Beyond the Null Ritual

Formal Modeling of Psychological Processes

Julian N. Marewski and Henrik Olsson

Max Planck Institute for Human Development, Berlin, Germany

Abstract. Rituals shape many aspects of our lives, and they are no less common in scientific research than elsewhere. One that figures prominently in hypothesis testing is the null ritual, the pitting of hypotheses against chance. Although known to be problematic, this practice is still widely used. One way to resist the lure of the null ritual is to increase the precision of theories by casting them as formal models. These can be tested against each other, instead of against chance, which in turn enables a researcher to decide between competing theories based on quantitative measures. This article gives an overview of the advantages of modeling, describes research that is based on it, outlines the difficulties associated with model testing, and summarizes some of the solutions for dealing with these difficulties. Pointers to resources for teaching modeling in university classes are provided.

Keywords: null hypothesis significance testing, formal models, model selection

> The arbitrariness of the .05 α level gives the author some flexibility in interpreting a *p*-value as indicating evidence against the null hypothesis. However, unless converging evidence is presented, effects that only approach this level (e.g., at $p < .06$ or even $p < .1$) should be interpreted with considerable caution. *Cognition* will require its authors to adhere to the convention that an effect *described* as 'statistically significant' must have a *p*-value below .05 (for better or for worse, this is the current convention).
> (Altmann, 2007, p. 6)

– *A village in Bavaria, Germany, at the beginning of the 19th century:* Whenever men sleep badly, they attribute this to the *Trut*, a ghost that visits them at night. To please this spirit, secret rituals are performed. When occasional visitors to the village challenge this behavior as ridiculous, the villagers reply that even their grandparents, long ago, followed the practice, otherwise they could not sleep. It seems impossible to eradicate the Trut rituals from the village.
– *Papua New Guinea, in the 1940s:* Some of the Eastern tribes sprinkle their blood on the soil each spring. Their reason: They just do whatever their ancestors did, otherwise something terrible will happen. This ritual is hard to abolish.
– *University departments, early 21st century:* Many academic psychologists' lives revolve around a small number, .05, which plays a crucial role in the way they test theories: They reject or accept their hypotheses based on this value. Researchers report it because it is always reported, and if they do not report it, then they must fear being unable to publish their work – an apprehension that is enforced by journals' editorial policies. Even if they are mindless, rituals are hard to escape. They shape our lives, in science and beyond.

This article is about how to overcome a prominent ritual involved in hypothesis testing in psychology: The null ritual, or *null hypothesis significance testing*, which pits one's unspecified hypothesis against "chance," that is, against a null hypothesis postulating a "zero correlation" or "no differences between two population means," rather than against a competing hypothesis. Its origins lie in parapsychology and education, where there was little interest in testing theories and more in detecting effects greater than those of chance (Danzinger, 1990). As far back as 40 years ago, editors of major psychology journals, such as A.W. Melton (1962), required this ritual to be carried out in order for a paper to be accepted for publication. As illustrated by the recent editorial in *Cognition*, cited above, the picture has changed little today – despite repeated demonstrations that null hypothesis significance testing is an incoherent methodology for statistical inference that is detrimental for the progress of psychology as science (e.g., Cohen, 1994; Edwards, Lindman, & Savage, 1963; Gigerenzer, 1991; Gigerenzer & Murray, 1987; Meehl, 1978; Morrison & Henkel, 1970; Schmidt, 1996).

This article is arranged as follows: First, we will argue that one way to avoid following this ritual consists of increasing the precision of theories by casting them as formal models. Second, we will give an example of a research program that is based on models. Third, we will point out difficulties associated with model tests and report a range of methods for coping with these difficulties. The article concludes with pointers to resources for teaching modeling in university classes.

DOI 10.1027/0044-3409.217.1.49

Beyond the Null Ritual

Rituals can be characterized by a range of attributes (e.g., Dücker, 2007) including (1) repetitions of the same action, (2) fixations on special features such as numbers, (3) anxieties about punishments for rule violations, and (4) wishful thinking. According to Gigerenzer, Krauss, and Vitouch (2004), each of these characteristics is reflected in null hypothesis significance testing: A repetitive sequence, a focus on the 5% (or 1%) level, fear of sanctions by journal editors, and wishful thinking about results (i.e., the meaning of the *p*-value). In its most extreme form, the null ritual reads as follows:

1. Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis.
3. Always perform this procedure.
   (Gigerenzer et al., 2004, p. 2)

Since the institutionalization of this ritual in psychology (around 1955, see Gigerenzer et al., 1989), several alternatives have been proposed to replace or supplement it. Most of these suggestions tackle the issue from a methodological perspective, focusing on the way data are analyzed. The prominent ones advocate the use of effect size measures (e.g., Rosnow & Rosenthal, 2009), confidence intervals (e.g., Cumming & Fidler, 2009; Loftus & Masson, 1994), graphical data displays (e.g., Masson & Loftus, 2003), and Bayesian hypothesis testing (e.g., Berger, 1985; Edwards et al., 1963; Wagenmakers, 2007). Other alternatives are meta-analysis (e.g., Schmidt, 1992), exploratory data analyses (e.g., Tukey, 1977), and resampling methods (e.g., Efron & Tibshirani, 1993). In fact, in the past some psychology journals – including *Cognition* – have encouraged the reporting of power, effect size measures, graphical data displays, or other information beyond *p*-values (e.g., Altmann, 2007; Loftus, 1993; see also Wilkinson and the Task Force on Statistical Inference, 1999). Possibly because of the writings of these authors and many others (e.g., Sedlmeier & Gigerenzer, 1989), the null ritual as described above may, thus, no longer accurately portray the practices of today's psychologists. However, this is not to say that the ritual has disappeared. Quite the contrary: Encouraged by the wide acceptance of statistical tools built around the general linear model, much psychological theorizing is still obsessed with formulating and testing hypotheses about simple differences – significant "main effects" and "interaction effects" populate our discipline's scientific outlets.

What are the reasons for the continuation of this state of affairs? Primarily, many psychological theories are simply too weak to do much other than make predictions about the direction of an effect. Such theories hardly deserve better tests than the null ritual, perhaps coupled with some statistical procedure based on the general linear model, and – in the best case – measures of effect size and power. Therefore, we will not present yet another alternative to the ritual that concentrates on the way data are analyzed. Rather, we will focus on directly addressing the vague nature of many psychological theories: One good way to make theories more precise is to cast them as formal models. In doing so, researchers can move beyond the problems of null hypothesis significance testing, and simple difference searching. Before we discuss in more detail why it is important to formally specify theories, a few introductory comments are warranted.

What Is a Model?

In the broadest sense, a model is a simplified representation of the world that aims to explain observed data. Countless verbal, that is, *informal*, explanations of psychological phenomena fit this definition. In a more narrow sense, a model is a *formal* instantiation of a theory that specifies the theory's predictions, for example, in mathematical equations or computer code.[1] This category also includes statistical tools, such as structural equation or regression models. Unless one believes that the mind works like a regression analysis or other statistical procedure, statistical tools are not typically meant to mirror the workings of psychological mechanisms, say, those determining how a person processes information (but see Gigerenzer, 1991, for examples of theories inspired by statistical tools). In this article, we mainly discuss *algorithmic level theories* (Marr, 1982), that is, formal instantiations of theories that are designed with the goal of reflecting psychological processes, although similar arguments could also be made for *computational level theories* that aim to explain the functional goals of behavior.

What Is the Scope of Modeling?

Modeling is not meant to be applied equally to all research questions. Each method has its specific advantages and disadvantages (for overviews on the pros and

[1] Sometimes mathematical models are contrasted with computer models. For instance, according to Fum, Del Missier, and Stocco (2007) mathematical models can be used to describe a phenomenon but they do not reproduce it, whereas computer models can produce observable behavior. Here, we do not distinguish between mathematical and computer models. Mathematical models can be implemented in computer code and sometimes computer models can be expressed in terms of mathematical equations. Similarly, sometimes one can derive a theory's predictions through both mathematical computation and computer simulation; however, for more complex theories one typically has to rely on computer simulations. But as Luce (1999) puts it: "Everyone agrees that when analysis is possible, it is far more satisfactory than numerical simulations" (p. 732).

cons of null hypothesis testing and some of its alternatives and supplements, see e.g., Frick 1996; Nickerson, 2000) and should be seen as a tool tailored to specific problems that researchers should pull out of their methodological toolbox whenever it is most advantageous to do so. For instance, when investigating which of two treatments for depression is more powerful, it might be pointless to model the processes underlying each treatment's effectiveness. Instead, it may be better to examine differences between groups of patients receiving one treatment or the other, using effect size measures, and meta-analyses as research tools.

Modeling helps researchers answer involved questions and understand complex phenomena. In the natural and social sciences, models are widely used: Economists, biologists, physicists, and many others employ them to understand economic behavior (e.g., von Neuman & Morgenstern, 1944), evolution (e.g., Maynard Smith, 1982), or emergence (e.g., Holland, 1998). In psychology, modeling is especially suited for basic and applied research about the cognitive system. Here, it has been used to investigate a large variety of phenomena, ranging from orthographic processing in visual word recognition (e.g., Grainger & Jacobs, 1996), to strategy selection (e.g., Rieskamp & Otto, 2006), to name just two. Modeling is also helpful beyond cognitive science. For example, in social psychology, models can be used to understand how segregation evolves in social networks (e.g., Schelling, 1969).

At the conclusion of these introductory remarks, a comment about the scope of this article is warranted. Modeling is a complicated enterprise that requires much skill and knowledge. Many have written about its merits and complications (e.g., Lewandowsky, 1993; Pitt, Myung, & Zhang, 2002; Roberts & Pashler, 2000; Sedlmeier & Renkewitz, 2007). Our intention, in this limited space, is to provide a rough orientation only.

Advantages of Formally Specifying Theories

In what follows, we will discuss four closely interrelated benefits of increasing the precision of theories by casting them as models. First, we will argue that models allow the design of strong tests of theories. Second, we will highlight that they can also sharpen research questions. Third, we will discuss how models can lead beyond theories built on the general linear model. Fourth, we will suggest that modeling helps to address real-world problems.

Designing Strong Tests of Theories

Models provide the bridge between theories and empirical evidence. They enable researchers to make competing quantitative predictions, which in turn lead to *strong* comparative tests of theories. That is, with modeling, academic psychologists can hope to identify better theoretical explanations for data. In doing so, they can come up with precise criteria for what "better" means, namely, quantitative measures of how much a model's predictions deviate from empirical observations, which can be, for instance, the R^2 or mean squared deviations between different models' quantitative predictions and the data. This precision makes comparative tests of theories strong, because any quantitative prediction can be systematically better or worse than any other.[2]

At the same time, as soon as one starts to compare quantitative predictions from different models, the use of null hypothesis testing can become inappropriate or meaningless (see Frick, 1996). To illustrate, it has long been acknowledged in the structural equation modeling literature that the χ^2 statistic is not an appropriate model evaluation measure as it will almost always be significant if it is applied to sufficiently large data sets (indicating a lack of fit; e.g., Bentler, 1990). In the section on model selection, we will give an overview of some of the techniques that are more appropriate than significance tests for selecting between formal explanations of data.

Sharpening Research Questions

Null hypothesis tests are often used to evaluate verbal, informal theories. However, if such theories are underspecified, then they can be used post hoc, to "explain" almost any observed empirical pattern. In the worst case, they become *one-word explanations*, labels that are broad in meaning and, hence, vague, and that, therefore, provide little or no specification of the underlying mechanisms or theoretical structure (see e.g., Gigerenzer, 1996, 1998a). Consider the *representativeness heuristic* in the field of judgment and decision making (Kahneman & Tversky, 1972). A probability assessed by this decision strategy, say whether a newly encountered animal is a dog, is derived from how representative this animal is of the target category – in this case, dogs. Even though this heuristic was proposed as an "explanation" for behavior over 30 years ago, until recently, exactly *how* the category is represented or *how* representativeness is derived was not defined. This vagueness made it possible to apply the notion of representativeness to a wide range of phenomena, such as misperception of regression, the conjunction fallacy, and base-rate ne-

[2] We use the term "predicting" (or "prediction") to refer to situations in which a model's free parameters are fixed such that they cannot adjust to the data on which the model is tested. In contrast, we use the term "fitting" to refer to situations in which a model's parameters are allowed to adapt to the test data. We use the term "quantitative" prediction in a broad sense to refer to both categorical and numerical statements. Note that there are also ways to evaluate models based on qualitative predictions (see Pitt, Kim, Navarro, & Myung, 2006).

glect. At the same time, this lack of formal specification made it hard – if not impossible – to test (see e.g., Ayton & Fischer, 2004). In fact, after the definition of the heuristic was finally strengthened (see Kahneman & Frederick, 2002), a number of studies found that models assuming different psychological processes outperform this heuristic in predicting people's behavior (e.g., Nilsson, Olsson, & Juslin, 2005).

Much research on heuristics beyond Kahneman and Tversky's program has formally specified decision-making processes (e.g., Payne, Bettmann, & Johnson, 1993). For instance, in the *fast and frugal heuristics research program* (e.g., Gigerenzer, Todd, and the ABC Research Group, 1999), such specifications have allowed researchers to ask counterintuitive questions such as whether and when ignoring information can help a person to make more accurate inferences than integrating all that is available (e.g., Brighton, 2006; Gigerenzer & Goldstein, 1996; Hogarth & Karelaia, 2007). Importantly, formal quantitative predictions are often not easy to grasp by intuitive reasoning. Frequently, the predictions being made by models can only be understood by running computer simulations. This holds true even if the models themselves seem to be simple: As is well known in the world of physics, a few very simple mechanisms can give rise to extremely complex data patterns. In short, often it is only when one starts modeling that one learns what a theory really predicts, and what it cannot account for (Hintzman, 1991), pointing to overlooked research problems. Thus, the goal of modeling is not only to find out which of competing explanations for data is preferred, but also to sharpen the questions to be asked.

Going Beyond Linear Theories

Many null hypothesis significance tests are only suited for the evaluation of simple hypotheses, for instance about linearly additive effects. This, in turn, encourages the formulation of hypotheses that can be subjected to such tests. Sometimes corresponding theories even read like a statistical procedure itself: A postulation of a few main effects, and interactions – an ANOVA-type theory. Gigerenzer (1991) subsumed this approach to theory building under the label *tools-to-theories*: Researchers take available tools, such as the analysis of variance (ANOVA), and turn them into a psychological explanation for data. A prominent example is attribution theory (Kelley, 1967). The theory assumes that just as experimenters use ANOVAs to infer causal relations between two variables, outside the lab people infer causal relations by unconsciously doing the same calculation. However, even though the general linear model – of which ANOVA is a special case – represents a precise methodological tool, it might not be the best starting point for theory building (see e.g., Gigerenzer & Richter, 1990). To illustrate, Loftus, Oberg, and Dillon (2004) compared traditional linear theory with dimensional theory, which is related to conjoint measurement, multidimensional scaling, and other concepts. Using the *face inversion effect* (i.e., processing disadvantages are greater for inverted faces than for other visual stimuli) as an example, they showed that depending on whether one uses linear theory or dimensional theory one can come to very different conclusions about when this effect occurs. Put in their words, "off-the-shelf linear theory is highly limited in its usefulness as a foundation for conceptualizing psychological problems and analyzing psychological data" (p. 860). By relying on modeling techniques, ideas beyond the ANOVA-type can be expressed and tested side-by-side with linear explanations for data.

Using More Externally Valid Designs to Study Real-World Questions

Since the publication of Woodworth's (1938) book *Experimental Psychology*, a frequent experimental practice has entailed manipulating a few variables while keeping all others constant or varying them at random. Such factorial designs help to ask, for instance, whether observed data patterns are the result of chance, or whether there are main effects or interactions, pointing to causal relations between variables. It is possible that such designs, the general linear model and corresponding significance tests, as well as the ANOVA-type theories that can be evaluated with them, have mutually aided each other in becoming widely accepted in the psychological research community.

However, just as the general linear model and null hypothesis significance tests are often inadequate for conceptualizing and evaluating a theory, factorial designs can lead to tests of theories under conditions that have little to do with the world outside the laboratory, which is where the explanatory power of theories should be proved. As argued by Egon Brunswik more than 50 years ago, factorial designs can destroy the natural covariation of variables, making it difficult to generalize from them to a world where organisms exploit the confounded relations between different pieces of information (Brunswik, 1956). In fact, a lack of external validity may be one reason why much basic research in psychology is of little use in the applied, real world: No person can randomize the people she is interacting with, and no organism can disentangle correlations between different pieces of life-sustaining information.

Modeling, in turn, allows researchers to deal with natural confounds without destroying them; they can be built into the models – for example as if-then rules or equations (Sedlmeier & Renkewitz, 2007). There are different ways to design and test corresponding models. One is to go out into the world. Observations made in the field can be used to build a model that can be tested by predicting new observations. For instance, Dhami (2003) observed judges in London courts to examine how punitive decisions are made. Based on her findings she constructed different models of these juridical judgments, which she then validated

by predicting new observations and by testing the models against each other. She found that a simple heuristic provided the best formal explanation for the judges' behavior. Yet another way is to take the world into the laboratory – the old idea of *representative design* (Brunswik, 1956). This entails turning an experimental task itself into a model of the world, for instance by ensuring that a task reflects the statistical structure of information inherent in natural environments. Given that there may often be different ways in which the world can be represented in the lab, this approach can require deciding between competing models of the world, just as one has to decide between alternative models of the mind.[3]

To conclude, modeling provides ways to increase the precision of theories. In doing so, it helps researchers to quantify the explanatory power of their ideas, allowing them to select between competing accounts of data without having to rely exclusively on the null ritual. Formal explanations of data can be linear or nonlinear; they can be simple and easy to understand, or complex and difficult to grasp by intuitive reasoning. Their precision aids in going beyond factorial designs, which in turn can make it possible to address real-world problems.

Illustrating Further Benefits of Formal Modeling: An Example of a Modeling Framework

There are many research programs that are based on models (e.g., Gigerenzer et al., 1999; Raaijmakers & Shiffrin, 1981; Ratcliff, Van Zandt, & McKoon, 1999; Rumelhart, McClelland, & the PDP Research Group, 1986). In what follows, we will give one example, the ACT-R architecture (*adaptive control of thought – rational*; e.g., Anderson et al., 2004). Note that while we chose this theory, others could also have been used. One advantage of ACT-R is that it illustrates (a) how cumulative knowledge can grow over several decades by developing models, (b) how modeling can aid one's understanding of how different kinds of behavior emerge in concert, and (c) how models can help to integrate disciplines.

The ACT-R Architecture

ACT-R is a broad, quantitative theory of behavior that covers much of human cognition. It has been applied to a large variety of phenomena in basic and applied research, including, for example, probability learning (Gaissmaier, Schooler, & Rieskamp, 2006), driving behavior (Salvucci, 2006), and the teaching of mathematics (Ritter, Anderson, Koedinger, & Corbett, 2007).

Cumulative Theory Building

Meta-analysis can be used to show that the reliance on significance testing retards the growth of cumulative knowledge (see Schmidt, 1996). ACT-R, in turn, is a good example of how knowledge can systematically accumulate over time. Since the 1970s, ACT-R has been repeatedly modified in order to be able to account for new phenomena, a fact that is also reflected in the small changes in the theory's name (e.g., ACT, ACT*, ACT-R). For instance, ACT has its origins in the *human associative memory theory* (HAM, Anderson & Bower, 1973). This theory did not deal with the many different types of knowledge a person can have about the world. In 1976, Anderson suggested distinguishing between *declarative knowledge* (knowing that), which HAM dealt with, and *procedural knowledge* (knowing how), which HAM ignored. Based on ideas from Newell (e.g., 1973b), *production rules* were proposed to implement such procedural knowledge. (Productions are if-then statements; the "if" part of the rule states a condition that must be fulfilled in order for the action to be carried out, which is specified in the "then" part of the rule.) The result was a production system called ACTE, which was later replaced by ACT* (Anderson, 1983). Among other things, this new system incorporated assumptions about how production rules might be acquired. The development of ACT-R followed, which embodied the insight that the cognitive system can give rise to adaptive processing by being tuned into the statistical structure of the environment (see e.g., Anderson & Schooler, 1991).

As of this writing, the latest version of this theory is called ACT-R 6. It consists of a set of modules, each of which is devoted to a different behavioral activity. To illustrate, there is a goal module for keeping track of intentions, a declarative module for information retrieval from memory, a visual module for identifying objects in the visual area, and a manual module for executing motor commands, such as hand movements. As in the earlier versions, these modules are coordinated through a production system.

Understanding How Different Kinds of Behavior Emerge in Concert

ACT-R models can be precise enough to allow for computational modeling of *outcomes* and *processes*. For instance, in a two-alternative choice situation, say, whether to read this article or another one, an ACT-R model could predict both *which* alternative will be chosen, and *how* different

[3] Let us stress that factorial and more externally valid designs, as well as linear and nonlinear data analysis techniques and models, are complementary means to investigate psychological phenomena. For instance, often research will start out with a factorial design and then move to more realistic settings, or vice versa, thereby examining the generalizability of a model.

reasons will be processed to derive a decision. In doing so, ACT-R allows the investigator to predict at what point in time which processes occur in parallel and which do not. For example, a person may be able to process visual information at the same time she is executing a hand movement, but she may not remember 10 ideas simultaneously. Specifically, with ACT-R researchers can derive predictions of at least three kinds of data: (1) overt behavior, such as the outcomes of decisions; (2) the temporal aspects of the behavior, such as time involved in making a decision; and (3) the associated patterns of activity in the brain, as measured with functional magnetic resonance imaging (fMRI) scanners. This variety of precisely predictable behavior can render standard null hypothesis tests of ACT-R models obsolete.

Theory Integration Beyond Disciplinary Boundaries

There are at least two approaches to modeling, entailing two kinds of models: Those that intend to capture the aspects of an isolated task (e.g., the Stroop task) or phenomenon (e.g., memory retrieval), and those that *in addition* are integrated into an overarching architecture that formally specifies the assumptions of a broader theory (e.g., about cognition in general). The ACT-R architecture formally integrates theories of perception, action, and other behavioral activities, letting artificial boundaries between different fields of psychological expertise dissolve. For instance, Schooler and Hertwig (2005) used ACT-R as a platform to tie theories of memory to the fast and frugal heuristics approach to judgment and decision making (Gigerenzer et al., 1999). In computer simulations, they derived quantitative predictions as to when the forgetting of memory records can help a person using different inference heuristics to make accurate decisions, illustrating how cognitive capacities such as human memory interplay with people's decision strategies (see also, e.g., Gaissmaier, Schooler, & Mata, 2008). It is difficult to see how such precise integrations of different theories can be achieved by subjecting informal theories to null hypothesis tests.

In short, modeling can foster the growth of cumulative knowledge, reveal how different behavioral activities are concentrated, and help to integrate psychological disciplines. Needless to say, overarching architectures such as ACT-R can also serve as a platform for comparative theory evaluation – the topic of the next section.

How to Select Between Competing Formal Models: A Short Overview

Consider two models that compete as explanations for a behavior in a task. How can one decide which model provides a better explanation for the data? This comparison of alternative models is called *model selection*. Model selection can have various technical meanings in different fields, but for our purposes it suffices to say that it is the task of choosing a model from a set of potential models, given available data. In what follows, we will briefly discuss some of the pitfalls encountered when deciding between competing models and give a short overview of different approaches to model selection.

A number of model selection criteria are available (see Jacobs & Grainger, 1994, for an overview). These include standards for psychological plausibility, such as whether the computations postulated by a model are tractable in the world beyond the laboratory (e.g., Gigerenzer, Hoffrage, & Goldstein, 2008), and falsifiability, that is, whether the model can be proven wrong or be shown to explain everything, and hence, nothing. Other criteria address the number of assumptions the models make. For instance, one could ask which of many competing models accounts for the data in the simplest way (e.g., Myung & Pitt, 1997). Yet another benchmark is whether a model is consistent with overarching theories of cognition. Integrative architectures, such as ACT-R, can impose precise theoretical constraints on which models represent acceptable developments of a theory. Given our earlier emphasis on testing models of cognition under realistic conditions, for instance, by going out in the field, we also list proof of practical relevance and applicability to real-world problems as selection criteria. There is much in the worlds of practicing lawyers, doctors, or other professionals that can inform academic psychologists about the usefulness of a theory, helping basic research to be grounded in the world, rather than hovering way above it, in beautiful, but abstract and useless imaginary castles of psychological theorizing. Possibly the most widely used model selection criterion is a model's *descriptive adequacy* – which is the yardstick for model selection we will focus on in the remainder of this section: When two or more models are compared, the model that provides the smallest deviation from existing data, that is, the best *fit,* is favored over a model that results in a larger deviation from data.

Besides standard goodness-of-fit measures such as R^2 or mean squared error, significance tests have been traditionally used to determine if a model fits data well. For example, running an F test is a standard procedure for determining which of two nested regression models fits the data better. Similarly, the χ^2 statistic is also widely used for nested models. However, as mentioned above, null hypothesis significance testing is not an appropriate tool for choosing between models; given enough power, such tests will almost always be statistically significant. Furthermore, most significance-based test procedures for model selection are only applicable to nested models, and not to comparisons of formal theories that do not represent subsets of each other. Yet, perhaps the most serious limitation of model selection procedures that are based exclusively on significance tests or goodness-of-fit indices such as R^2 is that on their own, these procedures do

not address a fundamental problem of deciding between competing theories: *overfitting*.

The Problem of Overfitting

To conclude that one model provides a better account of data than another based on R^2 or other standard goodness-of-fit indices might be reasonable if psychological measurements were noise-free. In science, however, noise-free data are practically impossible to obtain. Hence, researchers are confronted with the problem of disentangling the variation in data that is the result of noise from the variation that is the result of the psychological process of interest. Goodness-of-fit measures alone cannot make this distinction. As a result, a model can end up overfitting the data, that is, it can capture not only the variance that results from the cognitive process of interest but also that from random error.

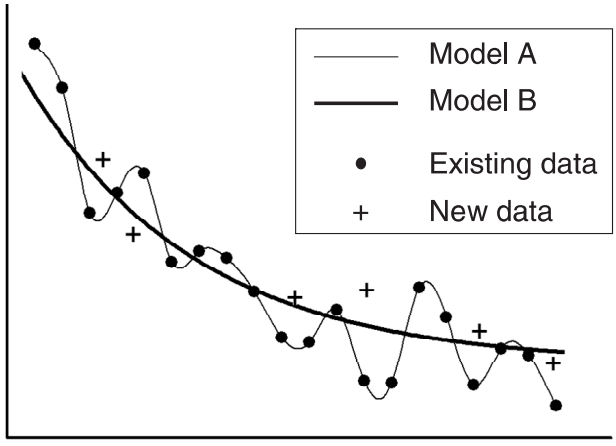

Figure 1. Illustration of how two models fit existing data (filled circles) and how they predict new data (pluses). Model A (thin line) overfits the data and is not as accurate in predicting new data as model B (thick line; see Pitt et al., 2002).

Figure 1 illustrates a situation in which one model, call it Model A (thin line), overfits existing data by chasing after idiosyncrasies in that data. This model fits the existing data (filled circles) perfectly but does a poor job of predicting new data (pluses). Model B (thick line), while not fitting the existing data as well as Model A, captures the main tendencies in that data and ignores the idiosyncrasies. This makes it better equipped to predict new observations, as can be seen from the deviations between the model's predictions and the new data, which are indeed smaller than the deviations for Model A.

The ability of a model to predict new data is called its *generalizability*, that is, the degree to which it is capable of predicting *all* potential samples generated by the same cognitive process, rather than to fit only a *particular* sample of existing data. The degree to which a model is susceptible to overfitting, in turn, is related to the model's *complexity* (Pitt et al. 2002), that is, a model's inherent flexibility that enables it to fit diverse patterns of data. Two factors contribute to a model's complexity: The number of free parameters it has, and how the parameters are combined in it – in other words, its *functional form*. The impact of many free parameters is illustrated in Figure 1, where the model that overfits the data (Model A) has more free parameters than the model that captures the main tendencies in the data (Model B). The impact of the functional form on the flexibility of a model's prediction can be illustrated by comparing Stevens's (1957) and Fechner's (1860/1966) models of the relationship between physical dimensions (e.g., the intensity of light, here called X) and their psychological counterparts (e.g., brightness, here called Y). In both models, there are two free parameters, a and b, but they are combined differently (Stevens's model: $Y = a\ X^{\ b}$; Fechner's model: $Y = a\ \ln[X + b]$). Townsend (1975) noted that Stevens's model is more complex than Fechner's model. Since it assumes that a power function relates the psychological and physical dimensions, Stevens's model can fit data that have negative, positive, and zero curvature. Fechner's model, in turn, can only fit data with a negative curvature because it assumes a logarithmic relationship.

The dilemma can be summarized in the following way. Increased complexity makes a model more likely to end up overfitting the data while its generalizability to new data decreases. At the same time, a model's generalizability can also increase positively with the model's complexity – but only to the point at which the model is complex enough to capture systematic variations in the data. Beyond that point, additional complexity can result in decreases in generalizability, because then the model may also start to absorb random variations in the data (Pitt et al., 2002). In short, a good fit to existing data does not necessarily imply good generalizability to new data, which can make it hard to tell which of two models provides a better explanation for data.

How to Select Between Models

The problem of overfitting is not limited to the selection between different theories of psychological processes, but applies equally to widely used statistical tools such as regression analyses. There are many different tools available for reducing the risk of selecting the wrong model. These model selection approaches can be roughly classified into *practical*, *simulation*, and *theoretical approaches* (see Shiffrin, Lee, Kim, & Wagenmakers, 2008).

Practical Approaches

Practical approaches rely on the intuition that in a comparison of models, the one that can predict unseen data better than other models should be preferred. Typically, corresponding procedures estimate how well models can generalize to all possible samples generated by the same process by dividing up available data into a *calibration* (or training) set and a

validation (or test) set. Each model's parameters are estimated on the calibration set. The resulting fixed parameters are then used to test the models on the validation set. The model that predicts the data best on the validation set (according to criteria such as mean squared error) is selected.

Most common is to use some form of cross-validation (Browne 2000; Stone, 1974, 1977). One general scheme is called *K-fold cross-validation*. Here, the data is partitioned into K subsets, and one of these K subsets is successively used as a calibration set and the remaining $K - 1$ subsets are used as the validation set. The overall prediction error is the average of the of prediction error on the K validation sets. As special cases, this scheme includes *split-half cross-validation* ($K = 2$) and *leave-one-out cross-validation* (where K equals the number of observations). Although variants of cross-validation are intuitively plausible and have high practical appeal, they have some drawbacks. For example, the balance between the size of the calibration set and the validation set is an open question, although this may be resolvable by simply taking all possible sizes into account. Perhaps a more fundamental limitation of most cross-validation methods is that they are not consistent; that is, as the number of observations increases, many cross-validation methods do not converge on the correct model, in the case that there is one (Stone, 1977).

Another practical way of dealing with the problem of overfitting consists of dispensing with as many free parameters as possible – either by fixing them or by designing simple models with few or no free parameters. ACT-R, for instance, comes with an array of free parameters that can make ACT-R models flexible in fitting data. However, rather than estimating parameters each time a study is run, many researchers use the default values for these parameters (i.e., values set by the ACT-R system) or try to estimate them in separate studies (see Newell, 1973a). In fact, testing models by pitting their respective out-of-study predictions against each other may sometimes be one of the best ways to perform model selection – especially when models are intended to be of practical value beyond describing psychological processes. For instance, a triage procedure developed in one hospital should also work in another. Other theories of cognition involve fewer free parameters. For example, several of the models that have been developed in the fast and frugal heuristics research program require no or only very few free parameters to be fitted (for a recent review, see Gigerenzer & Brighton, 2008).

Simulation Approaches

When one compares models with free parameters it can be difficult to ascertain, a priori, what the models actually predict, as the predictions are dependent on the specific values of their free parameters. By simulating the predictions of competing models for a specific task, one can gain insight into the behavior of the models and use the results to design the task to maximize the possibility of discriminating between the models. It might be that the models predict the same outcome for most items (e.g., a judgment) over most of their parameter spaces, but that there are some items for which they make divergent predictions. These items could then be included in the design of the task in order to be better able to distinguish between the models. A more advanced form of the simulation approach is called *landscaping* (e.g., Navarro, Pitt, & Myung, 2004). Here the focus is on the problem of *model mimicry*, which refers to a model's ability to fit not only data generated by its own process, but also data generated by some other model. By letting a model generate many data sets, and then fitting this model as well as competing ones to that data, one can evaluate the separability of models and the informativeness of data in distinguishing between them. Moreover, this way the relative flexibility of models in fitting different data can be determined, which in turn can be informative for assessing the models' risk of overfitting.

Theoretical Approaches

In most theoretical approaches to model selection, goodness-of-fit measures are combined with theoretically derived estimates of model complexity resulting in an estimate of generalizability. Overall, such estimates can usually be expressed as generalizability = goodness-of-fit + complexity. In most approaches, generalizability measures are based on the maximized log likelihood as a goodness-of-fit index. The complexity measure, in turn, takes different forms for different generalizability measures (for an overview see Pitt & Myung, 2000).

The two most widely used generalizability measures, or model selection criteria, are the *Akaike information criterion* (*AIC*; Akaike, 1973) and the *Bayesian information criterion* (*BIC*; sometimes called the *Schwarz information criterion*, see Schwarz, 1978). AIC, which is rooted in information theory, represents the complexity of a model as the number of parameters. BIC, which is rooted in Bayesian statistics, is also sensitive to the number of parameters in the model, but in addition, it takes the log of the sample size into account. As a result, BIC favors simpler models to a greater extent than AIC does (see e.g., Forster, 2000).

AIC and BIC are only sensitive to one dimension of complexity: The number of parameters. More advanced generalizability measures also take into account the functional form of a model's equation. Examples are *Bayesian model selection* (e.g., Myung & Pitt, 1997) and *minimum description length* (*MDL*; Pitt et al., 2002; see Grünwald, 2007, for a comprehensive treatment of MDL).

Choosing Between Model Selection Approaches

Choosing between different model selection approaches is not easy; they all have their pros and cons, and an approach

that works in one situation might not in another. For instance, the more advanced theoretical approaches such as Bayesian model selection and MDL outperform AIC, BIC, and practical approaches such as cross-validation in model recovery simulations (i.e., one model is used to generate data and other models are then fitted to that data; see e.g., Myung, Balasubramanian, & Pitt, 2000). However, applying these more advanced procedures usually requires a high level of mathematical ability on the part of the researcher, making it difficult for many investigators to rely on them in practice. Sometimes the nature of a model rules out certain approaches. To illustrate, when comparing extremely complex ACT-R models, it might not be possible to derive the equations necessary for Bayesian model selection or MDL. Instead, one may often have to rely on practical or simulation approaches. Also, the choice between simpler theoretical procedures is not straightforward. For example, mathematical and simulation results have fueled a long-running debate in the model selection literature about using AIC or BIC. Many researchers side with BIC, because it identifies the correct model when the number of observations approaches infinity, and in addition, it has outperformed AIC in many simulation studies (e.g., Wagenmakers, 2003). These findings, however, have been challenged, in particular by Burnham and Anderson (e.g., 2002). They argued that most of the simulation results are not relevant to realistic model selection problems.[4] In short, here we cannot give guidelines as to which method to choose. Rather, we recommend reporting the outcomes of model comparisons on as many selection criteria as possible, discussing the appropriateness of each.

Other Pitfalls of Model Selection

There are many other complications that can arise when designing and testing models. At the end of this section, let us highlight three of them. First, if precision is the major virtue of modeling, it can also be a curse. Modelers need to decide how to bridge the gaps between informal verbal descriptions of theories and formal implementations, which can lead to unintended discrepancies between theories and their various formal counterparts, otherwise known as the *irrelevant specification problem* (see Lewandowsky, 1993). Second, a problem that can arise in complex models is the *Bononi paradox*: When models become more complete and realistic they become less understandable and more opaque (see Dutton & Starbuck, 1971). For example, if one has a model of how the brain works and constantly adds more intricate layers of simulated neurons to this model, then it might end up as no more understandable than the workings of an actual brain. Third, there is the *identification problem*, namely, that for any behavior there may exist a universe of different models all of which are equally capable of reproducing and explaining the behavior (see Anderson, 1976). As a result it appears unreasonable to ask which of many models or theories is more "truthful"; rather, one needs to ask which model is better than another given a set of criteria, say, the models' practical relevance, simplicity, or usability. As Box (1979) put it, "All models are false, but some are useful" (p. 202).

Importantly, just as universes of functionally equivalent models may abound, there are an infinite number of vague, informal theories for which nobody will ever be able to decide whether one is better than another. In fact, just like their formal counterparts, informal theories are plagued by *all* of the aforementioned problems of modeling. In vague theories, these problems just do not become apparent, hence, one never learns to deal with them.

To summarize, there are a range of model selection criteria that allow researchers to pit competing theories against each other. Sometimes, however, different criteria may not unanimously favor the same model, making it difficult to determine which model is the best. Model selection is no ritual with fixed guidelines. Rather, it is an exercise that requires careful decision making on the part of the researcher. Moreover, model selection requires a range of skills that, as we will highlight next, should be taught to future generations of psychologists.

Beyond Methods 101: Teaching Formal Modeling of Psychological Processes

Is lack of skill one reason why few researchers pit competing models against each other, compared to the large number of psychologists who test informal theories against chance, fishing for main effects or interactions? There are a number of ways in which the null ritual might lose its prominence, including the teaching of alternative statistical procedures, for example, how to compute confidence intervals of various sorts. However, as we and others have argued (e.g., Gigerenzer, 1998b), replacing one statistical procedure with others may not be sufficient on its own. Instead, students must learn how to formulate precise, competing theories. In our view, methods education should, therefore, additionally encourage students to acquire modeling skills and offer corresponding math and programming courses.

One challenge for teachers is that few textbooks provide a comprehensive introduction to modeling for psychologists. There are some volumes that might be useful (e.g., Polk & Seifert, 2002), but these books need to be supple-

[4] For instance, according to Burnham and Anderson (2002), in model recovery simulations the model that generated the data is known, and, hence, perfectly recoverable, favoring BIC. In real model selection problems, however, the data generating process is unknown, and can at best only be approximated.

mented with other readings, some of which are cited above (e.g., Pitt et al., 2002). Another challenge for teachers is what software to use. Thorngate (2000) has argued that Matlab might act as a lingua franca for teaching simulation techniques. This program is easy to learn and not only allows the user to implement models of cognitive processes but also comes equipped with a range of data analysis tools. In addition, there are several open source Matlab clones (e.g., Scilab, http://www.scilab.org; Octave, http://www.gnu.org/software/octave). More advanced classes could also teach the basics for working with one or two architectures. For instance, there is a range of tutorials available for creating ACT-R models (http://act-r.psy.cmu.edu/).

Conclusion

If modeling is a good route to scientific progress, why then, do more researchers not engage in it? A pessimistic, almost cynical, answer to this question might look like this. Designing and testing models is an effortful and time-consuming enterprise that requires much skill. At the same time, the broad acceptance of using null hypothesis testing to examine vague theories in artificial study settings provides researchers with little incentive to elaborate their ideas by building models that are of theoretical value in the real world. Quite to the contrary, today academic psychologists can successfully make a career without even knowing how to come up with a formal explanation of data. Rituals shape our lives – sometimes for the worse, as in the case of the uncritical use of null hypothesis tests. Another, more optimistic, view is that with some effort most rituals can be overcome. All good science requires much skill, and with some training, modeling is not as painstaking as it sounds. Moreover, as pointed out above, there is often no better alternative to building and testing models, because informal theories are not only plagued by all of the aforementioned complications and problems of modeling, in addition they suffer from many others, fuelled by a lack of precision and reflected in the testing of vague hypotheses against chance. To conclude, psychologists' most prominent ritual might be abandoned when more researchers, trained in modeling, become reviewers and editors of our discipline's scientific journals. Hopefully then, there will be more emphasis placed on formulating psychological theories that are precise enough to be subjected to formal comparisons on model selection criteria beyond the null ritual, aiding the progress of psychology as science.

Acknowledgments

Our thanks go to Wolfgang Gaissmaier, Gerd Gigerenzer, Konstantinos Katsikopoulos, Stefan Lindner, Geoffrey Loftus, Lael Schooler, Peter Sedlmeier, and an anonymous reviewer for many constructive comments on earlier drafts of this article. We would also like to thank Julia Schooler and Anita Todd for editing the manuscript.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrox & F. Caski (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Altmann, G. (2007). Editorial. Journal policies and procedures. *Cognition*, *102*, 1–6.

Anderson, J.R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Erlbaum.

Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.

Anderson, J.R., & Bower, G.H. (1973). *Human associative memory*. Washington, DC: Winston and Sons.

Anderson, J.R., & Schooler, L.J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory and Cognition*, *32*, 1369–1378.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.

Box, G.E.P. (1979). Robustness in the strategy of scientific model-building. In R.L. Launer & G.N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic Press.

Brighton, H. (2006). Robust inference with simple cognitive models. In C. Lebiere & B. Wray (Eds.), *Between a rock and a hard place: Cognitive science principles meet AI-hard problems. Papers from the AAAI Spring Symposium* (AAAI Tech. Rep. No. SS-06–03, pp. 17–22). Menlo Park, CA: AAAI Press.

Browne, M.W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108–132.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie / Journal of Psychology, 217,* 15–26.

Danzinger, K. (1990). *Constructing the subject*. Cambridge, UK: Cambridge University Press.

Dhami, M.K. (2003). Psychological models of professional decision making. *Psychological Science*, *14*, 175–180.

Dücker, B. (2007). *Rituale. Formen, Funktionen, Geschichte. Eine Einführung in die Ritualwissenschaft* [Rituals. Forms, functions, history. An introduction into the science of rituals]. Stuttgart: Metzler.

Dutton, J.M., & Starbuck, W.H. (Eds.). (1971). *Computer simulation of human behavior: A history of an intellectual technology*. New York: Wiley.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Fechner, G.T. (1966). *Elements of psychophysics*. (H.E. Adler, Trans.). New York: Holt, Rinehart and Winston. (Original work published 1860).

Forster, M.R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205–231.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 4*, 379–390.

Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, *8*, 135–142.

Gaissmaier, W., Schooler, L.J., & Mata, R. (2008). An ecological perspective to cognitive limits: Modeling environment-mind interactions with ACT-R. *Judgment and Decision Making*, *3*, 278–291.

Gaissmaier, W., Schooler, L.J., & Rieskamp, J. (2006). Simple predictions fuelled by capacity limitations: When are they successful? *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *32*, 966–982.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*, 254–267.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Gigerenzer, G. (1998a). Surrogates for theories. *Theory and Psychology*, *8*, 195–204.

Gigerenzer, G. (1998b). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Gigerenzer, G., & Brighton, H. (2008). *Homo heuristicus: Why biased minds make better inferences*. Manuscript submitted for publication.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *104*, 650–669.

Gigerenzer, G., Hoffrage, U., & Goldstein, D.G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, & Thomas. *Psychological Review*, *115*, 230–239.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., & Richter, H.R. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development*, *5,* 235–264.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance. How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Grainger, J., & Jacobs, A.M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518–565.

Grünwald, P.D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Hintzman, D.L. (1991). Why are formal models useful in psychology? In W.E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Hillsdale, NJ: Erlbaum.

Hogarth, R.M., & Karelaia, N. (2007). Heuristics and linear models of judgment: Matching rules and environments. *Psychological Review*, *114*, 733–758.

Holland, J.H. (1998). *Emergence: From chaos to order*. Redwood City, CA: Addison-Wesley.

Jacobs, A.M., & Grainger, J. (1994). Models of visual word recognition. Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kelley, H.H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15). Lincoln: University of Nebraska Press.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, *4*, 236–243.

Loftus, G.R. (1993). Editorial comment. *Memory and Cognition*, *21*, 1–3.

Loftus, G.R., & Masson, M.E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review*, *1*, 476–490.

Loftus, G.R., Oberg, M.A., & Dillon, A.M.J. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*, 835–863.

Luce, R.D. (1999). Where is mathematical modeling in psychology headed? *Theory and Psychology, 9,* 723–737.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Masson, M.E.J., & Loftus, G.R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.

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*, 806–834.

Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology*, *64*, 553–557.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Myung, I.J., Balasubramanian, V., & Pitt, M.A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA*, *97*, 11170–11175.

Myung, I.J., & Pitt, M.A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4*, 79–95.

Navarro, D.J., Pitt, M.A., & Myung, I.J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.

Neumann, J. von, & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Newell, A. (1973a). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W.G. Chase (Ed.), *Visual information processing* (pp. 283–310). New York: Academic Press.

Newell, A. (1973b). Production systems: Models of control structures. In W.G. Chase (Ed.), *Visual information processing* (pp. 463–526). New York: Academic Press.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 600–620.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.

Pitt, M.A., Kim, W., Navarro, D.J., & Myung, J.I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57–83.

Pitt, M.A., & Myung, I.J. (2000). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421–425.

Pitt, M.A., Myung, I.J., & Zhang, S. (2002). Toward a method for selecting among computational models for cognition. *Psychological Review*, *109*, 472–491.

Polk, T.A., & Seifert, C.M. (Eds.). (2002). *Cognitive modeling*. Cambridge, MA: MIT Press.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.

Rieskamp, J., & Otto, P. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology*: *General*, *135*, 207–236.

Ritter, S., Anderson, J.R., Koedinger, K.R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin and Review*, *14*, 249–255.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rosnow, R.L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Zeitschrift für Psychologie / Journal of Psychology, 217,* 6–14.

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. I). Cambridge, MA: MIT Press.

Salvucci, D.D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, *48*, 362–380.

Schelling, T. (1969). Models of segregation. *American Economic Review, 59,* 499–493.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1172–1181.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schooler, L.J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*, 610–628.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Sedlmeier, P., & Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie* [Research methods and statistics in psychology]. Munich: Pearson Education.

Shiffrin, R.M., Lee, M.D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248–1284.

Stevens, S.S. (1957). On the psychophysical law. *Psychological Review, 64,* 153–181.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, *36*, 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, *64*, 29–35.

Thorngate, W. (2000). Teaching social simulation with Matlab. *Journal of Artificial Societies and Social Simulation*, *3*. Retrieved August 3, 2008, from http://www.soc.surrey.ac.uk/JASSS/3/1forum/1.html

Townsend, J.T. (1975). The mind-body problem revisited. In C. Cheng (Ed.), *Philosophical aspects of the mind-body problem* (pp. 200–218). Honolulu, HI: Honolulu University Press.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wagenmakers, E.-J. (2003). How many parameters does it take to fit an elephant? *Journal of Mathematical Psychology, 47*, 580–586.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Woodworth, R. (1938). *Experimental psychology*. New York: Holt.

Julian Marewski

Max Planck Institute for Human Development
Center for Adaptive Behavior and Cognition
Königin-Luise-Strasse 5
D-14195 Berlin
Germany
Tel. +49 30 82406-302
E-mail marewski@mpib-berlin.mpg.de

