Quantitative Explanation as a Tight Coupling of Data, Model, and Theory

Alexander Krüger, Jan Tünnermann, Katharina Rohlfing and Ingrid Scharlau

Abstract What does it mean to explain data patterns? Cognitive psychologists and other scientists face this question when observable phenomena have to be explained in theoretical terms. Frequentist null-hypothesis testing – one prominent approach in psychology – controls error rates. Machine learning – an alternative prominent outside of, but not yet inside psychology – focuses on precise predictions. However, both alternatives often provide little insight into the data. We propose a combination of formal modeling and Bayesian statistical inference to ground explanations in data analysis. We support this approach by reference to philosophy of science and discussions of the current methods crisis in several empirical sciences and illustrate it with an example from visual attention research.

Alexander Krüger · Katharina Rohlfing · Ingrid Scharlau
Paderborn University
✉ alexander.krueger@uni-paderborn.de
✉ katharina.rohlfing@uni-paderborn.de
✉ ingrid.scharlau@uni-paderborn.de

Jan Tünnermann
Marburg University
✉ jan.tuennermann@.uni-marburg.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/10
ISSN 2363-9881

1 Introduction

Empirical scientists seek to explain their observations or data. Providing insight into data is also a key requirement of practitioners in the field of data science. An evocative example of the difficulties of explaining data is provided by Lewandowsky and Farrell (2011) who discussed why Ptolemy's geocentric model was so rapidly replaced by the Copernican heliocentric model. It is commonly believed that the Copernican model provides the better account of the data. Its goodness of fit is, however, only slightly better. Its advantages are its simplicity and elegance. Lewandowsky and Farrell use this example to conclude that:

> " 1. *Data never speak for themselves but require a model to be understood and to be explained.* 2. *Verbal theorizing alone ultimately cannot substitute for quantitative analysis* 3. *There are always several alternative models that vie for explanation of data, and we must select among them.* 4. *Model selection rests on both quantitative evaluation and intellectual and scholarly judgment"* (Lewandowsky and Farrell, 2011, p. 5).

We would like to add another point: 5. Theory development, modeling and data collection lead to explanations in an iterative process. Such a process may start with a vague verbal theory and some loosely connected observations but develop into an abstract mathematical description of theoretically relevant, measurable variables. To provide a rigorous argumentation for these 5 points, we will present positions from the philosophy of science. Initially, we will review what actually *is* an explanation in psychology. This approach reveals the importance of theory. Different theories entail different data patterns. We then present a framework in which *models link data and theory*. *The quality of this link is what makes a good explanation.*

2 Explanations in Psychology

In epistemology, the fact to be explained is called the *explanandum* whereas the part of the explanation doing the explaining is called the *explanans*. For psychology – and possibly other sciences dealing with data – finding a causal relationship between variables (an effect) is no explanation, but the thing to be explained (Cummins, 2000). A good explanans here would capture the functional properties of the processes that cause the observed relationship.

These functional properties are latent causes of the data. Psychology thus is a science that usually asks "How does it work?" (instead of "What are the laws?" in axiomatic frameworks such as classical physics) and uses *functional analysis* to answer this question (Cummins, 2000). Functional analysis means explaining a disposition of a system by resorting to simpler mechanisms. There are different frameworks to do so. Examples from psychology include computational explanations which frame the function as an information processing system that follows certain algorithms. Within this framework, the cognitive psychologist may ask what algorithm would warrant an observed effect. A further example is the neuroscientific framework that resorts to biological properties of neurons and their interactions. These frameworks have in common that they allow to explain a system in terms of simpler and less problematic subsystems.

Models play a crucial and often overlooked role in bridging the gap between hypothetical causes of data and the data themselves. Of course, model is a broad term whose meaning differs depending on the domain and historic context. We mean models in the sense of scientific models for which Bailer-Jones (2009) provides a historical review. Summing up her findings, models were initially considered "a poor man's theory", a form of description or explanation of a phenomenon that is entirely superseded by a theory if such a theory becomes available. In present times, models have gained relevance in philosophy of science because it has been shown that theories cannot be tested directly. Instead, many assumptions have to be made on different levels to develop a concrete experiment for testing an abstract theory (Suppes, 1966).

For example, take Newton's laws of motion. Dropping a feather and a ball of equal mass from a tower will apparently refute the laws: Although the mass is equal, the speed of fall will differ. It is only through our theoretical model of the experimental situation that we know that in this particular situation air drag an skin friction will be different to a degree that severely affects the outcome. Even if both objects are equal in mass and surface properties, a precise empirical measure of speed of fall is highly unlikely to yield exactly Zero as predicted for the speed difference. Thus, data of a phenomenon is not the same as the phenomenon itself.

Tackling the difficult relation between empirical phenomenon and theory, Bailer-Jones (2009) proposed a framework that identifies a hierarchy of models as the missing link. According to her analysis, "a model is an interpretative description of a phenomenon that facilitates access to that phenomenon" (p. 1).

Although neither the term “data” nor the term “theory” is part of this definition, her analysis reveals that they are what models connect in scientific research. Models apply abstract theories to concrete phenomena and do this by satisfying abstract logical constraints of a theory and concrete empirical constraints of a phenomenon, although as interpretative descriptions they can still be abstract and incomplete. In other words, models are customizations of theories such that these become applicable to some of the concrete properties of the phenomena by filling in the gaps between latent causes and data.

As model and theory can be distinguished, so can phenomenon and data. A phenomenon is a fact or event in nature – “things happening” in Bailer-Jones’s words (2009, p. 1). It becomes apparent from observation and is at least suspected to be stable and not a random fluke. The discovery of a phenomenon can be theory-laden. To give an example from attention research: When searching for a unique target item among homogeneous distracting items, search difficulty does not always stay the same if target and distracting items are swapped. For instance, it is easier to search for a Q among Os than for an O among Qs (Treisman and Gormican, 1988). This phenomenon, search asymmetry, would not have been recognized without previous research on visual attention, especially in the field of visual search.

The phenomenon can change with investigation. Closer inspection and analysis may refine a phenomenon or show that it is not as stable as initially suspected so that it does not warrant investigation. One prominent psychological example is the face in the crowd effect according to which angry faces are found quicker in a crowd of neutral faces than happy faces. It is usually explained by one or the other version of preferential detection/processing of relevant stimuli. Although it has been under investigation by now for 30 years, it is not yet clear whether the effect is an artifact of the faces in the crowd (Horstmann et al, 2006) or some confound in schematic faces such as inward-pointing lines (Coelho et al, 2010; Kennett and Wallis, 2019); a discussion is provided by Savage et al (2013).

What separates data from the phenomenon is that data arises from specific ways of observation or experimentation (e.g., a psychological experiment, a log-file, or a health record), often with a particular goal in mind. The data collected thus are a limited aspect of the phenomenon and they may be affected by more than the phenomenon itself. This creates uncertainty.

Having distinguished models and theory as well as phenomena and data, we now turn to the question how models link data to theory and why this link is necessary for science. A *theoretical model* is necessary as a model of the situation in which the data is observed. A *data model* is necessary to deal with the uncertainty arising from data collection. Consider the initial example of the Copernican and Ptolemaic models of planetary motion – two different theories. Both had a comparable fit. Only Kepler provided a better explanation of the phenomenon of planetary motion by linking the observed data to the heliocentric theory by a theoretical model that uses ellipses instead of circles such that a nearly perfect fit was achieved. In visual attention research one might, for instance, ask whether location is to be conceptualized as a property which organizes visual information or whether it is a feature of visual objects such as their color or shape. Both theoretical models accord well with some phenomena and less well with others. To decide between these theories, a tight coupling between data and the respective theory in formal models is necessary (following this logic, Nordfang et al (2018) indeed showed that location is special).

Bailer-Jones's (2009) example for the second type of model, the data model, are measurement errors in physics: When measuring the melting point of lead, the true melting temperature may be not read off the thermometer once during an experiment. If a normally distributed error is assumed, the mean of many measured values will be a good indicator for the true melting temperature. To give a psychological example, the minimal response time to a warning signal in a semiautonomous car has to be inferred from a sample of responses from different participants. Because response time cannot be negative, a hierarchical model based on ex-Gaussian or log-normal distributions represents the functional properties of the data-generating process adequately.

One may wonder whether measurement theories are an alternative to the data model in the proposed hierarchy of models. For measurement in psychology, (Buntins et al, 2017) analysed whether the two prevalent positions on measurement in psychology can guarantee that a measurement is valid in the sense that it measures indeed what it is supposed to measure. These positions are

1. that measurement is a rule-driven assignment of numbers to objects, and
2. that measurement is based on a homomorph relationship between an empirical and a numerical relational system (p. 704).

The authors find that both measurement theories cannot resolve competing theoretical explanations of results and consequentially, psychometrics alone does not guarantee a connection of theory and the measurement (data) via the formal concepts of validity. As such, measurement theories may help to develop parts of the proposed hierarchy but do not provide a complete link between data from measurements and theory.

If a scientific hypothesis is to be tested, this cannot be done directly on the data because of measurement errors and variance from nuisance sources in the data and because data always needs to be interpreted under a theoretical model. Data models deal with these inference problems. Such statistical inference does not directly test a *theory*, but the prediction of the *theoretical model*. A theoretical model represents an experimental or observational situation and satisfies empirical as well as theoretical constraints. The *data model* deals with the uncertainty that collecting data about a *phenomenon* introduces. Connecting data model and theoretical model, thus, connects data and theory.

Figure 1(A) depicts Bailer-Jones's (2009) framework complemented by the link between theoretical model and phenomenon. This bidirectional connection symbolizes that a certain phenomenon may compel us to come up with a theoretical conceptualization of the situation in which it occurs – the theoretical model. A theoretical conceptualization of a situation may lead the researcher to observe an unexpected event or fact that needs explaining – a new phenomenon.

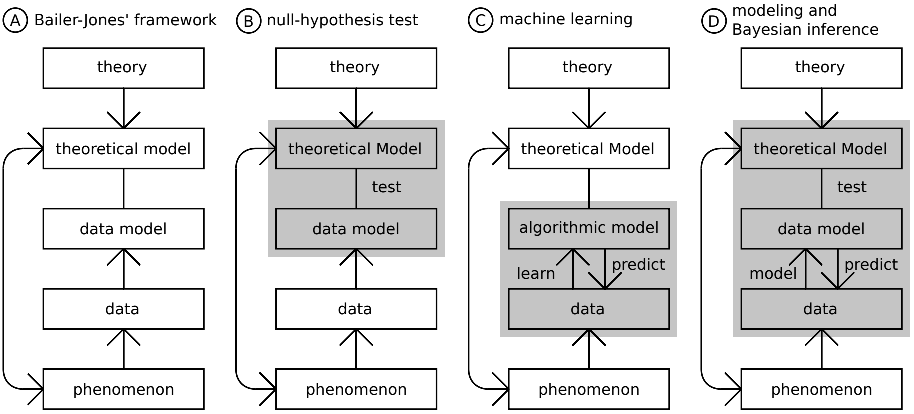

Figure 1: (A) Bailer-Jones's (2009) framework of how models link data and theory. (B) Framework parts centrally involved in null-hypothesis testing. (C) Framework parts centrally involved in ML. (D) Framework parts centrally involved in modeling and Bayesian inference.

Conjointly, Bailer-Jones's (2009) framework and Cummins's (2000) analysis provide a philosophical underpinning for the five points based on Lewandowsky and Farrell (2011): Because data cannot be understood by themselves and there are always different possible explanations of data, model selection requires quantitative evaluation of theoretical ideas. The quality of the explanation offered by a model can then be evaluated by taking into account the fit and how well the model satisfies empirical and theoretical constrains. Together, these authors describe the main requirement for scientific data analysis in psychology and other sciences dealing with data – a tight coupling of data and theory. In the following, we will analyze how statistical null hypothesis testing, machine learning and modeling in conjunction with Bayesian inference techniques meet this requirement.

3 Explaining Data by Linking it to Theory

3.1 Coupling Data and Theory with Frequentist Hypothesis Testing

The predominant way to couple data and theory in psychology is to derive a prediction about the data from theory – the hypothesis, a preliminary answer to a scientific question. Whether the answer is supported by the data or not is tested by a statistical hypothesis test. The test provides a decision procedure for which the long-term error rates are controlled. Very often, the goal of the statistical inference is to check whether an independent variable Y has a effect on a potentially dependent variable X. In practice, such an analysis is conducted by first setting up a null hypothesis. This null hypothesis describes the 'no mean difference' or 'zero correlation' hypothesis. The exact prediction of the hypothesis which the researcher is actually interested in – a substantial theoretical alternative – is not specified. As a second step, a conventional level for the long-term false-positive error rate (in practice mostly 0.05) is used to try to reject the null hypothesis. If this is indicated by a p-value of less than 0.05, the result is called significant and the alternative hypothesis – whose prediction is not explicitly stated – is accepted (for an analysis of the actual use of this technique, see Gigerenzer, 2004; for a historical and theoretical analysis, see Dienes, 2008).

During the last years, many reported findings (effects) could not be replicated, for instance in psychology, cancer biology, or medicine. A variety of reasons contributes to this (e.g., Chambers, 2017), but one main cause is application of statistical practices which critics as Gigerenzer (2004) call "ritualized", emphasizing that hypothesis testing is often done mindlessly and that more inferential power is ascribed to the procedure than it actually has. One problem is that researchers often "follow" the data. Data trimming is one of the examples: removing outliers or apparently inadequate data or grouping data by some criterion. Simmons et al (2011) showed via simulation that such seemingly minor analysis decisions can increase the chance of falsely significant results to as high as 60% if several questionable practices are combined, and John et al (2012) found that some questionable practices actually seem to be the norm rather than the exception. Even without the intention to apply such practices, many analysis decisions are contingent on the data and thus affect the reproducibility and theoretical conclusion. In the context of null hypothesis testing, this means that researchers use undisclosed freedom in data analysis. Such practices violate assumptions of the frequentist model used in these tests: This model presupposes that the hypotheses are not informed by the data and that all aspects of data sampling (stopping rule, measurement, variables to be included in tests) are fixed in advance.

To sum up (see Figure 1(B)), the rationale of frequentist hypothesis testing is to decide on a theoretical idea by setting up another hypothesis and rejecting it. In the terms of Bailer-Jones (2009), a data model is chosen from a toolbox of models. The theoretical model is only verbally formulated (if at all). The connection between theoretical model and data model exists only by the single decision made against the (often uninteresting) null hypothesis. Because the coupling of data and theory does not include the theoretically interesting alternative to the null hypothesis, no strong coupling of data and theory is achieved. For instance, when applying a t-test, the t-distribution is chosen for its robust description of the data under the null model while the alternative models – the ones in which the researcher is actually interested – remain unspecified. The advantage that this procedure controls the long-term error rates of the false-positive error is countered by the undisclosed flexibility in data analysis that can even unconsciously distort the control of error rates. Of course, the problems of ritualized and mistaken application of frequentist hypothesis testing can be mitigated. More important is, however, to realize that all statistical analyses depend on a specific model (Rodgers, 2010), make certain presuppositions and are more or less adequate or useful (although there may not be a true model).

3.2 Coupling Data and Theory with Machine Learning

Machine learning (ML) is a much less canonical way of analyzing data for psychologists. Yarkoni and Westfall (2017) emphasize the benefit that ML explicitly tackles the variance-bias trade-off, that is the trade-off between precision and systematic errors of predictions: Whereas in psychology bias is usually avoided at all cost, ML is very aware of the costs of such a minimized bias for prediction errors. ML searches for a trade-off between bias and variance that minimizes prediction errors.

ML can be put into perspective with Bailer-Jones's (2009) framework (see Figure 1(C)): It improves a program's performance – usually in predicting unobserved data – by learning from examples. Whether the model corresponds to the data-generation process is irrelevant to this improvement. The approach of choosing a model according to performance instead of a similarity to the data generation process is called algorithmic modeling as opposed to data modeling which aims to resemble the data-generating process. The focus on model performance instead of model similarity to a supposed function of the system makes functional analysis difficult because the connection between theoretical ideas and the actual data pattern is not spelled out as a model. We would like to remark here that functional analysis and algorithmic modeling are not necessarily mutually exclusive: If the algorithmic model is designed according to a theory about the observed phenomenon, algorithmic modeling can be insightful. However, such an approach requires knowledge about algorithm design as well as relevant theories (see e.g., Jäkel et al, 2008).

ML is often viewed with skepticism by psychologists. This is problematic not only because its predictive power is so high, but also because common ideas of the predictive power of canonical psychological analyses may be unjustifiably optimistic. Yarkoni and Westfall (2017) argue that psychologists' undisclosed flexibility in data analysis and goodness of fit measures are comparable to overfitting and psychology as a whole could profit from ML techniques such as cross validation and regularization to avoid overly optimistic beliefs in the predictive power of models.

The perceptive reader may notice that we did not discuss unsupervised learning. Unsupervised learning is characterized by self-organization that is data-driven. Thus, it may be particularly valuable to come up with a theoretical model in cases where there is no formal theoretical model available as of yet. However, it is particularly this powerful data-driven self-organization that can give rise to "ghosts" as Carlson et al (2018) argue for cognitive neuroscience. By ghosts the authors mean that results about phenomena are apparently interpretable in terms of a theory. However, if the link between data and theoretical argument is required to be spelled out, assumptions are revealed that are superimposed on the actual ML procedure. These assumptions concern the source ambiguity of the data, the perceived neutrality of the the ML procedure, and the underconstrained representational interpretation of results. Although Carlson et al (2018) make a suggestion on how to cope with these problems, their analysis also reveals that the connection between phenomenon and theory is neither created automatically nor does it become obsolete when unsupervised methods are applied.

To sum up, a ML perspective may improve psychology's predictions by choosing models according to a variance-bias trade-off that is favorable for prediction. These models are, however, not required to resemble the data-generating process proposed by a particular theory. Thus, it is difficult to provide a quantitative explanation of the data in terms of a theory.

3.3 Coupling Data and Theory with Modeling and Bayesian Inference

We propose a modeling scheme that is different from the two approaches described above. It is based on combining formal modeling with Bayesian inference for explanations grounded in data analysis. The need for this originates from weaknesses in the two previous modeling views that can be identified using Bailer-Jones's (2009) framework: Machine learning intentionally disconnects theoretical model and data model. This allows to optimize the prediction performance, but it hinders causal explanation and theoretical advancement. Frequentist statistics controls error rates, but its inferential power is often low. Hypotheses are not directly connected to the theoretical model that the scientist has in mind. In the following, we spell out a different and more substantial modeling approach based on explicit and formal models and Bayesian inference.

Our focus is on integrating the theoretical model into data analysis. As Taagepera (2008) shows, data models can be very simple, for instance when forbidden areas or anchor points in the data space are taken into account or when adequate probability distributions are used. For example, response times – an almost ubiquitous variable in psychology – cannot be negative and their distribution is likely to be skewed. Thus, for example, ex-Gaussian or log-normal distributions may fit them well. Taagepera calls such models that take the specific logical properties of the data-generating process into account *logical models*. They are more specific than *descriptive models* that are merely a convenient description of the data, although both are data models.

One approach to logical modeling of cognitive functions are mathematical models (Moore, 2015). An example from psychophysics is Steven's power law that relates different aspects of physical stimulus strength to perceived stimulus intensity, resulting in a power function that describes this relation across a wide range of stimulus features. A yet closer link between theory and data model can be established when fine-grained mathematical models are available for different aspects of a theory. These models not only describe the observed data patterns but the processes that lead to them. These fine-grained components can be used to assess explanations. For instance, Bundesen's (1990) theory of visual attention (TVA; see Bundesen et al, 2015, for a recent account) describes the encoding processes of visual stimuli in a way that expected data patterns for different psychophysical tasks can be derived. Moreover, the description is hierarchical: the encoding process itself can be mathematically described as a combination of sub-processes, which again can be further dissolved. Of course, this reductionism cannot be pursued unlimitedly. However, in practice, an observed data pattern can be explained as arising of relatively simple components. In section 4 of this article we will illustrate a concrete example of how TVA can be used in visual perception research, highlighting the merits of such a fine-grained mathematical model.

Even though such formal models can be treated with various methods (e.g., maximum likelihood estimation), we believe that Bayesian estimation is particularly well-suited. As graphical Bayesian models, the different components of the theoretical model can be linked with deterministic or probabilistic connections. Moreover, the hierarchical structure of the data (e.g., groups →

participants → conditions → etc.) can be expressed and posterior parameter distributions and their uncertainties at the different levels are accessible to the researcher (Figure 3(F)).

Bayesian inference allows to estimate latent variables, compare models, and predict data. When parameters are estimated (a decision between models can be considered a parameter), the researcher assigns prior distributions to the parameters. The estimation procedure updates these distributions according to the data, leading to the posterior, a distribution of probable parameter values. Doing this analytically is challenging and often impossible. However, developments in probabilistic programming allow to specify models as computer code and to estimate posteriors by numerical approximations (cf. Salvatier et al, 2016).

Specification of a prior may strike the reader as an unwanted source of discord and divergence from objective analysis, and the specification of a substantial model is a somewhat subjective endeavor. However, you may also regard this as an advantage: The Bayesian approach forces the researcher to make her implicit assumptions explicit (Rouder et al, 2016). It thus also allows to falsify theory-derived statements (Gelman and Shalizi, 2013). In fact, as Bailer-Jones (2009) argues, a model is necessary to state how the theory is interpreted by the researcher in the particular situation that is supposed to falsify it.

Bayesian statistics are also more intuitive than frequentist methods: Prior and posterior distributions reflect belief as a degree of confidence and not the long-run frequencies of outcomes that give frequentist approaches their name. Long-run asymptotic behavior of a chance experiments seems to be difficult to grasp (this can be tested by checking one's intuitions about data analysis (Dienes, 2008, p. 121)). Given the fact that creating, testing, and improving models is a creative and central process in modern science, it is helpful if researchers can follow their intuition here (although we do not want to imply that intuition is without problems).

We would like to remark that the Bayesian and the frequentist understanding of probability are often identified with "subjective" and "objective" probabilities, respectively. In his comprehensive book on inductive logic, Hacking (2001, p. 131) emphasizes that these are ideology-loaded terms and recommends to avoid them in debates. It is true that Bayesian notion of probability is belief-centered whereas frequentist notion is event-centered. However, dismissing all beliefs as merely subjective can be misleading: Reasoning about the probability of an asteroid hitting the earth causing the extinction of the dinosaurs based on the physical laws and iridium traces is not what we commonly mean by a subjective opinion. However, integrating the evidence for or against an event in the past is a prototypical example of the Bayesian notion of probability. Similarly, in physics research, Bayesian data analysis is not understood as particularly subjective because the researchers usually have to select a model and decide upon the likelihood, based on scientific judgment independently of whether Bayesian priors are selected by scientific judgment as well (von Toussaint, 2011). One reason why the event-centered probability concept may be so dominant when discussing the normatively right and objective way of doing statistics is given by Gigerenzer (1991): He proposes that theories of rationality and cognitive abilities are not conceived independently of the methods used. With frequentist statistics being the method of choice since the cognitive revolution, theories on rationality and cognitive abilities often treated a deviation from the frequentist analysis of uncertainty as subjective and irrational even though it can be rational under a Bayesian perspective. Consequentially, we think that associating one type of probability with objective facts and the other type with subjective opinions is an oversimplification that should be avoided.

In sum, formal modeling goes well together with Bayesian inference. A tight coupling of the data and theory can be achieved by explicitly stating and connecting the data model and the theoretical model. This allows quantitative comparisons of models, model predictions and latent causes of data.

3.4 Comparison of the Three Approaches

In the previous sections, we introduced three approaches to explain data by linking it to theory with different analysis methods that are currently applied. Also, we argued that a combination of formal modeling and Bayesian inference is particularly well suited for this task. In the present section, we aim at providing a concise comparison.

There are different levels on which these methods can be compared. The most abstract level is inductive logic, a part of philosophy. On this level, the implications of different understandings of probability are discussed (for an introduction see Hacking, 2001). These include belief-type probability, the Bayesian understanding of the term, and frequency-type probability as in the common event-based statistics. Inductive logic has implications for the understanding of science and for the understanding of statistics (for two opposing views see Mayo, 1996; Gelman and Shalizi, 2013). Bayesian inference and null-hypothesis testing have fixed understandings of probability. ML does not buy into one of these understandings and is seldom discussed on this level. However, the idea of different types of quantifiable uncertainty, i.e. probability, can be used in ML to distinguish uncertainty that stems from the model used from uncertainty that stems from a lack of knowledge (Senge et al, 2014). An exception that discusses science as a possible instance of meta-learning is provided by Korb (2004).

If we want to compare approaches on a level connected more tightly to psychological research, an interesting level concerns the inference techniques that are actually used in a community. Oakes (1986), for instance, analyses how statistical methods have been used in social, biological, and behavioral sciences. More recently, Dienes (2008) discussed null-hypothesis tests, maximum likelihood estimation and Bayesian inference for the field of psychology. This is different from the conceptual understanding because these debates revolve not only around the theoretical properties of methods but their actual usage. In psychology, Bayesian inference and frequentist hypothesis testing have been compared by e.g. Dienes (2011); recommendations for their respective usage are provided by statisticians (e.g., Little, 2006; Efron, 2005). Our evaluation of the fledgling use of ML in psychology has been drawn from Yarkoni and Westfall (2017).

Within psychology, three criteria are especially relevant for researchers. These are the estimation of theoretically relevant parameters (e.g., the speed of visual processing, or the capacity of visual short term memory), the possibility to predict new data given some observations (e.g., for a particular participant or the population), and the ability to compare different theories (e.g., theories that propose different functions for the data model). Table 1 summarizes the evaluation of the approaches with respect to these criteria.

Null-hypothesis testing offers the outstanding advantage of controlling error rates (Mayo, 2016). In domains like quality assurance, this property is indispensable. When linking theoretical parameters to data, maximum likelihood estimations is used that is – roughly speaking – equivalent to the Bayesian approach without using a prior – hence the brackets around the check mark in Table 1. Because null-hypothesis testing works with the rejection of the eponymous null models without specifying the alternative explicitly, it is not possible to predict data for the scientifically interesting alternative model. Also, theory comparison is indirect: If the null model cannot be rejected for theory A but can be rejected for theory B, this may serve as a comparison in favor of theory B. However, the two theories are not compared directly. This is advisable only if no model can be derived from the theories. Otherwise, a direct comparison of both models given the observed evidence is prudent.

ML is usually not able to estimate a particular theoretical parameter. The reason for this is also the reason for ML's outstanding capacity for prediction: ML is not bound to find the stochastic process that generates the data but to apply algorithmic models while it treats the data mechanism as unknown (Breiman, 2001). However, we do not want to withhold that algorithmic modeling as used in ML can be used to compare theories. Doing so, however, requires deep understanding of algorithmic modeling as well as the theories to be tested so that the theoretical model can be translated into an algorithm. In the study of cognition, algorithmic modeling has gained less interest than the more abstract computation level (e.g., models of the different parts of human memory such as long-term or working memory) as well as the more concrete implementation level (e.g. neuroscientific evidence) on which researchers often focus to advance understanding (Peebles and Cooper, 2015).

Bayesian inference and modeling allows to estimate parameters of interest because it requires the analyst to spell out the connection between theoretical model and data. This estimation proceeds from effects to causes by the Bayes rule. However, for data prediction, parameter distributions can also be assumed for the causes to derive a prediction. Data prediction and parameter estimation depend on the model used. That is, an inadequate model leads to bad predictions and estimations. It may come as a surprise that this is actually advantageous for the search for a good explanation of the data: Because a bad model is – within the framework of Bailer-Jones (2009) – derived from a theoretical model, it will not be unknown why the model performs poorly: A bad model represents poor theoretical ideas about the data generation process. If another theoretically motivated model performs better, the result can be counted as a comparison of theories. Such comparisons are made possible by the Bayes factor (e.g., Rouder et al, 2009) or information criteria (e.g., WAIC, Watanabe, 2010). The resulting comparisons facilitate good explanation because neither their fit nor their theoretical origin is evaluated independently.

Table 1: Summary of the comparison of null-hypothesis testing, machine learning, and a combination of modeling and Bayesian inference.

	Null-hypothesis Testing	Machine Learning	Modeling and Bayesian Inference
Parameter Estimation	(✓)	✗	✓
Data Prediction	✗	✓	✓
Theory Comparison	(✓)	(✓)	✓

4 Application in Cognitive Psychology

We have seen that Bailer-Jones's (2009) framework can accommodate different modeling perspectives. Moreover, we have argued in favor of a close coupling between data and theory via formal theoretical models and data models and Bayesian inference. However, it may have remained somewhat unclear how these concepts can be applied in practice. In the following, we look at concrete common and often interlocked research activities: interpreting published studies, modeling, simulation, and obtaining inferences. For this purpose, we will return to TVA, the theory of visual attention, which was briefly introduced above, and show that it is a well-specified logical model with a deep and broad theory integrating neural interpretations and applications in fundamental as well as clinical research (Bundesen et al, 2015). For another, detailedly spelled-out example from organizational psychology, we refer the reader to a yet unpublished study by Ballard et al.

Bailer-Jones's (2009) framework provides a road map that can guide research at different stages. As we will see, it is less important to follow all directions exactly as specified in Figure 1(D). In fact, depending on the current goal – casting a study from the literature into our theoretical framework, or drawing conclusions from our own experiments – we may need to wander in different directions along the connections of Bailer-Jones's diagram. However, even though we can exploratively go back and forth between phenomenon, data, and the different theoretical levels, we must make sure that our roads are well paved: All the connections should be as firm as possible. The concepts of a possibly verbal or metaphorical theory should be turned into mathematical entities whose relations can be formally defined. Together with the phenomenon and the context that generates the data (e.g., the experiment), the theoretical model and, taking statistical aspects into account, the data model can be formalized.

As a research topic, we look at processing speed differences in the visual hemifields. In particular, we are interested whether processing speed is higher in the left visual hemifield, the so-called left visual field (LVF) advantage. This phenomenon has been reported by different researchers with different explanations, for instance by Matthews and Welch (2015). They related it to the system responsible for motion perception, in their study specifically an attentional motion system. One of their experimental tasks is the temporal-order judgment (TOJ) in which participants judge which of two stimuli was shown first (see Figure 3(A) for an illustration of the TOJ stimulus presentation). While Matthews and Welch include further tasks in their study, we limit the discussion here to the TOJs, to provide a concise picture of the methodological aspects on which our article focuses. Moreover, the descriptions below are just the first steps of a more thorough study (Tünnermann and Scharlau, 2019), which can be consulted for more background, interpretations and follow-up experiments.

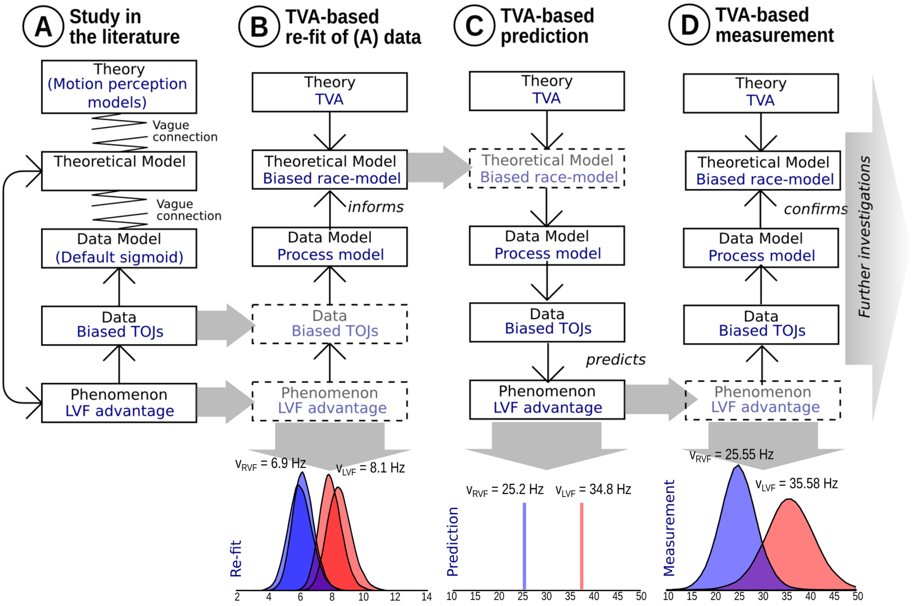

Figure 2: Investigating left visual field advantages with Bundesen's theory of visual attention. The different research stages are mapped to the Bailer-Jones Framework in subfigures A to D.

Matthews and Welch's (2015) study was guided by substantial theoretical considerations about the origin of the possible LVF advantage in systems responsible for motion perception and the way to measure them by temporal-order judgments and other tasks. The study discusses findings on a quantitative level exceeding that of many experimental psychology studies which only detect the presence or absence of certain effects. Matthews and Welch discuss temporal thresholds from the TOJ task (that reflect how much earlier the LVF stimulus finishes processing) and compare them to earlier findings from different, but comparable tasks.

Casting the study in Bailer-Jones' (2009) framework, we see two main origins of vagueness (cf. Figure 2(A)):

1. The authors draw connections between the LVF advantage (phenomenon), the TOJ task, and theories of motion perception which are merely qualitative, building on the idea that TOJs are computationally similar to motion perception.
2. The data model they use is a default psychometric function.

They apply a logistic function that *describes* the S-shaped data pattern (cf. Figure 3(D)) typically observed in psychophysical judgments. It is not derived from a theoretical model of the processes as in the TVA-based perspective we have described. Although it is a model of differential perceptual latencies, and thus theoretically justified, it is impossible to *analytically* trace effects evident in its parameters to deeper meaningful concepts of the theory.

How can we improve on, or at least substantially contribute to Matthews and Welch's (2015) analysis? In section 3.3 we described how mathematical models can be fruitfully combined with Bayesian parameter estimation (see also section 3.4). We will now apply one such model, TVA, to investigate the potential LVF advantage in TOJs. With TVA, the components of Bailer-Jones's (2009) framework can be linked firmly. TVA's concepts are mathematically formalized, enabling a theoretical TOJ model with a precise data model (see Figure 3(B)–(E) for a brief overview and Tünnermann et al, 2017).

TVA describes how visual stimuli presented to a human observer (metaphorically) race for encoding into visual short-term memory (VSTM; see part (A) of Figure 2). VSTM is an early storage system which holds the visual information active for the task at hand or for forwarding it to more permanent memory structures. In healthy adults it is limited to three to four visual items. TVA includes several meaningful parameters that characterize this process. For instance, K is the VSTM limit mentioned above, and C is the overall capacity that determines the overall speed with which the items in the visual field race for VSTM entry. Parameter v_x is the rate with which one particular stimulus x races to be encoded. As TVA considers encoding to occur in "exponential races" (see Bundesen, 1990), it is mathematically described by the cumulative density function in Figure 2, part (B). This function models whether a stimulus' race has already finished at time t.

In broad terms, TVA assumes that overall capacity C can be distributed across the objects in a visual scene. As an example, imagine the children playing in a playground when you look for your daughter. Your capacity C is allocated to the stimuli depending on weights, for instance a higher weight of green objects when you know that your daughter wears a green coat. Thus, the stimulus that is your daughter will be processed at a higher rate than, for example, her friend who wears a blue jacket. In the case of Matthews and Welch's (2015) study, one might assume that weights (or rates) are higher in the left than in the right visual field which is their prediction in TVA terms. Exceeding their approach, we can then trace the origin of weight (or rate) differences in three further parameters, the (objective) visual evidence for a target (η), the (subjective) pertinence (π), and the (subjective) bias (β). Pertinence values reflect favoring object features such as certain colors, and biases pertain to categorization. As indicated at the bottom of panel (B) in Figure 2, speed can be decomposed into visual evidence, bias and weight, and weight into pertinence and evidence (note that the two evidences differ in their subscript). These parameters specified by TVA have well-defined meanings which are established across different attention-related phenomena; they have furthermore been interpreted within the neural interpretation of the theory. Further details of the these components and how they interact can be found in fundamental descriptions of TVA (Bundesen et al, 2015).

In the present context, we derive a model for TOJs (the "which stimulus appeared first" task, see Part (C) of Figure 2) from TVA. In TOJs, participants of the experiment judge which of two similar stimuli appeared first. They do so for a number of different intervals between the stimuli and a large number of repetitions (typically in the hundredth). The judgment may be difficult if the stimulus onsets are separated by only a very small time interval, but is easy to understand and perform.

Part (D) of Figure 3 illustrates a typical data pattern gained by judging temporal order: Across stimulus onset asynchronies (SOAs), judgment probability follows an S-shaped function. With large SOAs, observers make few mistakes, whereas around zero, mistakes are common. To be more precise, in the present example, the observers saw two letters appearing one in the left and one in the right visual field and named the first letter. This judgment is transferred into the probability that the stimulus in the left visual field was judged first. This probability is high when the left letter was indeed first by a large interval and low when in was second by a large interval.

If there is a LVF advantage in processing speed, observers might see the left stimulus first in a high proportion of trials, even when the two stimuli are presented at the same time (SOA zero) or the right stimulus leads with a small interval. Indeed, the whole psychometric function would be shifted horizontally (cf. Figure 3(D)). This is what Matthews and Welch (2015) observed. Modelling with TVA can now help to pin down how this shift arises from lower-level processes.

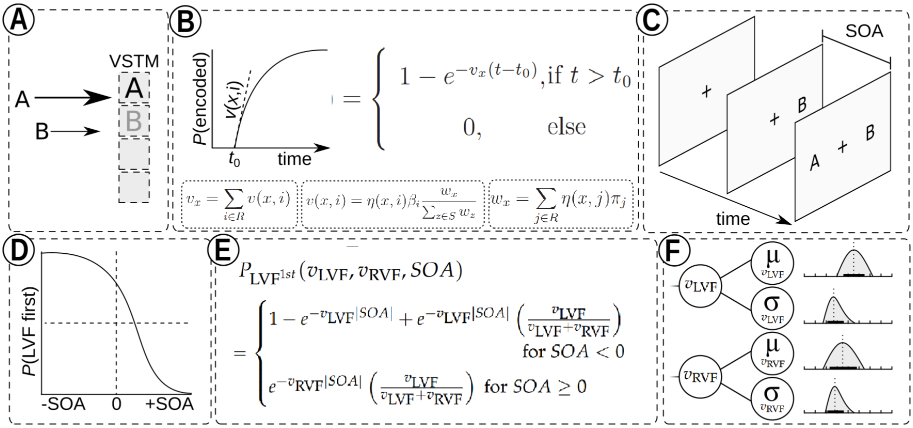
A
VSTM
A
B
A
B
B
P(encoded)
v(x,i)
t_0
time
$= \begin{cases} 1 - e^{-v_x(t-t_0)}, \text{if } t > t_0 \\ 0, \quad \text{else} \end{cases}$
$v_x = \sum_{i \in R} v(x,i)$
$v(x,i) = \eta(x,i)\beta_i \frac{w_x}{\sum_{z \in S} w_z}$
$w_x = \sum_{j \in R} \eta(x,j)\pi_j$
C
SOA
+
+ B
A + B
time
D
P(LVF first)
-SOA
0
+SOA
E
$P_{\mathrm{LVF}^{1st}}(v_{\mathrm{LVF}}, v_{\mathrm{RVF}}, SOA)$
$= \begin{cases} 1 - e^{-v_{\mathrm{LVF}}|SOA|} + e^{-v_{\mathrm{LVF}}|SOA|} \left(\frac{v_{\mathrm{LVF}}}{v_{\mathrm{LVF}}+v_{\mathrm{RVF}}}\right) \text{ for } SOA < 0 \\ e^{-v_{\mathrm{RVF}}|SOA|} \left(\frac{v_{\mathrm{LVF}}}{v_{\mathrm{LVF}}+v_{\mathrm{RVF}}}\right) \text{ for } SOA \geq 0 \end{cases}$
F
v_{LVF}
$\mu_{v_{\mathrm{LVF}}}$
$\sigma_{v_{\mathrm{LVF}}}$
v_{RVF}
$\mu_{v_{\mathrm{RVF}}}$
$\sigma_{v_{\mathrm{RVF}}}$

Figure 3: Components of formal cognitive modeling. (A) Metaphorical model in which two stimuli race for encoding into the VSTM (visual short-term memory) and (B) its formal description according to Bundesen's theory of visual attention. (C) The temporal-order judgment task (SOA = stimulus onset asynchrony). (D) Typical data pattern of temporal-order judgments (LVF = left visual field). (E) Formal model of (D) based on (C). (F) Part of the hierarchical graphical model and exemplary group-level estimates (μ = mean; σ = standard deviation).

A TOJ model can be derived from TVA by turning the probabilities of one stimulus' encoding finishing before time t (as in Figure 2 Part B) into the probability of one stimulus being encoded earlier as another one (see Figure 2 Parts (D) and (E)). Thereby, this model relates the data to assumed processes and parameters, most importantly v, the assumed rate of processing (which, depending on the experimental conditions, can be broken down further in the components mentioned above).

We combined a TVA-based psychometric function with hierarchical Bayesian estimation (Figure 3(F)). What can we then learn about the alleged LVF advantage? To ask how Matthews and Welch's (2015) data is explained by TVA we fitted them with the model. The LVF bias is reflected in a higher processing rate for LVF stimuli $v_{\mathrm{LVF}} = 8$ Hz compared to the RVF $v_{\mathrm{RVF}} = 6$ Hz. In TVA, the overall processing rate is then $C = 14$ Hz ($v_{\mathrm{LVF}} + v_{\mathrm{RVF}}$) and the left-side attentional weight $w_{\mathrm{LVF}} = 0.57$ ($v_{\mathrm{LVF}}/C = w_{\mathrm{LVF}}$) is increased. What can cause higher attentional weights? TVA tells us (cf. Figure 3(C)) that higher weights result from stronger visual evidence for a target (η), a higher pertinence (π), or a stronger bias (β). Because Matthews and Welch (2015) used targets with constant visual evidence and the same importance to report them, η and β cannot vary, thus the difference must be in π. In TVA, π modulates the filtering of stimuli based on their attributes (here locations) and thus represents a spatial attention effect.

Because TVA is applied to various research questions and domains – with the same theoretical model – the estimates from re-fitting Matthews and Welch's (2015) data can be put into perspective: The attentional weight w_{LVF} of 0.57 is typical for attention shifts (caused by, for instance, increased conspicuousness), but the overall rate C of 15 Hz is exceptionally low compared to rates we usually find (see Tünnermann, 2016, p. 153–154, for an overview of typical TVA estimates). Because C depends on the visual evidence, it is likely that the faint gratings used by Matthews and Welch (2015) have low visual impact. Another difference between this experiment and many TVA-based TOJs is that both stimuli were always presented in one hemisphere whereas we often use judgments across hemispheres. If both hemispheres have different attentional weights, the effect should also occur across hemispheres. We thus decided to replicate Matthews and Welch's (2015) finding in our usual TOJ procedure with our usual stimuli to confirm the phenomenon in general and the predictions sketched above.

Because of TVA's formal theoretical model, we can make concrete quantitative predictions how the phenomenon should show up in the new experiment: If we assume an overall processing rate of 60 Hz (which we expect based on the experience with our stimuli) and take the attentional weight of $w_{\mathrm{LVF}} = 0.57$ estimated from Matthews and Welch' 2015 data, we can predict the left-side processing rate $v_{\mathrm{LVF}} = C \cdot w_{\mathrm{LVF}} = 60\ \mathrm{Hz} \cdot 0.57 = 34.2\ \mathrm{Hz}$ and the right-side rate as $v_{\mathrm{RVF}} = C \cdot (1 - w_{\mathrm{LVF}}) = 60\ \mathrm{Hz} \cdot 0.43 = 25.8\ \mathrm{Hz}$ (cf. Figure 2(B)). Note that instead of this quick calculation we could also generate expected distributions, taking the typical dispersions of the parameters (and group-level parameters) into account. We could also then fit multiple such simulations to estimate the power of our new experiment. These informative steps are possible because of the graphical Bayesian data model. However, in the present case we only predicted the expected parameter values based on the quick calculations above.

We conducted bilateral TOJs with 17 participants and obtained the processing rate posteriors depicted at the bottom of Figure 2(D). The estimates are very close to the predictions, with $v_{\mathrm{LVF}} = 35.48$ Hz (maximum of the posterior distribution; 34.2 Hz predicted) and $v_{\mathrm{RVF}} = 24.55$ Hz (25.9 predicted). This replicates the LVF advantage with a weight of $w_{\mathrm{LVF}} = 0.59$ and an overall rate $C = 60$ Hz. Admittedly, predictions rarely turn out as accurately as this one. With its consistent results the present experiment provides further support for TVA as a theory and the TVA-based TOJ model (theoretical model). It also provides a new perspective on the LVF advantage, which now appears as an attentional phenomenon rooted in filter criteria (TVA's π). Because πs are adjustable and task-dependent, the phenomenon could result from biases in the typical reading direction. It remains open how a detailed explanation may look (how do the πs change?) or whether alternatives are more likely.

5 Conclusion

An important point that our examples make is the following: When we set up formal models that provide well-defined connections in Bailer-Jones's (2009) framework, data and theory can be explored relatively freely. For instance, we ask questions like "What does the data show according to our model?" and make decisions for follow-up analyses contingent on outcomes not anticipated in advance. In the process of this, much can be learned about the problem under investigation and the applied theory in general. Intriguingly, this may sound just like how research is typically conducted or at least like how researchers typically would like to proceed. However, there is a difference, and it is an important one. The typical “effect-based” studies, those that confirm the existence (or rather reject the absence) of phenomena with no close link to theory beyond a tool-box frequentist null model do not warrant this degree of freedom. Researchers have to commit in advance to statistical tests (including many implicit assumptions). Positive results often provide little insight beyond the existence of the phenomenon under certain conditions without any formal link to related problems. What researchers take home from negative results is even thinner. Of course they inform the researcher’s intuition and thereby influence subsequent studies, but none of this happens in a form that allows to effectively accumulate knowledge. We therefore suggest to turn to formally specified models derived from theory combined with Bayesian statistics.

Optimally, when much research is conducted in this manner, it may be possible to link models from different domains. This may still be a long – but as we believe an enlightening – way to go.

References

Bailer-Jones DM (2009) Scientific Models in Philosophy of Science. University of Pittsburgh Press, Pittsburgh (USA). DOI: 10.2307/j.ctt5vkdnq.

Ballard T, Palada H, Griffin M, Neal A (2019) An Integrated Approach to Testing Dynamic, Multilevel Theory: Using Computational Models to Connect Theory, Model, and Data. DOI: 10.1177/1094428119881209.

Breiman L (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199–231, Zhang CH, Sriram T, Kelly P (eds), The Institute of Mathematical Statistics (IMS), Bethesda (USA). DOI: 10.1214/ss/1009213726.

Bundesen C (1990) A Theory of Visual Attention. Psychological Review 97(4):523, American Psychological Association (APA), Washington D.C. (USA). DOI: 10.1037/0033-295x.97.4.523.

Bundesen C, Vangkilde S, Petersen A (2015) Recent Developments in a Computational Theory of Visual Attention (TVA). Vision Research 116:210–218, Elsevier B.V., Amsterdam (The Netherlands). DOI: 10.1016/j.visres.2014.11.005.

Buntins M, Buntins K, Eggert F (2017) Clarifying the Concept of Validity: From Measurement to Everyday Language. Theory & Psychology 27(5):703–710, SAGE Publications, Thousand Oaks (USA). DOI: 10.1177/0959354317702256.

Carlson T, Goddard E, Kaplan DM, Klein C, Ritchie JB (2018) Ghosts in Machine Learning for Cognitive Neuroscience: Moving from Data to Theory. NeuroImage 180(A):88–100, Elsevier B.V., Amsterdam (The Netherlands). DOI: 10.1016/j.neuroimage.2017.08.019.

Chambers C (2017) The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice. Princeton University Press, Princeton (USA). DOI: 10.2307/j.ctvc779w5.

Coelho CM, Cloete S, Wallis G (2010) The Face-in-the-crowd Effect: When Angry Faces are just Cross(es). Journal of Vision 10(1):1–14, Watson AB, Brainard DH, Gegenfurtner K, Mamassian P, Rosenholtz R (eds), Association for Research in Vision & Ophthalmology (ARVO), Rockville (USA). DOI: 10.1167/10.1.7

Cummins R (2000) "How does it work?" versus "What are the laws?": Two Conceptions of Psychological Explanation. In: Explanation and Cognition, Keil F, Wilson RA (eds), Keil F, Wilson RA (eds). MIT Press, Cambridge (USA), chap. 5, pp. 117–144.

Dienes Z (2008) Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Palgrave-Macmillan, Houndmills (United Kingdom).

Dienes Z (2011) Bayesian Versus Orthodox Statistics: Which Side are you on? Perspectives on Psychological Science 6(3):274–290, Association for Psychological Science (APS), Washington, D.C. (USA). ISSN: 1745-6916, DOI: 10.1177/1745691611406920.

Efron B (2005) Bayesians, Frequentists, and Scientists. Journal of the American Statistical Association 100(469):1–5, Taylor & Francis, London (United Kingdom). DOI: 10.1198/016214505000000033.

Gelman A, Shalizi CR (2013) Philosophy and the Practice of Bayesian Statistics. British Journal of Mathematical and Statistical Psychology 66(1):8–38, Cheng Y (ed), John Wiley & Sons, on behalf of the British Psychological Society. DOI: 10.1111/j.2044-8317.2011.02037.x

Gigerenzer G (2004) Mindless Statistics. The Journal of Socio-Economics 33(5):587–606, Elsevier B.V., Amsterdam (The Netherlands). DOI: 10.1016/j.socec.2004.09.033.

Hacking I (2001) An Introduction to Probability and Inductive Logic. Cambridge University Press, New York (USA). ISBN: 978-0-521775-01-4, DOI: 10.1017/CBO9780511801297.

Horstmann G, Scharlau I, Ansorge U (2006) More Efficient Rejection of Happy Than of Angry Face Distractors in Visual Search. Psychonomic Bulletin & Review 13(6):1067–1073, Hickok G (ed), Springer, The Psychonomic Society, Inc. DOI: 10.3758/BF03213927.

John LK, Loewenstein G, Prelec D (2012) Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. Psychological Science 23(5):524–532, Association for Psychological Science (APS), Washington, D.C. (USA). DOI: 10.1177/0956797611430953.

Jäkel F, Schölkopf B, Wichmann FA (2008) Generalization and Similarity in Exemplar Models of Categorization: Insights from Machine Learning. Psychonomic Bulletin & Review 15(2):256–271, Hickok G (ed), Springer, The Psychonomic Society, Inc. DOI: 10.3758/PBR.15.2.256.

Kennett MJ, Wallis G (2019) The Face-in-the-crowd Effect: Threat Detection Versus Iso-feature Suppression and Collinear Facilitation. Journal of Vision 19, Watson AB, Brainard DH, Gegenfurtner K, Mamassian P, Rosenholtz R (eds), Association for Research in Vision & Ophthalmology (ARVO), Rockville (USA). DOI: 10.1167/19.7.6

Korb KB (2004) Introduction: Machine Learning as Philosophy of Science. Minds and Machines – Journal for Artificial Intelligence, Philosophy and Cognitive Science 14(4):433–440, Taddeo M (ed), Springer Nature B.V., Kluwer Academic Publishers. ISSN: 1572-8641, DOI: 10.1023/B:MIND.0000045986.90956.7f.

Lewandowsky S, Farrell S (2011) Computational Modeling in Cognition: Principles and Practice. SAGE Publications, Thousand Oaks (USA). URL: https://us.sagepub.com/en-us/nam/computational-modeling-in-cognition/book233316.

Little RJ (2006) Calibrated Bayes: A Bayes/Frequentist Roadmap. The American Statistician 60(3):213–223, Taylor & Francis, London (United Kingdom). DOI: 10.1198/000313006X117837.

Matthews N, Welch L (2015) Left Visual Field Attentional Advantage in Judging Simultaneity and Temporal Order. Journal of Vision 15(7):1–13, Watson AB, Brainard DH, Gegenfurtner K, Mamassian P, Rosenholtz R (eds), Association for Research in Vision & Ophthalmology (ARVO), Rockville (USA). DOI: 10.1167/15.2.7

Mayo DG (1996) Error and the Growth of Experimental Knowledge. University of Chicago Press, Chicago (USA). ISBN: 978-0-226511-97-9, URL: https://www.press.uchicago.edu/ucp/books/book/chicago/E/bo3637756.html.

Mayo DG (2016) Don't Throw out the Error Control Baby with the Bad Statistics Bathwater: A Commentary. URL: https://errorstatistics.com/2016/03/07/dont-throw-out-the-error-control-baby-with-the-bad-statistics-bathwater/ [accessed 2019-11-23].

Moore J (2015) Pragmatism, Mathematical Models, and the Scientific Ideal of Prediction and Control. Behavioural Processes 114(Supplement C):2–13, Elsevier B.V., Amsterdam (The Netherlands). DOI: 10.1016/j.beproc.2015.01.007.

Nordfang M, Staugaard C, Bundesen C (2018) Attentional Weights in Vision as Products of Spatial and Nonspatial Components. Psychonomic Bulletin & Review 25(3):1043–1051, Hickok G (ed), Springer US, The Psychonomic Society, Inc. DOI: 10.3758/s13423-017-1337-1.

Oakes MW (1986) Statistical Inference. Epidemiology Resources Incorporated, Newton Lower Falls (USA). ISBN: 978-0-917227-04-2.

Peebles D, Cooper RP (2015) Thirty Years After Marr's vision: Levels of Analysis in Cognitive Science. Topics in Cognitive Science 7(2):187–190, Gray WD (ed), John Wiley & Sons, on behalf of the Cognitive Science Society, Inc. DOI: 10.1111/tops.12137.

Rodgers JL (2010) The Epistemology of Mathematical and Statistical Modeling: A Quiet Methodological Revolution. American Psychologist 65(1):1–12, American Psychological Association (APA), Washington, D.C. (USA). DOI: 10.1037/a0018326.

Rouder J, Morey R, Wagenmakers EJ (2016) The Interplay between Subjectivity, Statistical Practice, and Psychological Science. Collabra: Psychology 2(1), University of California Press, Society for the Improvement of Psychological Science, Berkeley (USA). DOI: 10.1525/collabra.28.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t-Tests for Accepting and Rejecting the Null Hypothesis. Psychonomic Bulletin & Review 16(2):225–237, Hickok G (ed), Springer, The Psychonomic Society, Inc. DOI: 10.3758/PBR.16.2.225.

Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic Programming in Python using PyMC3. PeerJ Computer Science 2:e55, Elkan C (ed). DOI: 10.7717/peerj-cs.55.

Savage RA, Lipp OV, Craig BM, Becker SI, Horstmann G (2013) In Search of the Emotional Face: Anger Versus Happiness Superiority in Visual Search. Emotion 13(4):758–768, Pietromonaco PR (ed), American Psychological Association (APA), Washington, D.C. (USA). DOI: 10.1037/a0031970.

Senge R, Bösner S, Dembczyński K, Haasenritter J, Hirsch O, Donner-Banzhoff N, Hüllermeier E (2014) Reliable Classification: Learning Classifiers that distinguish Aleatoric and Epistemic Uncertainty. Information Sciences 255:16–29, Pedrycz W, Wang PP (eds), Elsevier B.V., Amsterdam (The Netherlands). DOI: 10.1016/j.ins.2013.07.030.

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22:1359–1366, Association for Psychological Science (APS), Washington, D.C. (USA). DOI: 10.1177/0956797611417632.

Suppes P (1966) Models of Data. In: Studies in Logic and the Foundations of Mathematics, Nagel E, Suppes P, Tarski A (eds), Logic, Methodology and Philosophy of Science, vol. 44, Nagel E, Suppes P, Tarski A (eds). Elsevier B.V., Board of Trustees of the Leland Stanford Junior University, Amsterdam (The Netherlands), pp. 252–261. DOI: 10.1016/S0049-237X(09)70592-0.

Taagepera R (2008) Making Social Sciences more Scientific: The Need for Predictive Models. Oxford University Press, Oxford Press Scholarship Online, Oxford (United Kingdom). ISBN: 978-0-199534-66-1, DOI: 10.1093/acprof:oso/9780199534661.001.0001.

von Toussaint U (2011) Bayesian Inference in Physics. Reviews of Modern Physics 83(3):943–999, American Physical Society (APS), College Park (USA). DOI: 10.1103/RevModPhys.83.943.

Treisman A, Gormican S (1988) Feature Analysis in Early Vision: Evidence from Search Asymmetries. Psychological Review 95(1):15–48, American Psychological Association (APA), Washington D.C. (USA). DOI: 10.1037/0033-295x.95.1.15.

Tünnermann J (2016) On the Origin of Visual Temporal-Order Perception by Means of Attentional Selection. PhD thesis, Paderborn University, Faculty of Arts and Humanities Psychology. URL: `https://core.ac.uk/download/pdf/50521662.pdf`.

Tünnermann J, Scharlau I (2019) Left vs. Right Visual Field: Assessment of Hemifield Processing Speed Differences with the Theory of Visual Attention [in preparation].

Tünnermann J, Krüger A, Scharlau I (2017) Measuring Attention and Visual Processing Speed by Model-based Analysis of Temporal-order Judgments. Journal of Visualized Experiments 119:e54856. DOI: 10.3791/54856.

Watanabe S (2010) Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research 11:3571–3594, Association for Computing Machinery (ACM), New York (USA). URL: `http://dl.acm.org/citation.cfm?id=1756006.1953045`.

Yarkoni T, Westfall J (2017) Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning. Perspectives on Psychological Science 2:1100–1122, SAGE Publications, Association for Psychological Science (APS), Washington, D.C. (USA). DOI: 10.1177/1745691617693393.