Towards Transparency and Open Science

A Principled Perspective on Computational Reproducibility and Preregistration

Abstract

TBD

Contents

Introduction

Psychology is a difficult science (Meehl, 1978). Though there might be some disagreement on why exactly this is the case, I doubt there is disagreement about the fact itself. Laying open the difficulties and attempting to overcome them is not a recent trend, though it has been invigorated by the so-called replication crisis in psychology (Ioannidis, 2005; Open Science Collaboration, 2015) that begins to ripple through other empirical sciences. As psychology grapples with this crisis of confidence in its empirical results, several causes and remedies have been proposed. We can roughly divide the causes into two categories: misuse of statistical methods (e.g., Bakan, 1966; Benjamin et al., 2018; Cohen, 1994; Gigerenzer, 2004; Wagenmakers et al., 2011) and sociological factors (Bakker et al., 2012; John et al., 2012; Rosenthal, 1979; e.g., Simmons et al., 2011).

In my view, both categories emerge from a lack of transparency about the inductive process in the empirical test of a theory. Testing a theory empirically is often viewed as deductive, since the theory is making statements about the data. Empirical scientists, however, often simultaneously engage in induction by deriving general statements from data. The inductive element in the empirical test leads to overconfidence in the empirical results, if unaccounted. An accurate judgment of the empirical support of a theory is only possible if the inductive process is made transparent.

The above distinction arises because some parts of the inductive process are well defined in the form of statistical methods, while other parts happen more informally. Statistical methods make induction quantifiable; while open science measures reduce uncertainty about sociological factors in induction. To understand why transparency is crucial, we must understand that these two forms of induction are integral to empirical sciences. Any science must be able to communicate how it generates its knowledge. However, transparency has an outstanding role in psychology and other empirical sciences, because empirical statements lose their value without transparency about the inductive process that contributed to them. Therefore, transparency is more than a virtue that somewhat improves empirical sciences but is an indispensable property.

To function as an empirical science, psychology must be able to make statements about the world that can be compared to the true conditions. In psychology, this is not a purely deductive endeavor. Very few psychological theories are precise enough to derive testable statements. While it is tempting to claim that a theory makes deductively testable claims by, e.g., implying a mean difference between two groups (Lee & Pawitan, 2021), such a statement is not testable on its own. Consider a simple t-test of mean differences. The decision to reject the null hypothesis still depends, besides true mean difference and sample size, on the observed variance. However, the variance is, of course, estimated from the data. So the threshold of the deductive decision depends on a quantity that must be induced.

Induction is necessary and, in the present case, innocuous for psychology as a science. Here, it is innocuous because the introduced bias can be accounted for

and vanishes with increasing sample size, but we will see later that this is not always the case. In this specific example, we use a t-test instead of a z-test to account for variance estimation from data. It is widely known that even without this correction, the z-test is a good approximation when sample sizes are large (Student, 1908). Induction is also necessary because it is almost unimaginable to ask psychological researchers to specify every detail, such as the variance, a priori from their theory. If they had to, there would probably be no psychological theory that could survive an empirical test.

In other words, induction gives our theories some slack to be imprecise and contain "blank" spaces, later to be filled with data. It allows researchers to concentrate on the essential statements of their theories and choose some of the assumptions to fit the data well. In some ways, it is the empirical researchers' answer to the Duhem–Quine problem (Duhem, 1976; van Orman Quine, 1976), which states that any empirical test of a theory is testing the conjunction of theory, auxiliary assumptions, and conjectures (Meehl, 1990, 1978). By inducing some quantities, psychological researchers can remove them from the conjunction. If researchers use induction for some necessary but under the theory arbitrary assumptions, their theory will not be refuted because of these assumptions. Since empirical researchers often cannot derive every assumption from their theory, avoiding refutation because of those assumptions is a desirable property.

By the same token, whole theories may escape refutation by replacing every ill-fitting statement deduced from theory with statements induced from data. Such a strategy of post hoc changing a theory in light of facts has been called "Lakatosian Defense" (Meehl, 1990). If pushed to the limit, we arrive at a "theory" governed by the data. Such a theory, full of empirically induced statements, is almost empty of statements that have been empirically verified. The data used for induction can not refute these statements, so they are not tested at all.

So what to think of such, yet untested, theory? Researchers and philosophers of science differ considerably in their opinion about how to appraise theories, e.g., judging the long-term performance (if they are frequentists), degrees of belief (if they are Bayesians), or probativeness (if they are severe testers, Mayo, 2018, p. 14) of a hypothesis. Whatever measure they subscribe to, an untested theory should not score very high.

So empirical researchers find themselves in a pinch. On the one hand, they need induction to test their imprecise theories. On the other hand, induction may render any test of a hypothesis ineffective. The problem, I argue, therefore, is not induction but making induction transparent. The replication crisis can be traced to a misjudgment of how much induction has been going on in psychology and hence, how well-tested the empirical claims are. Therefore, the question of this dissertation is what must be made transparent, and how can we make it transparent?

The first question (the "what") is theoretical in nature. It is addressed in the thesis itself, which supplies the theoretical framework of the articles written as part of the dissertation. Under this framework, induction is split into a process that can be formally analyzed (statistical methods) and a part that is much more difficult

to judge (sociological factors).

Based on this split, the articles answer the second question (the "how"). I argue that transparency about statistical methods is enabled by computational reproducibility, while transparency about sociological factors is facilitated by preregistration. The conceptualization of computational reproducibility and preregistration as means for transparency is supplemented by practical guidance on how researchers may implement these tools.

What necessitates transparency

The need for transparency is closely tied to the use of induction in the empirical test of a theory. There has been a long and vigorous debate about what it means to test a theory empirically. I do not attempt to rehash the debate about what constitutes an empirical test but aim to lay open the role of transparency in two frameworks that lend themselves to investigate induction. The first framework motivates transparency when the aim of an empirical test is to evaluate the verisimilitude ("truth likeness") of a theory. The second framework motivates transparency under a science that wants to select a theory according to its expected predictive performance.

Both frameworks show how unaccounted induction leads to overconfidence in empirical results and give us some theoretical tools to asses and control this overconfidence. Because both are quite technical, they are followed by a more conceptual summary of these tools. These sections provide the basis to understand how computational reproducibility and preregistration enable a proper assessment of an empirical test on a conceptual level.

An information-theoretic perspective

Information theory provides a rigorous mathematical measure that can be understood as the verisimilitude of a theory. That is, how much information about the truth is lost when the theory is used to model reality. Expressed mathematically, we have a function $f(x)$ that gives us the liklihood to observe the state of the world x where f represents the ground truth. We are now interested in how much information is lost if we use $g(x)$, our theory, instead of $f(x)$, the reality, over all possible states $\mathcal{X}$. Expressed as lost bits of information, a measure known as Kullback–Leibler divergence (Kullback & Leibler, 1951), we get:

$$\mathcal{L}_{KL}(f, g) = \int_{\mathcal{X}} f(x) \log\left(f(x)\right) \mathrm{d}x - \int_{\mathcal{X}} f(x) \log\left(g(x)\right) \mathrm{d}x$$

$$\mathcal{L}_{KL}(f, g) = \mathbb{E}_{X\sim f}\left[\log(f(x))\right] - \mathbb{E}_{X\sim f}\left[\log(g(x))\right]$$

Most readers will recognize that this information-theoretic setup and the derivation below follow closely Burnham et al. (2002), Ch. 7.2, in their derivation of the Akaike Information Criterion in its general form. What is of interest here is

not the derivation but how this conceptualization can help us to understand what happens when data is simultaneously used to induce quantities of a theory and test the theory.

This setup is, of course, highly theoretical. $\mathcal{L}_{KL}(f, g)$ is unknowable, since the truth is unobserved. However, this fact does not impede us from getting closer to the truth because we can still compare two theories relative to each other. Because the expectation for f remains constant (left-hand expectation), we only need to estimate the relative expected loss of information (right-hand expectation) to make a comparative judgment. To make a relative judgment about several competing theories it suffices to estimate:

$$\mathbb{E}_{X\sim f}\left[\log(g(x))\right]$$

To allow for quantities to be induced, we must assume that our theory is parameterized, e.g., $g(x|\theta)$. That means our theory implies a multitude of possible probability distributions that may describe reality. This parameterization captures the idea that some assumptions necessary for a theory to make testable statements are arbitrary. Of those arbitrary assumptions, we want to find those that fit the reality with the least amount of information lost. The best parameterization is achieved by:

$$\theta_* = \arg\min_{\theta} \mathcal{L}_{KL}(f, g(\cdot|\theta))$$

The inference goal is, therefore:

$$\mathbb{E}_{X\sim f}\left[\log(g(x|\theta_*))\right]$$

Of course, we usually do not know θ_*. That is why it is necessary to induce it from data, denoted as $\widehat{\theta}(y)$, where Y is independently sampled from $X \sim f$. The crucial point here is to understand what happens when we can not derive θ deductively but must substitute it inductively with an estimate $\widehat{\theta}(y)$. Any estimated parameters $\widehat{\theta}(y)$ would almost surely not be equal to θ_* (assuming θ may take an infinite number of values, i.e., is continues). It follows that almost surely:

$$\mathcal{L}_{KL}(f, g(\cdot|\widehat{\theta}(y)) > \mathcal{L}_{KL}(f, g(\cdot|\theta_*))$$

Or ignoring the constant:

$$\mathbb{E}_{X\sim f}\left[\log(g(x|\theta_*))\right] > \mathbb{E}_{X\sim f}[\log(g(x|\widehat{\theta}(y)))]$$

That is to say, any induced estimate will be sub-optimal. The inference goal, however, is to compare the theory g to reality f, not to evaluate the estimates of $\widehat{\theta}$. The point is to make a statement about the theory, not to make a statement about the data in combination with the theory. If the estimate of $\widehat{\theta}(y)$, i.e., the inductive

process, is unbiased in the sense that it converges towards θ_*, we may form an expectation over the data Y:

$$\mathbb{E}_{Y\sim f^n}\mathbb{E}_{X\sim f}[log(g(x|\hat{\theta}(y)))]$$

Forming this expectation over data is a crucial step; it requires us to think beyond the data we observed to all the data we could have observed. There are two ways to get at this expectation. One is the use of Taylor series expansion, which follows in this section, and another is the use of cross-validation discussed in the next section.

We might notice that we often get an unbiased estimate for the observed distance on the data used to induce $\hat{\theta}$. With slight abuse of notation, $log(g(y|\theta)) \equiv \sum_{i=1}^{n} log(g(y_i|\theta))$, so that $\mathbb{E}_{y\sim f^n} log(g(y|\hat{\theta}(y)))$ refers to the observed likelihood. The observed likelihood is often easy enough to obtain, e.g. in maximum likelihood estimation.

The expectation over the data together with Taylor series expansion yields:

$$\mathbb{E}_{Y\sim f^n}\mathbb{E}_{X\sim f}[log(g(x|\hat{\theta}(y)))] \approx \mathbb{E}_{Y\sim f^n}[\log(g(y|\hat{\theta}(y)))] - tr[J(\theta_*)I(\theta_*)^{-1}]$$

Where J is the Fisher information matrix with regard to g, and I for f respectively. For more details see of this derivation see Burnham et al. (2002), Ch. 7.2.

The observed likelihood $log(g(x|\hat{\theta}(y)))$ is, therefore, a biased estimate of the distance to the truth. Note that $tr[J(\theta_*)I(\theta_*)^{-1}]$ is usually negative definitive and, therefore, we may conclude that substituting deduced quantities by induced estimates leads to some overconfidence about how close one is to the truth. This overconfidence is directly related to how much induction a model entails. This bias is often called the complexity or capacity of a model, i.e., how much the data is influencing the results and, hence, how detailed the model is representing the data (Goodfellow et al., 2016b, 2016a). I therefore denote it as as $\mathcal{C}$.

$$\mathbb{E}_{X\sim f}\mathbb{E}_{Y\sim f^n}[log(g(x|\hat{\theta}(y)))] \approx \mathbb{E}_{Y\sim f^n}[\log(g(y|\hat{\theta}(y)))] + \mathcal{C}$$

If we want to induce quantities and correctly appraise a theory on the same data, we must know how much we have to correct our appraisal for the induction. Fortunately, it is possible to approximate the complexity of a model under some conditions. One condition is, since θ_* is unknown, that we know the properties of the inductive process that generated $\hat{\theta}$. We then can formally analyze the behavior and derive a mathematical expression for $\mathcal{C}$, e.g., corrections for a large class of statistical models, most famously the class of linear models, are available.

A future performance perspective

In addition to closeness to truth, there is another line of argumentation about why transparency about the process of induction is important. Instead of

verisimilitude, one might be concerned with future performance. That is, how well does a theory do in predicting novel facts. Please note that the information-theoretic setup above has made no appeal to the expected performance on unseen data. Verisimilitude and expected performance are different motivations for transparency, though they can be linked. In the future performance setup, we do not appeal to ground truth, replace the Kullback–Leibler divergence with an arbitrary loss function, and no longer require $g(x)$ to return a likelihood:

$$\mathcal{L}(x, g(x|\theta)) = \mathbb{E}_x L(x, g(x|\theta))$$

Again we can define the loss we have observed in the sample used to estimate $\hat{\theta}$ (Soch et al., 2020):

$$\mathbb{E}_y L(y, g(y|\hat{\theta}(y))) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(y_i|\theta))$$

However, what we are interested in is not how well the theory did on data that informed it but on future, yet unseen, data:

$$\mathbb{E}_x \mathbb{E}_y L(x, g(x|\hat{\theta}(y)))$$

This expectation over, what is often called training and test data, is termed generalization error or expected prediction error (Bengio & Grandvalet, 2004) and resembles the expectation over data discussed in the information-theoretic setup (Stone, 1977). Instead of using the Taylor series expansion, we can repeatedly sample data and repeat the inductive process. That is, we make use of cross-validation where the data is partitioned, and the inductive process is repeated on all permutations of a subset of the partitions. For each subset, the resulting model is then compared to the complement that was not used for induction, which is indicated by $y_{-i} = y \setminus \{y_i\}$.

$$\mathbb{E}_x \mathbb{E}_y L(x, g(x|\hat{\theta}(y))) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(y_i|\hat{\theta}(y_{-i})))$$

As stated earlier, using cross-validation it is possible to estimate $\mathbb{E}_{X\sim f}\mathbb{E}_{Y\sim f^n}[log(g(x|\hat{\theta}(y)))]$ as well; which connects the information-theoretic setup with this approach (Stone, 1974, 1977). To make the link to the first approach even more clear, we may define complexity in these terms (Hauenstein et al., 2018):

$$\mathcal{C} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(y_i|\hat{\theta}(y_{-i}))) - L(y_i, g(y_i|\hat{\theta}))$$

Instead of a formal analysis to derive $\mathcal{C}$, we can simply repeat the inductive process, i.e., using cross-validation. That drastically expands the set of inductive

processes for which we can estimate the inductive bias since it is considerably easier to repeat the process than to derive the complexity analytically.

A conceptual perspective

Now that we have established the need for transparency about the inductive process, we can drop a few of the more technical details to get a more straightforward about what we have to make transparent. It bears repeating that simply laying open what has been done is not enough. Showing the inductive results instead of the process is insufficient to appraise the theory. On a conceptual level, we want to compare:

$$\mathcal{L}(\text{Theory}, \text{Reality})$$

But to allow for induction to happen, we replace theory with a model (not necessarily a statistical one) or, put differently, a multitude of implications about the data from the theory.

$$\mathcal{L}(\text{Model(Reality)}, \text{Reality})$$

The idea is that we choose from the multitude the version of our theory that best fits reality. However, we are forced to rely on a limited sample of reality. We are misled because these two factors, induction and limited sample size, interact.

$$\mathcal{L}(\text{Model(Reality)}, \text{Reality}) > \mathcal{L}(\text{Model(Sample)}, \text{Sample})$$

$$\mathcal{L}(\text{Model(Reality)}, \text{Reality}) \approx \mathcal{L}(\text{Model(Sample)}, \text{Sample}) + \mathcal{C}$$

Transparency is necessary because induction leaves researchers overly optimistic regarding their theories' fit to the data. The extent of this optimism depends on the inductive process, not merely its results. Specifically, it depends on the complexity, i.e., the ability of the inductive process to adept to data. Without knowing the inductive process, researchers can not judge the overconfidence, so the inductive process ought to be made transparent.

How to establish transparency

The above sections aimed to motivate the observation that the apparent fit of a theory to data is often overly optimistic, if it has inductive elements. This observation is only of limited use if we do not now the extend of this optimism. However, both setups show that the extent of the optimism ($\mathcal{C}$) is closely related to the inductive process and give us two starting points for making this bias transparent. The first requires a formal analysis of the inductive process to compute the complexity (e.g., using information criteria), and the second merely requires

that the process is repeatable (e.g., using cross-validation). Both approaches require researchers to make the process of induction transparent instead of merely publishing the results.

Even a casual consideration of the above formalization should strike anyone who has ever worked with empirical data as unrealistic. No one can actually expect researchers to be inductive only in ways that are formally analyzable or even strictly repeatable. The point is to set the goal post and have a yardstick to measure how well a method aimed at improving empirical sciences does.

It is without question that researchers sometimes engage in inductive behavior that is neither formally analyzable nor repeatable. This fact implies that for theses situations the optimism bias can not be fully quantified and that full transparency can not be archived. To enable proper judgment of the whole theory, the imperative is simple: induce only what is necessary and what you induce should, if at all possible, be done formally. Otherwise, the supposedly objective test of the theory using hard data must still be judged subjectively.

We, therefore, have two bounds that limit transparency about the inductive process. First, some things can not be made transparent in principle because some induction happens informally, in the sense that we can not estimate the optimism with certainty. Second, even formalized inductive processes only allow to estimate optimism theoretically but they have to be communicated effectively to do so in practice.

In the following, I propose preregistration, as means to move induction into the formal domain, and computational reproducibility to make formal induction transparent.

Transparency about statistical models: Computational Reproducibilibiltiy

Transparency about human researcher: Preregistration

Discussion

Limitations

Theoretical

Practical

Future Research

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423. https://doi.org/10.1037/h0020412

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bengio, Y., & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *The Journal of Machine Learning Research*, *5*, 1089–1105.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Burnham, K. P., Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed). Springer.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Duhem, P. (1976). Physical Theory and Experiment. In S. G. Harding (Ed.), *Can Theories be Refuted? Essays on the Duhem-Quine Thesis* (pp. 1–40). Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0_1

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Goodfellow, I., Bengio, Y., & Courville, A. (2016a). Capacity, Overfitting and Underfitting. In *Deep learning* (p. 110). The MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016b). *Deep learning*. The MIT Press.

Hauenstein, S., Wood, S. N., & Dormann, C. F. (2018). Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation. *Communications in Statistics - Simulation and Computation*, *47*(5), 1382–1396. https://doi.org/10.1080/03610918.2017.1315728

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lee, Y., & Pawitan, Y. (2021). Popper's Falsification and Corroboration from the Statistical Perspectives. In Z. Parusniková & D. Merritt (Eds.), *Karl Popper's Science and Philosophy* (pp. 121–147). Springer International Publishing. https://doi.org/10.1007/978-3-030-67036-8_7

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Open Science Collaboration. (2015). Estimating the reproducibility of psycho-

logical science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Soch, J., Faulkenberry, T. J., Petrykowski, K., & Allefeld, C. (2020). Law of the unconscious statistician. In *The Book of Statistical Proofs*. Zenodo. https://doi.org/10.5281/ZENODO.4305950

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 44–47.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1. https://doi.org/10.2307/2331554

van Orman Quine, W. (1976). Two Dogmas of Empiricism. In S. G. Harding (Ed.), *Can Theories be Refuted? Essays on the Duhem-Quine Thesis* (pp. 41–64). Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0_2

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790