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). The debates about psychology's shortcomings and remedies fall roughly into two categories: 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).

I propose to view both categories through the lens of transparency about the inductive process. One side of transparency is that statistical methods make induction quantifiable; the other is that open science measures remove uncertainty about sociological factors in induction. Any science must be able to communicate how it generates its knowledge, but transparency has an outstanding role in psychology and other empirical sciences. To understand why transparency is crucial, we must understand that induction is integral to empirical sciences, formally using statistical procedures or informally as researchers' judgment. However, 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 psychological sciences but is an indispensable property.

To function as a 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 testable claims by, e.g., implying a mean difference between two groups, 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 the estimation of variance from data, and 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 assump-

tions so that they 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 researcher 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" that is 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 how they think theories should be appraised, 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. But whatever measure they subscribe to, their appraisal of the empirical support of a yet untested theory would be relatively low.

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. The question of this dissertation is, therefore, 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 in practice.

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 sciences that wants to select a theory according to its expected predictive performance.

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 probability to observe the state of the world x where f represents the full reality. We are now interested, how much information is lost if we use $g(x)$, our theory, instead of $f(x)$, the reality, over all possible states x. Expressed as lost bits of information, a measure known as Kullback–Leibler divergence (Kullback & Leibler, 1951), we get:

$$\mathcal{J}(f,g) = \mathbb{E}_x\left[log(f(x))\right] - \mathbb{E}_x\left[log(g(x))\right]$$

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

$$\mathbb{E}_x\left[log(g(x))\right]$$

To allow for quantities to be induced, we must assume that our theory is parametrized, e.g., $g(x|\theta)$. That means our theory implies a multitude of possible probability distribution that may describe reality. Though, this assumption may be lifted, we assume each implied version of the theory equally probable a priori. This parametrization captures the idea that some assumptions necessary in 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 parametrization is archived by:

$$\theta_o = \arg\min_{\theta} \mathcal{I}(f, g(\cdot|\theta))$$

The inference goal is therefore:

$$\mathbb{E}_x \left[log(g(x|\theta_o)) \right]$$

Of course, we usually do not know θ_o. That is why it is necessary to induce it from data, denoted as $\widehat{\theta}(y)$. 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_o. It follows that, almost surely:

$$\mathcal{I}(f, g(\cdot|\widehat{\theta}(y)) > \mathcal{I}(f, g(\cdot|\theta_o))$$

Or ignoring f as constant:

$$\mathbb{E}_x[log(g(x|\widehat{\theta}(y)))] > \mathbb{E}_x \left[log(g(x|\theta_o)) \right]$$

That is to say, any induced estimate will be suboptimal. 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 we may form an expectation over the data y:

$$\mathbb{E}_y \mathbb{E}_x[log(g(x|\widehat{\theta}(y)))]$$

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 follows in this section, another is discussed in the next section.

We might notice that we often get an unbiased estimate for $\mathbb{E}_x[log(g(x|\widehat{\theta}(y)))]$, e.g., in maximum likelihood estimation. The expectation over the data together with Taylor series expansion yields:

$$\mathbb{E}_y \mathbb{E}_x[log(g(x|\widehat{\theta}(y)))] \approx \mathbb{E}_x[log(g(x|\widehat{\theta}(y)))] - tr[J(\theta_o)I(\theta_o)^{-1}]$$

The observed likelihood $log(g(x|\widehat{\theta}(y)))$ is, therefore, a biased estimate of the distance to the truth. So substituting deduced quantities by induced estimates leads to some overconfidence about how close one is to the truth.

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, this bias $tr[J(\theta_o)I(\theta_o)^{-1}]$ can be approximated under some conditions. One condition is, since θ_o is unknown, that we know the properties of the inductive process that generated $\widehat{\theta}$. That is, it is insufficient to know the end result of the induction, we must understand the inductive process that generated those estimates. That requires that the inductive process can be formally described, e.g., corrections for a large class of statistical models, most famously the class of linear models, are available.

To summarize, what does this mean for researchers evaluating theories? A researcher who empirically evaluates a theory without inductive elements gets a good estimate how close they are to the truth (relative to other theories evaluated on the same data). If we believe they did not make any mistakes and that they are truthful, we may take their assessment at face value. We might still want transparency on how they did gather the data and how they compared it to the theories. However, it is a simple kind of transparency, about what did happen. If the researcher induces some quantities, transparency about what did happen is not enough; instead a transparency about what could have happened, had the data looked different, is required.

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. It is important to realize that the information theoretic setup above has made no appeal to expected performance on unseen data. Verisimilitude and expected performance are different motivations for transparency, though they can be linked. If predictions and observed facts are compared using a sufficient statistic, a close link to information theory exists. The expectation over data $\mathbb{E}_y\mathbb{E}_x[log(g(x|\widehat{\theta}(y)))]$, can be estimated by repeatedly sampling data and repeating the inductive process. This insight connects the information theoretic setup with cross validation, where the data is partitioned and the inductive process is repeated on all permutations of a subset of the partitions (Stone, 1974, 1977). For each subset, the resulting model is then compared to the complement that was not used for induction. Of course, there is no strong reason to use a sufficient statistic for this comparison beyond its link to information theory. So researcher might differ in what they find to be a good prediction (replacing the Kullback–Leibler divergence with an arbitrary loss function) and we do not longer appeal to a concept of ground truth:

$$\mathcal{L}(x, g(x|\theta))$$

Again there exists some optimal θ_o:

$$\theta_o = \arg\min_{\theta} \mathcal{L}(x, g(\cdot|\theta))$$

Choosing $\hat{\theta}(y)$ based on a

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 answer about the how. 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 by a model (not necessarily a statistical one), or put differently a multitude of implications about the data from the theory. This multitude is a result of arbitrary assumptions.

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

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 mislead because these two factors, induction and limited sample size, interact. There are two principled approaches for this problem. Choosing the version of our theory according to the best fit on one sample and use another sample to asses it:

Note, this is not strictly correct. No expectation over the inferred parameters is taken.

$$\mathcal{L}(\text{Model}(\text{Train}), \text{Test})$$

Or use the same sample for both, but correct for the process of induction:

$$\mathcal{L}(\text{Model}(\text{Train}), \text{Train}) + \mathcal{C}(\text{Model})$$

Where $\mathcal{C}$ denotes complexity or equivalently the extend to which data is influencing the results. To deal with this bias, I propose to split the term in complexity that can be formally described and a part that can only evaluated subjectively. Formal, in the strict sense means that the complexity can be mathematically derived and in a looser sense that the process can be repeated at will. If a researcher employs a linear model, the complexity is calculable and even if it were not, we could fit the linear model on a large set of other data to asses its inductive behavior. However, if the researchers reconsiders the model based on the results, and fits a model with, e.g., a predictor added or removed, the complexity can not be formally judged. We can not ask the researcher to repeat the process on

all possible data sets nor is the behavior mathematically well defined. What we know, however, is that the resulting linear model is more complex than a linear regression formally implies.

$$\mathcal{C}(\text{Model}) = \mathcal{C}(\text{formal}) + \mathcal{C}(\text{informal})$$

That is not to say that we have no basis to judge the informal induction. We can find the inductive decision reasonable and judge it unlikely that just about any variable would be added if the data suggests it. Or the opposite, we might find it not well justified on theoretical grounds and deem it a purely data driven decision, which implies higher complexity. What one can say, however, is that this judgment is debatable and, therefore, subjective.

How to establish transparency

Using a formally defined inductive process to preemptively escape refutation differs in at least one important way from altering the hypothesis after refutation. Induction through statistical models can be formally analyzed and accounted for. Therefore, the researchers may adjust their appraisal of the partially induced hypothesis. Adjusting the appraisal for induction outside of statistical methods is much more uncertain. Here the process of induction can not be separated from the data itself, since we observe such induction only on a single dataset. 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.

If this imperative is adopted, it still remains to make this process transparent to other researchers. Not an easy feat as we will see. We, therefore, start with the part that is easier made transparent. The complexity due to the formally defined inductive process may be made transparent by employing computational reproducibility. For the more informally made inductive decisions, we have a more limited toolbox that includes preregistration.

Transparency about statistical models: Computational Reproducibilibiltiy

Even if the results stand at the end of a formal inductive process, this process must be made transparent to other researchers if they are to judge the empirical test. Such transparency can be achieved by computational reproducibility. Computational reproducibility is usually defined as the ability to recreate the same results from the same data set. This definition only suffices if the inductive process is strictly formal, that is the complexity can be calculated mathematically. If we must rely on repetition to assess the complexity, we must demand computational reproducibility to extent to other data sets.

Transparency about human researcher: Preregistration

Discussion

Limitations

Theorethical

Practical

Future Research

Accounting for how much induction was used to deductively test a statement is necessary for correct appraisal of evidential support of a hypothesis. Researcher (and philosophers of science) differ considerably in what appraisal entails, e.g., judging the long-term performance (if they are frequentists), degrees of believe (if they are Bayesians), or probativeness (if they are severe testers, Mayo, 2018, p. 14) of a hypothesis. But whatever measure they subscribe to, induction can shield shield any hypothesis from refutation by data altogether.

Assume that all auxiliary assumption, regarding proper quantification of behavior and situation in valid measurements, causality, individual differences, nuisance variables, distributional, etc, hold.

While not using statistical methods would lead to the rejection of almost all psychological theories, pushed to the extreme statistical methods can shield theories from refutation altogether. Both situations are uninformative and, therefore, rather unhelpful for the generation of knowledge. The later situation, however, becomes less than helpful when one is under the impression that theories are well tested, when in fact there was little chance that the data could refute the theory if wrong.

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.

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

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

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 psychological 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

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