. (2023). On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal, 00*, e2200212. https://doi.org/10.1002/bimj.202200212

Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Ickstadt, K., Kestler, H. A., Lederer, J., Leitgöb, H., Pauly, M., Steland, A., Wilhelm, A., & Friede, T. (2022). Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification, 16*(3), 823-846. https://doi.org/10.1007/s11634-021-00455-6

Grimes, D. R. (2019). *The irrational ape*. Simon & Schuster.

Gundersen, O. E., Gil, Y., & Aha, D. W. (2018). On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI Magazine*, *39*(3), 56-68. https://doi.org/10.1609/aimag.v39i3.2816

Haibe-Kains, B., Adam, G.A., Hosny, A. et al. Transparency and reproducibility in artificial intelligence. Nature 586, E14–E16 (2020). https://doi.org/10.1038/s41586-020-2766-y

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Hensel, P. G. (2021). Reproducibility and replicability crisis: How management compares to psychology and economics–A systematic review of literature. *European Management Journal*, *39*(5), 577-594. https://doi.org/10.1016/j.emj.2021.01.002

Hernández Sampieri, R., Fernandez Collado C., & Baptista Lucio, P. (2014). *Metodología de la Investigación*. México: McGraw-Hill.

Hoang, A.-T., & Dickhaus, T. (2022). Combining independent p-values in replicability analysis: A comparative study. *Journal of Statistical Computation and Simulation*, *92* (10), 2184–2204. https://doi.org/10.1080/00949655.2021.2022678

Hong, S., Lin, Y., Liu, B., Liu, B., Wu, B., Zhang, C., Wei, C., Li, D., Chen, J., Zhang, J., Wang, J., Zhang, L., Zhang, L., Yang, M., Zhuge, M., Guo, T., Zhou, T., Tao, W., Tang, X., Lu, X., Zheng, X., Liang, X., Fei, Y., Cheng, Y., Gou, Z., Xu, Z., & Wu, C.

(2024). Data Interpreter: An LLM Agent For Data Science. arXiv. https://arxiv.org/pdf/2402.18679

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8). https://doi.org/10.1371/journal.pmed.0020124

Izbicki, R., Cabezas, L. M. C., Colugnatti, F. A. B., Lassance, R. F. L., de Souza, A. A. L., & Stern, R. B. (in press). REACT to NHST: Sensible conclusions for meaningful hypotheses. *The Quantitative Methods for Psychology*.

REACT to NHST: Sensible conclusions for meaningful hypotheses

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R (2nd ed.)*. Springer.

Jeffreys, H. (1961). *Theory of probability (3rd ed.).* Clarendon Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23*, 524–532. https://doi.org/10.1177/ 0956797611430953

Jones, M. C., & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika, 96* (4), 761–780, https://doi.org/10.1093/biomet/asp053

Kaplan, D. (2004). *The Sage handbook of quantitative methodology for the social sciences*. Sage.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9). https://doi.org/10.1016/j.patter.2023.100804

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kogani, M., Almasi, S. A., Ansari-Mogaddam, A., Dalvand, S., Okati-Aliabad, H., Tabatabaee, S. M., & Almasi, S. Z. (2020). Relationship between using cell phone and the risk of accident with motor vehicles: An analytical cross-sectional study. *Chinese journal of traumatology = Zhonghua chuang shang za zhi*, 23(6), 319–323. https://doi.org/10.1016/j.cjtee.2020.08.002

Kruschke, J. K. (2015). *Doing Bayesian data analysis. A tutorial with R, JAGS, and Stan (2nd ed.)*. Academic Press.

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. Behavior Research Methods, 53(6), 2528-2543. https://doi.org/10.3758/s13428-021-01546-0

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet policy review*, 9(2), 1-16. https://doi.org/10.14763/2020.2.1469

Lassance RF, Izbicki R, Stern RB. (2025) Adding imprecision to hypotheses: A Bayesian framework for testing practical significance in nonparametric settings. *International Journal of Approximate Reasoning*. 2025 Mar 1;178:109332. https://doi.org/10.1016/j.ijar.2024.109332

Lehmann, E. L. (2008). On the history and use of some standard statistical models. *IMS Collections, Probability and Statistics: Essays in Honor of David A. Freeman, 2*, 114-126. http://doi.org/10.1214/193940307000000419

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). Deepseek-v3 technical report. arXiv preprint. https://arxiv.org/abs/2412.19437

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv. https://arxiv.org/abs/2408.06292

Luijken K., Lohmann A., Alter U., Claramunt Gonzalez J., Clouth F. J., Fossum J. L., Hesen L., Huizing A. H. J., Ketelaar J., Montoya A. K., Nab L., Nijman R. C. C., Penning de Vries B. B. L., Tibbe T. D., Wang Y. A. and Groenwold R. H. H. (2024) *Replicability of simulation studies for the investigation of statistical methods: the RepliSims projectR.* Soc. Open Sci.11231003. http://doi.org/10.1098/rsos.231003

Marmolejo-Ramos, F., Simon, T., & Abadia, R. (2023, August 28). The power of large language models to augment human learning. *Digital Dexterity*. https://digitaldexterity.edublogs.org/2023/08/28/the-power-of-large-language-models-to-augment-human-learning/

Marmolejo-Ramos, F., Tejo, M., Brabec, M., Kuzilek, J., Joksimovic, S., Kovanovic, V., González, J., Kneib, T., Bühlmann, P., Kook, L., Briseño-Sánchez, G., & Ospina, R. (2023). Distributional regression modelling via GAMLSS. An overview through a data set from learning analytics. *WIRES Data Mining and Knowledge Discovery, 13* (1), e1479. https://doi.org/10.1002/widm.1479

Matos, E., Vilensky, M., Boffetta, P., & Kogevinas, M. (1998). Lung cancer and smoking: a case-control study in Buenos Aires, Argentina. *Lung cancer (Amsterdam, Netherlands), 21*(3), 155–163. https://doi.org/10.1016/s0169-5002(98)00055-5

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.

Meho, L. I., & Akl, E. A. (2025). Using bibliometrics to detect questionable authorship and affiliation practices and their impact on global research metrics: A case study of 14 universities. *Quantitative Science Studies*, 1-36. DOI: https://doi.org/10.1162/qss_a_00339

Metropolis, N. Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1091. https://doi.org/10.1063/1.1699114

Min, J., Song, X., Zheng, S., King, C. B., Deng, X., & Hong, Y. (2024). *Applied statistics in the era of artificial intelligence: A review and vision*. https://arxiv.org/abs/2412.10331

Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., ... & Ziv-Crispel, N. (2022). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review*, *29*(2), 613-626. https://doi.org/10.3758/s13423-021-01927-8

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika*, *20A*, 175–240. https://doi.org/10.2307/2331945

Neyman, J. (1935). On the problem of confidence intervals. *The Annals of Mathematical Statistics, 6,* 111–116. https://doi.org/10.1214/aoms/1177732585

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society, 105,* 292–327. https://doi.org/10.2307/2980436

Neyman, J. (1967). R. A. Fisher (1890-1962): an appreciation. *Science, 156,* 1456–1460. https://doi.org/10.1126/science.156.3781.1456

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. Communications in Statistics - Theory and Methods, 5(8), 737–751. https://doi.org/10.1080/03610927608827392

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), https://doi.org/10.1126/science.aac4716

O'Grady, C. (2024, October 21). Springer Nature retracts 75 papers connected to Spanish university head. Science. https://doi.org/10.1126/science.zvmx7id

Pawel, S., Kook, L., & Reeve, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, 2200091. https://doi.org/10.1002/bimj.202200091

Perezgonzalez, J. D. (2015). Confidence intervals and tests are two sides of the same research question, *Frontiers in Psychology 28*(6). https://doi.org/10.3389/fpsyg.2015.00034

Piccolo, L. R., Merz, E. C., He, X., Sowell, E. R., Noble, K. G., & Pediatric Imaging, Neurocognition, Genetics Study. (2016). Age-related differences in cortical thickness vary by socioeconomic status. *PLOS ONE, 11*(9), e0162511. https://doi.org/10.1371/journal.pone.0162511

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... & Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115-126. https://doi.org/10.1038/nrn.2016.167

Popper, k. (1962). *Conjectures and refutations: The growth of scientific knowledge*. Basic Books.

Pownall M, Pennington CR, Norris E, et al. Evaluating the Pedagogical Effectiveness of Study Preregistration in the Undergraduate Dissertation. *Advances in Methods and Practices in Psychological Science*. 2023;6(4). https://doi.org/10.1177/25152459231202724

Prieto, D. (2022). Ciencia Abierta: desafíos y oportunidades para Uruguay y el Sur Global. *Information, 27*(1), 253-283. https://doi.org/10.35643/Info.27.1.5

Rigby, R. A., & Stasinopoulos, D. M., Heller, G., & de Bastiani, F. (2020). *Distributions for Modeling Location, Scale, and Shape*. Taylor & Francis.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics, 54*, 507-554.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, *3*(3), 199-217. https://doi.org/10.1038/s42256-021-00307-0

Rubin, M. (2022). The costs of HARKing. *British Journal for the Philosophy of Science*, *73*(2), 535-560. https://doi.org/10.1093/bjps/axz050

Rugamer, D., Kolb, C., & Klein, N. (2024). Semi-structured distributional regression. *The American Statistician, 78* (1), 88-99. https://doi.org/10.1080/00031305.2022.2164054

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S. R., Busa, E., Morris, J. C., Dale, A. M., & Fischl, B. (2004). Thinning of the cerebral cortex in aging. *Cerebral Cortex, 14*(7), 721–730. https://doi.org/10.1093/cercor/bhh032

Sapkota, R., Raza, S., & Karkee, M. (2025). Comprehensive analysis of transparency and accessibility of chatgpt, deepseek, and other sota large language models. arXiv preprint. https://arxiv.org/abs/2502.18505

Sasaki, K. & Yamada, Y. (2023). SPARKing: Sample-size planning after the results are known. *Front. Hum. Neurosci. 17*:912338. https://doi.org/10.3389/fnhum.2023.912338

Schuirmann, D. J., 1987: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of

pharmacokinetics and biopharmaceutics, 15, 657–680. https://doi.org/10.1007/bf01068419

Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. The MIT Press.

Shmueli, G., & Tafti, A. (2023). How to "improve" prediction using behavior modification. *International Journal of Forecasting, 39*(2), 541-555. https://doi.org/10.1016/j.ijforecast.2022.07.008

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman and Hall/CRC.

Stasinopoulos, M. D., Kneib, T., Klein, N., Mayr, A., & Heller, G. Z. (2024). *Generalized Additive Models for Location, Scale and Shape: A Distributional Regression Approach, with Applications*. Cambridge University Press.

Stasinopoulos, M. D., Rigby, R. A., & de Bastiani, F. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling, 18* (3-4), 248-273. https://doi.org/10.1177/1471082X18759144

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science, 11(5), 702-712. https://doi.org/10.1177/1745691616658637

Steinruecken, C., Smith, E., Janz, D., Lloyd, J., & Ghahramani, Z. (2019). The Automatic Statistician. In Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.), *Automated Machine*

*Learning. The Springer Series on Challenges in Machine Learning* (pp. 161-173). Springer. https://doi.org/10.1007/978-3-030-05318-5_9

Stodden, V., Leisch, F., & Peng, R. D. (Eds.). (2014). *Implementing reproducible research*. CRC Press.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., & Williams, R.M. (1949). *The American Soldier: Adjustment during Army Life, Vol. 1.* Princeton University Press, Princeton, NJ.

Taleb, N. N. (2001). *Fooled by randomness*. Random House.

Taleb, N. N. (2007). *The black swan. The impact of the highly improbable*. Random House.

Thielmann, A. F., Kruse, R.-M., Kneib, T., & Säfken, B. (2024). Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean. In S. Dasgupta, S. Mandt, and Y. Li (Eds.), *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (Vol. 238*). PMLR. https://proceedings.mlr.press/v238/frederik-thielmann24a.html

Toribio-Flórez, D., Anneser, L., de Oliveira-Lopes, F., Pallandt, M., Tunn, I., Windel, H., & Max Planck PhDnet Open Science Group (2021). Where Do Early Career Researchers Stand on Open Science Practices? A Survey Within the Max Planck Society. *Frontiers in Research Metrics and Analytics, 5*, 586992. https://doi.org/10.3389/frma.2020.586992

Trafimow, D. (2019). A taxonomy of model assumptions on which ‘p’ is based and implications for added benefit in the sciences. *International Journal of Social*

*Research Methodology*, *22*(6), 571-583. https://doi.org/10.1080/13645579.2019.1610592

Tu, J., Liu, W., & Mao, X. (2024). Distributed Estimation on Semi-Supervised Generalized Linear Model. *Journal of Machine Learning Research*, *25*, 1-41. http://jmlr.org/papers/v25/22-0670.html

Umlauf, N., Klein, N., & Zeileis, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, *27*(3), 612–627. https://doi.org/10.1080/10618600.2017.1407325

van Buuren, S., & Fredriks, M. (2001). Worm plot: simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, *20*, 1259-1277. https://doi.org/10.1002/sim.746

Van der Laan, M. J., & Starmans, R. J. C. M. (2014). Entering the era of data science: Targeted learning and the integration of statistics and computational data analysis. *Advances in Statistics*, 502678. https://doi.org/10.1155/2014/502678

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *25*(3), 337-372. https://doi.org/10.1142/S0218001411008683

Vidal-Piñeiro, D., Parker, N., Shin, J., et al. (2020). Cellular correlates of cortical thinning throughout the lifespan. *Scientific Reports*, *10*, 21803. https://doi.org/10.1038/s41598-020-78471-3

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C.,

Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wasserstein, R. & Lazar, N. (2016) The ASA Statement on p-Values: Context, Process, and Purpose, *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Westlin, C., Theriault, J. E., Katsumi, Y., Nieto-Castanon, A., Kucyi, A., Ruf, S. F., Brown, S. M., Pavel, M., Erdogmus, D., Brooks, D. H., Quigley, K. S., Whitfield-Gabrieli, S., & Barrett, L. F. (2023). Improving the study of brain-behavior relationships by revisiting basic assumptions. *Trends in Cognitive Sciences, 27*(3), 246–257. https://doi.org/10.1016/j.tics.2022.12.015

Yarkoni, T. (2022) The generalizability crisis. *Behavioral and Brain Sciences 45*, e1: 1–78. https://doi.org/10.1017/ S0140525X20001685

Zhao, J., & Liu, N. (2021). A safe semi-supervised classification algorithm using multiple classifiers ensemble. *Neural Processing Letters, 53*(4), 2603–2616. https://doi.org/10.1007/s11063-020-10191-1