American Psychological Association

Psychological Review

© 2020 American Psychological Association
ISSN: 0033-295X

2020, Vol. 127, No. 1, 136–145
http://dx.doi.org/10.1037/rev0000167

COMMENTARY

Strong Theory Testing Using the Prior Predictive and the Data Prior

Wolf Vanpaemel
University of Leuven

In their seminal article, Roberts and Pashler (2000) highlighted that providing a good fit to empirical data does not necessarily provide strong support for a theory. For a good fit to be persuasive and for a theory to be strongly supported, the theory should have survived a strong test, in the sense that it is plausible that the theory might have failed the test. The most common way to accommodate the problem of the limited value of a good fit alone is to not only report a measure of goodness-of-fit, but also a measure of the complexity. A recent example of this line of reasoning is provided by Veksler, Myers, and Gluck (2015). In this article, I argue that whereas considering complexity provides useful information when theory testing, using complexity to gauge the severity of a test, or, equivalently, the persuasiveness of a good fit, is misguided. The reason is that complexity only provides information about the possibility of a bad fit, which does not guarantee a strong test. A condition for a test to be strong and a good fit to be persuasive is the demonstration of the plausibility of a bad fit. I provide a worked example of a more complete answer to assessing whether a good fit is persuasive. Providing a strong theory test requires the use of what can be called a data prior, which quantifies—before taking the empirical data into account—which outcomes are plausible.

Keywords: theory testing, model evaluation, model complexity, goodness of fit, data prior

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

An important goal of science is making sense of the world by constructing strong theories, which are hoped to provide precise accounts of phenomena or processes of interest. To a large extent, the success of this strategy is dependent on the availability of valid strategies for testing theories. Theory testing often involves instantiating the theory into a formal model and considering whether the model is consistent with empirically observed data. If that is the case, the model is said to provide a good match or fit to the data.

Theory testing based on how well the model fits empirical data might seem highly reasonable, as a good fit between theory and data has considerable rhetorical force. However, in their seminal article, the late Seth Roberts and Hal Pashler (Roberts & Pashler, 2000, see also Roberts & Pashler, 2002) forcefully warned against the blind use of fit for testing theories. They argued that a good fit alone is not necessarily meaningful, impressive or persuasive, and therefore does not necessarily increase the confidence in a theory. Using refreshingly frank words, they called the use of a good fit as support for a theory "a mistake" (Roberts & Pashler, 2000, p. 358) and even "rotten to the core" (Roberts & Pashler, 2002, p. 605).

Their main point is that a good fit between theory and data is not impressive if the theory has survived a weak test, in the sense that it is almost guaranteed to pass the test. For a good fit to be persuasive, the test should be strong. Knowing that a theory has passed a test—that is, that it provided a good fit—is of limited value only if one does not know whether the test was a strong one or a weak one. Meaningful support, or a persuasive fit, can only be provided by a strong test. If only the goodness of fit is reported, it is difficult if not impossible to assess the severity of the test, and to draw any meaningful conclusion about the persuasiveness of the good fit.

In sum, to assess the persuasiveness of a good fit, some additional information is necessary. But which exact information do we need? The dominant strategy is going beyond reporting just goodness-of-fit by additionally reporting some measure of model complexity. This practice is reflected in many well-known model selection measures, such as Akaike's information criterion, Bayesian information criterion, and minimum description length, which combine a measure of fit with a measure of complexity (e.g., Myung, 2000). Veksler et al. (2015) propose another implementation of this same idea, and suggest that the persuasiveness of a

Portions of this work (titled "Five Routes to Better Models of Cognition") were presented at the 44th Annual Meeting of the Society for Mathematical Psychology (July 16–18, 2011), the European Mathematical Psychology Group Meeting (August 29–31, 2011), the 26th Annual Convention of the Association for Psychological Science (May 22–25, 2014), and the 47th Annual Meeting of the Society for Mathematical Psychology (July 18–21, 2014). The author appreciates the kind encouragement of Francis Tuerlinckx, Paul De Boeck, Zoltan Dienes, Peter Grünwald, Eric-Jan Wagenmakers, Richard Shiffrin, and Jay Myung, and especially thanks Michael Lee for his continuous support and guidance and Johan Vanpaemel for early exposure.

The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (C14/19/054).

Correspondence concerning this article should be addressed to Wolf Vanpaemel, Faculty of Psychology and Educational Sciences, University of Leuven, 3000 Leuven, Belgium. E-mail: wolf.vanpaemel@kuleuven.be

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

good fit should be gauged in the light of a newly proposed stand-alone measure of model complexity, called ϕ.[1] This measure, which, roughly, estimates the proportion of potential outcomes a model can account for, is essentially a non-Bayesian version of the prior predictive complexity (PPC; Vanpaemel, 2009). Both ϕ and PPC are stand-alone measures of complexity that measure complexity as the ratio of the width of the interval containing the outcomes the model predicts and the width of the interval containing all possible outcomes that could potentially be observed.[2] Veksler et al. (2015) consider a good fit to be persuasive only if ϕ is sufficiently low. When ϕ is high, a good fit is not persuasive, and provides no support for the theory. This reasoning basically implements the famous principle of *falsifiability*, which states that a theory can only be supported if it can be proven wrong.

First, I argue that a measure of complexity like ϕ lacks crucial information for gauging the severity of a test and the persuasiveness of a good fit. In particular, it focuses on the possibility of being proven wrong, rather than on the plausibility of being disproven. By focusing on possibility rather than on plausibility, it is still unclear whether the theory has been subjected to a weak or a strong test, whether the good fit is persuasive or unconvincing, and whether the support for the theory is weak or strong.[3] Second, I provide a worked example of what is needed to assess the persuasiveness of a good fit. The key ingredient of my proposal is the use of what I refer to as a *data prior*—a quantification of which outcomes are plausible and implausible. Without such a data prior, it is impossible to assess the severity of a test and, hence, the persuasiveness of a good fit.

Complexity Is Not Sufficient for Gauging the Persuasiveness of a Good Fit

To evaluate whether complexity is sufficient to evaluate the persuasiveness of a good fit, one should consider the proposal from Roberts and Pashler (2000). Inspired by Lakatos (1978), they argued that a good fit is persuasive and provides support for a theory only if a bad fit was plausible. They were careful to distinguish their position from a position inspired by Popper's (1959) view, who suggested that a good fit is persuasive if it was possible that the theory might not have fit. The difference between the possibility and the plausibility of a bad fit is critical. To drive home this crucial difference, Roberts and Pashler discussed an example, taken from Lakatos, focusing on the Newtonian theory of gravity. This theory predicts that a stone, when dropped, falls toward the earth. This prediction rules out some outcomes, such as the stone staying put or rising when dropped. Thus, a bad fit is clearly possible. If, the possibility of a bad fit is sufficient for a good fit to be persuasive, the observation of a falling stone would provide support for Newtonian gravity. If, however, the plausibility of a bad fit is the necessary condition for a good fit to be persuasive, observing a falling stone does not provide support for Newton's theory at all. The reason is that the theory does not rule out any plausible outcomes. Although it is possible that the prediction that the stone will fall will turn out otherwise, it is not very plausible that the prediction will be wrong, and the stone will not fall. From earlier experiences with stones, one expects that stones fall when dropped. Thus, because the outcomes that were ruled out were implausible, the prediction of a falling stone is without any real risk. Its confirmation is almost inevitable, and therefore the good fit between data and model should not be considered as persuasive.

In the early 1700s, however, Newtonian theory also made the prediction that Halley's comet would return. This prediction again rules out some outcomes (i.e., the comet not returning), so a bad fit is clearly possible. But unlike in the case of falling stones, it was highly plausible that the prediction of the returning comet would turn out otherwise. The reason is that, at the time, the return of the comet was unexpected, stunning, and novel. Thus, because the outcome that was ruled out (i.e., the comet not returning) was plausible, the prediction of the comet returning was risky. Consequently, the return of Halley's comet, on Christmas night in 1758, provided spectacular and compelling support for Newton's theory.

How does a stand-alone measure of complexity like ϕ relate to the plausibility of a bad fit? First, consider the situation where Veksler et al. (2015) claim a good fit to be persuasive, which happens when ϕ is low. This means that the theory rules out many possible outcomes. If, however, none of the outcomes the theory rules out are plausible, the good fit is not persuasive, despite the low complexity. If all the excluded outcomes are very unlikely to occur, the test was very easy to pass. A good fit is almost as inevitable as the good fit of a theory that is maximally complex and can fit all potential outcomes. A low ϕ does not reveal what one really needs to know: that the theory cannot fit any plausible outcome. In sum, a theory can be low in complexity, but still not be supported by a good fit.

Second, consider the situation where Veksler et al. (2015) consider a good fit to be not persuasive. This occurs when ϕ is

[1] Veksler et al. (2015) prefer the term *flexibility* over the term *complexity*, to avoid misunderstanding the term as "being difficult to understand." However, because the term *complexity* has been often used in the literature, I think this fear is unfounded, and I use the term *complexity*. It is important to note that although early definitions of *complexity* restrict it to the number of parameters, I, like Veksler et al. (2015), use a more general definition to include other aspects of complexity, such as functional form (e.g., Myung, 2000; Vanpaemel, 2009).

[2] When discussing PPC, Veksler et al. (2015) not only use an incorrect equation, but also an incorrect name. By referring to PPC as *prior predictive evaluation*, they give the wrong impression that PPC can be used for theory testing or model evaluation rather than as a measure for model complexity. This is especially unfortunate, since the bottom line of this article is precisely that complexity provides insufficient information to evaluate the meaning of a good fit of a theory. Further, they correctly observed that, up until now, PPC has only been applied under the assumption of independence of conditions, where the joint prior predictive distribution can be approximated reasonably well by the marginalized prior predictive distributions. But as they correctly note themselves, in their Footnote 5, there is no theoretical barrier for applying PPC in conditions where this assumption is less defensible. The idea underlying PPC straightforwardly applies to the full, joint distribution over several dependent variables. The only real difference between PPC and their proposed ϕ is the way the predictions are derived. Whereas PPC uses a Bayesian framework and relies on a likelihood and a parameter prior to make predictions in the form of the prior predictive distribution, ϕ ignores the likelihood (and thus sampling variability) and the prior (and thus any prior information about the parameters that cannot be expressed in a range restriction).

[3] Evans, Howard, Heathcote, and Brown (2017) have also pointed out some limitations of the Veksler et al. (2015) proposal, by arguing that ϕ does not provide a reliable way to measure model complexity. My criticism is more general; I argue that even a proper measure of complexity is insufficient for assessing the severity of a test and the persuasiveness of a good fit.

high, indicating that the theory rules out a few of the possible outcomes only. If, however, a large enough subset of these excluded outcomes are plausible ones, the good fit is (at least weakly) persuasive. Whereas a high ϕ makes it unlikely that the theory rules out plausible outcomes, it doesn't make it impossible. In sum, a theory can be high in complexity, but still be (weakly) supported by a good fit.

Both scenarios together make it clear that complexity by itself provides limited guidance for gauging the persuasiveness of a good fit. Complexity, as assessed by ϕ or by any other standard model complexity measure, such as geometric complexity (Myung, Balasubramanian, & Pitt, 2000; Pitt, Myung, & Zhang, 2002), parameter space partitioning-based complexity (Pitt, Kim, Navarro, & Myung, 2006; Pitt, Myung, Montenegro, & Pooley, 2008), or prior predictive complexity (Vanpaemel, 2009) is only informative about the extent to which a theory rules out possible outcomes. Crucially, it does not convey whether the theory rules out plausible outcomes. It is the latter, not the former information that is useful for gauging the severity of the test and thus the persuasiveness of a good fit.

Gauging persuasiveness of a good fit by complexity implements Popper's (1959) view, focusing on the possibility of a bad fit, but fails to implement Lakatos' (1978) view, focusing on the plausibility of a bad fit. Admittedly, a theory that is low in complexity rules out a lot of outcomes and is therefore quite likely to rule out plausible outcomes; and a highly complex theory rules out a limited set of outcomes only, and so is quite unlikely to rule out plausible outcomes. But complexity falls short of guaranteeing the plausibility of a bad fit and is therefore of limited use for gauging the severity of a test and the persuasiveness of a good fit.

Outcome Plausibility Is Crucial for Gauging the Persuasiveness of a Good Fit

In this section, I demonstrate what is needed to evaluate the persuasiveness of a good fit. This proposal has two key ingredients. Like Veksler et al. (2015), a first key ingredient are predictions. Unlike Veksler et al. (2015), I rely on Bayesian prior predictions, which are based on a likelihood and a prior (see also Footnote 2), but this difference is not crucial. The second ingredient is the crucial addition, and the one that Veksler et al. (2015) missed in their proposal. Irrespective of how predictions are made (Bayesian or non-Bayesian), evaluating the persuasiveness of a good fit requires the use of what I call a *data prior*. The data prior quantifies—before the data have been observed—which outcomes are plausible for a certain task and design. The data prior is a distribution over outcomes and should not to be confused with the parameter prior or model prior from the Bayesian statistical framework, which are distributions over parameter values and models, respectively.[4]

Model and Experimental Task

As a worked example, I focus on the generalized context model (GCM; Nosofsky, 1986), a seminal exemplar model of category learning. According to the model, classification decisions are based on the similarity of the to-be-classified stimulus to the relevant categories. Categories are assumed to be represented by individual exemplars, which means that, in determining the similarity of a stimulus to a category, every exemplar in that category is considered. Stimuli are assumed to be represented as points in a multidimensional psychological space, the structure of which is systematically modified by selective attention.

The data to evaluate the model are collected in a category learning task, where, in the training phase, participants are presented with stimuli and their accompanying category labels and, in the subsequent test phase, participants are required to assign old and novel stimuli to one of two different categories. Panels A and C in Figure 1 show a graphical depiction of two category structures to be learned, based on the same set of 12 two-dimensional stimuli. The squares and circles correspond to those stimuli that were assigned to either Category A or Category B, whereas the numbers without shapes correspond to stimuli that are unassigned to a category. During the training phase, only the assigned stimuli are presented, and corrective feedback is provided after each response, indicating the true category membership of the presented stimulus. During the test phase, both the assigned and unassigned stimuli are presented, and corrective trial-by-trial feedback is provided for assigned stimuli but is withheld on trials in which unassigned stimuli are presented. In the current application, I focus on the proportion of Category A responses in the test phase, for each stimulus separately, as the relevant empirical data.

Deriving the Core Predictions

My proposal uses two key ingredients: predictions and a data prior. For the first ingredient, the predictions, I rely on the prior predictive distribution, a key concept in the Bayesian statistical framework (see, e.g., Box, 1980; Edwards, Lindman, & Savage, 1963; Lee, 2018), which gives the probability of unseen, future, yet-to-be observed outcomes arising under the assumption that the model is true.[5] Formally, the prior predictive distribution involves integrating the probability of the outcomes for each of the possible parameter values, as weighted by the prior probability of each of these values: $p(d \mid M) = \int_{\Omega} p(d \mid \theta, M)p(\theta \mid M)d\theta$, with d the data, M the model, Ω the range of the parameter vector θ, $p(d \mid \theta, M)$ the data-generating function formalized in the likelihood, and $p(\theta \mid M)$ the prior, specifying, without taking observed data into account, which parameter values are likely. The predictions are made before data are observed and are thus predictions in the strict sense.[6]

Because predictions formalized by the prior predictive distribution are made probabilistically, all possible, yet-to-be-observed outcomes will generally be assigned at least some prior mass. This implies that in a strict sense, no outcomes are ruled out by

[4] Recently, Chandramouli and Shiffrin (2016) proposed to extend Bayesian induction by assigning priors to data distributions for the experiment under consideration. However, unlike the approach taken in this article, their approach does not take the severity of a test into account.

[5] It is called a *predictive* distribution, because it shows the probabilities of the outcomes. It is called a *prior* distribution, because it is based on the model alone, without taking any observed data into account, and can be specified before the data have been observed. The prior predictive distribution is often called a marginal likelihood, but in the present context its alternative name of prior predictive distribution is more apt.

[6] The term *prediction* is often used, confusedly, for quantities that are not genuine predictions but that are actually *postdictions* (Lee, 2018; Taagepera, 2007) in the sense that they are derived from the same data that one attempts to predict. In the Discussion section, I explain why predictions, rather than postdictions, should be used.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

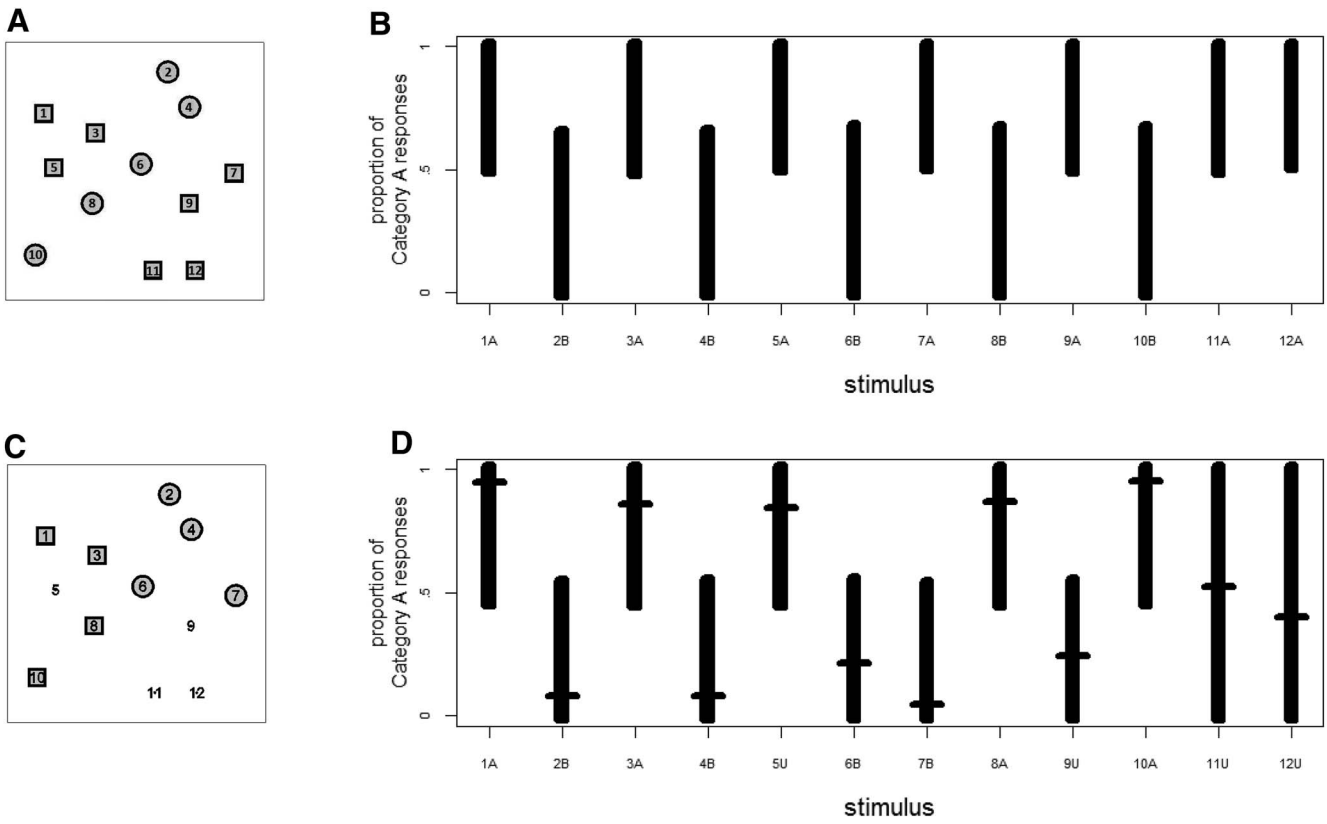


Figure 1. Panels A and C show two category structures that are based on the same set of 12 stimuli. In each panel, a numbered location corresponds to a stimulus in a two-dimensional psychological space. Squares denote stimuli assigned to Category A, circles denote stimuli assigned to Category B. The remaining stimuli are unassigned. Panels B and D show the 99.99% core predictions of the Category A responses of the GCM for the structures of Panels A and C, respectively, using a sample size of 240 and 400, respectively. The predictions in Panel B are for a model version assuming a uniform prior for both the sensitivity parameter and the attention parameter (using 10 and 1 as upper bounds, respectively), to reflect the absence of strong expectations about sensitivity and attention in this design. The predictions in Panel D are for a model version assuming a uniform prior bounded by 10 for the sensitivity parameter, and a beta (9, 1) prior for the attention parameter to capture the expectation that in this design learners will mainly focus on the relevant, vertical dimension (Nosofsky, 1986; Vanpaemel & Lee, 2012). The horizontal lines in Panel D indicate the empirically observed proportions, as reported in Nosofsky's (1987) saturation (B) condition.

probabilistic models, and as Lakatos (1978, p. 25) observed, a rule is needed to make some outcomes inconsistent with the theory. Therefore, I will make use of core predictions, which correspond to the smallest possible interval of outcomes that includes some predetermined amount of the prior mass. For example, the 99% core predictions are defined as those outcomes with the highest prior mass that together make up 99% of the prior mass. As such, it contains those outcomes the model can provide a good fit to and excludes those outcomes the model fits poorly. Following Popper (1959, p. 182) I consider an outcome to be ruled out if it is highly improbable, in the sense that its predictive mass is so small the outcome falls outside of the core predictions.

Panels B and D in Figure 1 show the core predictions of the model for the category structures shown in Panels A and C, respectively (the horizontal lines on each black bar in Panel D can be ignored for now). In each panel, the *y*-axis covers the full range of outcomes that can possibly be observed in the task. In this application, the Category A response proportion is the dependent variable, so this interval ranges from 0 to 1. If the main interest would, for example, be the difference between two proportions, the interval containing the full range of outcomes should be changed accordingly. The *x*-axis indexes each of the 12 stimuli, together with an indication of whether the stimulus belongs to Category A or B or is unassigned, denoted by U. The 99.99% core predictions are represented by the black bar for each stimulus separately.

It is important to appreciate that the prior predictive distribution is, in this case, a single 12-dimensional distribution over the configuration of outcomes for all 12 stimuli. The black bars are merely summaries of the full joint prior predictive distribution—useful for the current application, but not necessarily for all possible applications—in at least two aspects. First, each black bar corresponds to a marginalized prior predictive distribution, focusing on a single stimulus only. In other situations, with other goals, or with other background information, other summaries might be

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

more appropriate. For example, when a key interest is the combination of responses, or when relevant information regarding the plausibility of outcomes for such a combination is available, the prior predictive distribution should be summarized differently. For a researcher who is particularly interested in the difference in responses to, say, Stimuli 1 and 11 in the design of Panel C, a more useful summary of the same 12-dimensional prior predictive distribution is given in Figure 2. Second, each black bar corresponds to a probability distribution that has been flattened out. In the prior predictive distribution, some outcomes are assigned more prior mass than others, but knowing which outcomes are more likely than others is not necessary for determining whether the model rules out plausible outcomes and assessing whether the fit is good or bad.

Determining the Data Prior

The second ingredient in the proposed strategy for theory testing is the data prior. In the category learning application, the dependent variable is the Category A response proportion, so the possible outcomes range from 0 to 1. The data prior, however, only contains the plausible outcomes, not the possible outcomes. Consider now the design shown in Panel C in Figure 1. In this design, stimulus 1 has been assigned to Category A, as indicated by the square. Suppose now that, as is typical in a category learning task, the participants receive corrective feedback every time Stimulus 1 is presented in the training phase, and that the training phase was sufficiently long. In this scenario, it does not seem very plausible that participants would assign this stimulus to Category B when it is presented again in the test phase. If one is willing to assume that participants have learned something from the extensive feedback during the training phase, it seems reasonable to expect that the Category *A* response proportion for stimulus 1 will be above 0.5 in the test phase. This line of reasoning results in $\left(\frac{1}{2},1\right)$ as the data prior for Category A stimuli, and $\left(0,\frac{1}{2}\right)$ for Category B stimuli, expressing that the plausible outcomes correspond to above chance performance. Now consider Stimulus 5, which is unassigned to a category and has never been presented during training, as indicated by the absence of a square or circle. Assuming that this stimulus is sufficiently discriminable from the training stimuli, it seems reasonable to consider both a low and a high proportion of Category A responses to be plausible, which translates to a data prior of (0, 1) for unassigned stimuli.

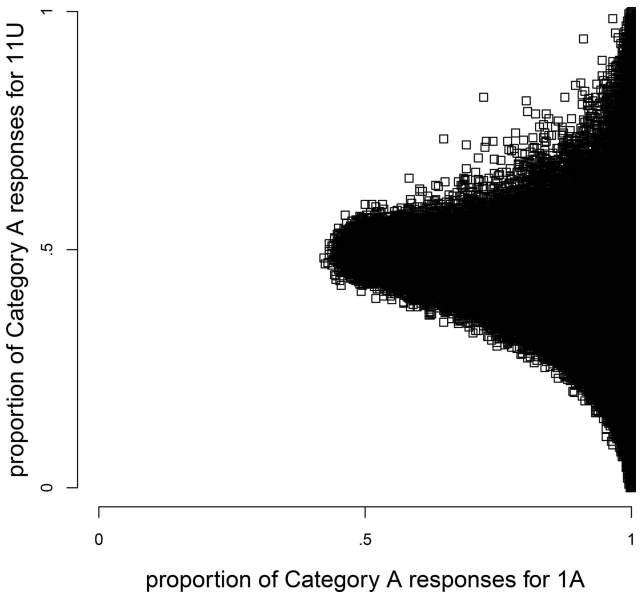


Figure 2. Joint predictions of the GCM for Stimuli 1 and 11 for the second category structure.

Assessing Test Severity and Fit Persuasiveness

With these two ingredients in place, it is possible to evaluate the severity of the test the theory is facing in the chosen experimental design, and thus, whether an eventual good fit should be considered as impressive, persuasive and providing strong support for the theory. First, I focus on the core predictions of the model for the category structure of Panel A, as shown in Panel B of Figure 1. Inspection reveals that there are outcomes outside the model's core predictions. Taking the simplest operationalization of what a good or bad fit means—either being within or outside the core predictions, respectively—this means that a bad fit is possible. For example, for Stimulus 1, a proportion of Category A responses (roughly) below $\frac{1}{2}$ is ruled out. However, as argued above, the possibility of a bad fit does not make a test strong and a good fit meaningful. What matters is the plausibility of a bad fit.

It is now the responsibility of the modelers to convince themselves and their audiences that there are plausible outcomes that are not (strongly) predicted by the model, or, put simply, that the model excludes plausible outcomes. This is where the data prior comes in. Using the data prior derived earlier, it is clear that for Stimulus 1, all outcomes outside of the core predictions are outside of the data prior and are thus implausible. Because the outcomes ruled out by the model —proportions below $\frac{1}{2}$ — are not very plausible, a bad fit is not plausible for Stimulus 1.

A similar argument can be made for the remaining 11 stimuli. For all 12 stimuli, the model does not rule out any plausible outcomes. The predictions of a proportion above chance for Category A stimuli and below chance for Category B stimuli corresponds to the prediction of a stone falling, in the sense that the outcomes it rules out are all implausible. Even before any observation is made, one virtually knows for sure that the prediction will be borne out, and that the model will survive the test. This experimental design does not provide a strong test for the model. Because it is almost impossible that the model will provide a bad fit, the model is immune to testing, and it makes no sense to collect data in this design and assess whether they agree with the predictions of the model. Because observing a good fit is inevitable, it is not persuasive and does not provide support for the model.

As a second example, consider Panel D of Figure 1, showing the 99.99% core predictions of the model for the design shown in Panel C. Note that, in contrast to the design from the previous example, the current design contains unassigned stimuli, which have not been presented during training, as indicated by *U*. Again, to be able to gauge the severity of the test and the persuasiveness of a good fit, one needs to show that a bad fit was plausible, or, equivalently, that the model rules out plausible outcomes. Using the data prior established earlier, it is clear that for the eight assigned stimuli, the model does not rule out plausible outcomes.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

For two of the unassigned stimuli (11 and 12), the model does not rule out any outcomes at all, let alone plausible ones. However, for the remaining two unassigned stimuli (5 and 9), the model rules out a low (for Stimulus 5) or high (for Stimulus 9) proportion of Category A responses. Thus, using this data prior, the model does not just rule out possible outcomes, it rules out plausible outcomes. The prediction of a high (for Stimulus 5) or a low (for Stimulus 9) proportion of Category A responses corresponds to the prediction of Halley's comet returning, in the sense that the ruled-out outcomes are plausible. Since it is plausible that the model makes the wrong predictions, it can be supported by empirical data. This experimental design provides a strong test for the model, so it makes sense to collect the data, and assess the fit.

The vertical lines in Panel D show that the empirically observed response proportions (as reported in Nosofsky, 1987, Saturation [B] condition). All observed proportions fall inside the core predictions, which means that the model provides a good fit. Because for two stimuli a bad fit was plausible, the good fit for these stimuli is persuasive and provides empirical support for the model. The model has been faced with and withstood a strong test.

Practice

In practice, the strategy for theory testing I propose involves the following steps. First, define a data prior, which is suited for the experimental design and model under consideration. Second, compute the core predictions, relevant to the particular application (either using the prior predictive or any other way, e.g., in Veksler et al. (2015)). If probabilistic predictions are used, as in the current worked example, decide on the threshold to define the core predictions. The most interesting information obtained in the second step is not what the model predicts, but rather what it rules out. Third, combine both sources of information to assess whether a bad fit is plausible. Only when the predictions do not completely overlap with the data prior, a bad fit is plausible, implying that the experimental design provides a strong test, and a good fit (if observed) is persuasive.

The next step depends on the outcome of the third step. If the model only rules out implausible outcomes (as in Panel B), the theory is challenged by a weak test only. It makes no sense to collect data in this design for this model. No empirical data will ever provide support for the model, even if the model's predictions and the data fit extremely well. Testing the model will require a different, more constrained version of the model, or a more informative experimental design. The only reasonable course of action is to change the model and/or the design, and to start over with the first step. If, in contrast, the model rules out plausible outcomes (as in Panel D), the theory is faced with a strong test, and can, in principle be supported by empirical data. As collecting empirical data will be informative about the empirical success of the model, it makes sense to collect data and evaluate the fit. If the observations fall inside the predictions, the fit is good and the data support the model, meaning that the data offer no reason to discard the model. If one or more observations fall outside the core predictions, the fit is bad, and the data discredit the model.

Discussion

Freeman Dyson famously quipped that "it is better to be wrong than to be vague" (Dyson, 1999, p. 48). If a theory is so vague that it does not exclude any outcomes, the theory is guaranteed to provide a good fit to any observed data pattern, but, precisely because of the inevitability, this good fit is meaningless. No empirical data can provide support for this vague theory. What is less appreciated is that being very precise, in the sense of excluding a large set of outcomes, is not enough for making a good fit impressive and persuasive. A precise theory might not be much better off than a vague theory, in the sense that it is possible that even the precise theory cannot be supported by empirical data. Roberts and Pashler (2000, 2002) point out that if the outcomes the theory rules out are implausible, the good fit of the precise theory is almost as inevitable, and therefore just as meaningless, as the good fit of the vague theory. Simply ruling out many outcomes is not enough for a theory to be falsifiable, for a test to be strong, and for a good fit to be meaningful. No empirical data can provide support for a precise theory, unless it has been shown that the theory rules out plausible outcomes, so that a bad fit is plausible.

Augmenting a reported fit with a measure of complexity, as many researchers, including Veksler et al. (2015), propose is somewhat useful, as it separates the vague theory from the precise theory. However, it does not provide sufficient information for evaluating the severity of a test and hence the persuasiveness of a good fit. Without establishing that the precise theory rules out plausible outcomes, there is no fundamental difference between the precise theory and the vague theory. The problem of focusing on complexity is that it deals with how much the theory rules, whereas what we really need to know is what the theory rules out. Evaluating the persuasiveness of a good fit requires more than an assessment of complexity. It depends critically on the quantification of which outcomes are plausible and implausible for the experiment and the theory at hand. Theory testing requires a data prior.

Determining a Data Prior

Testing a theory is not simply a matter of combining the predictions of the theory with empirical observations. It also requires background knowledge, which takes the form of the data prior. Various objections might be raised against this view. One is that determining the plausibility of outcomes is challenging, presents an extra burden to modelers, and is a very difficult and time-consuming task. All of this is true; but being hard does not negate the fact that it is what we need to do and does not imply that we would be better off not doing it at all.

I can offer no definite guidelines, let alone automatic rules to assist in setting a data prior. I can only express my hope that if data priors are becoming more commonly used when testing theories, the research community will come up with good practices for doing so. To start such a conversation, I offer some general thoughts on what could guide these practices. These thoughts notwithstanding, it is important to keep in mind that, just like building models and designing experiments, setting a data prior is a creative process that requires substantial expertise and careful judgment, and will likely escape automation.

One aspect researchers should keep in mind when specifying a data prior, is that what is considered plausible depends on the exact research design, instructions, and so on. The worked example highlights that the data prior should reflect the nature and design of the task of interest. If, for example, in a particular implementation

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

of a category learning task, training was absent or not extensive, or if an unassigned stimulus is perceptually very close to the training stimuli or is a category prototype, this information should be reflected in the data prior. The data prior obviously also depends on which (combination of) dependent variables are being considered. In the example, the dependent variable was a proportion for each stimulus in isolation, and so it was specified for each stimulus separately. If, in contrast, the main interest is in the combination of two dependent variables, as shown in Figure 2, the data prior should indicate which combined outcomes are plausible.

Further, one should set a data prior in such a way that it is sensitive to the theory under consideration. The data prior used in the worked example discussed here is independent of the model under consideration. It is as reasonable (or as unreasonable) for evaluating the GCM as it is for evaluating any other model of category learning, such as the prototype model (Nosofsky, 1987; Smith & Minda, 1998), SUSTAIN (Love, Medin, & Gureckis, 2004), ALCOVE (Kruschke, 1992), or the rational model of categorization (RMC; Anderson, 1991). In more general situations, however, this should not be the case. If one is interested in testing a theory of category learning that, for some exotic reason, assumes that the training stimuli are not learned by the participants, but rather that categorization is based on the features of the test stimuli alone, the data prior used in the current example is not appropriate.

An intriguing example of how outcome plausibility, and thus the data prior, should reflect the theory under consideration is offered by Lakatos and Zahar (1976), focusing on Einstein's General Theory of Relativity. This theory correctly predicts the exact amount of precession of Mercury's orbit. However, the exact amount of precession had already been empirically observed well before Einstein developed his theory, so the precession amounts that were ruled out were not plausible at all at Einstein's time. Still, Lakatos and Zahar argued, the fit between the empirical observation and the General Theory of Relativity is highly impressive and provides strong support for the theory. The reason is that the theory was not devised to predict this observation in particular, making a bad fit plausible. Thus, what is considered implausible should be considered with respect to the exact theory under evaluation. This means that the data prior for testing the General Theory of Relativity is different from the data prior for testing another (as far as I am aware, fictional) theory that is explicitly designed to accommodate the exact amount of precession.

Possible sources for specifying a data prior include theoretical considerations, previously observed empirical data, and expert knowledge, not unlike the sources typically used for determining parameter priors in Bayesian statistics (see Lee & Vanpaemel, 2018, for an overview). One practical way to construct a data prior is by considering the joint predictions of a set of alternative established models for explaining the same behavior. For example, when evaluating the GCM, the data prior could be defined by the union of the core predictions of the prototype model, SUSTAIN, ALCOVE, and the RMC, although deciding which models count as established is opening another can of worms.

Given that there is currently no single, automatic rule for determining what is plausible, and that there most likely never will be such a rule, modelers can, and most likely will, differ in what they consider plausible. In this light, a second objection against the data prior could be that the data prior may seem to introduce an unwanted source of subjectivity and arbitrariness, thereby corrupting the desired objectivity of theory testing. It seems inevitable that the set of plausible outcomes contained in the data prior are likely to be subject to substantial disagreement among modelers. Even with the same information, some modelers could, for example, disagree that (0, 1) is an appropriate data prior for unassigned stimuli, but consider the plausible range to be smaller. Others might disagree that $\left(0,\frac{1}{2}\right)$ and $\left(\frac{1}{2},1\right)$ are reasonable data priors for Category B and A stimuli, respectively, arguing for wider ones. These different views on plausibility may imply different conclusions. If, in an extreme case, all outcomes are considered plausible for all stimuli, a bad fit in the design of Panel A is plausible, so a good fit should be considered persuasive in that design.

The fact that model evaluation potentially depends on the judgment of the modelers might seem undesirable. However, it is important to realize that what are considered plausible outcomes is directly related to the severity of the test. Wide data priors make for lenient tests; narrow data priors imply strong tests. The fact that the result of a test depends on the severity of the test seems logical, and even desirable. Mild tests are easy to pass, whereas severe tests are hard to survive. It is up to the modelers, and to the readers, to decide on the desired degree of severity, and this judgment will likely depend on the context and the exact goal of the test. At least, when a data prior is used, the decisions made are open and fully in view. By account of this transparency, readers can decide for themselves whether they agree with the data prior used and whether they are satisfied with the implied severity of the test. If they are not, they are free to judge the persuasiveness of a good fit based on a data prior constructed by themselves.

Finally, it might be reassuring to realize that relying on a data prior for theory testing is not as radical as it may seem at first sight. Implicitly, the plausibility of outcomes plays an important role in model comparison. One scientific ideal is the notion of a critical or crucial experiment. When there are two competing, alternative theories, a crucial experiment is one in which both theories make contrasting predictions (Platt, 1964). Suppose that Theory 1 predicts Outcome A but not B, and Theory 2 predicts Outcome B but not A. If Outcome B is observed, then the good fit of Theory 2 is considered strong support for the theory. Why is this the case? The reason is that, assuming Theory 1 is a sensible theory, Outcome A was plausible because it is predicted by Theory 1. Because Outcome A was plausible and not predicted by Theory 2, Theory 2 clearly rules out a plausible outcome, making a good fit persuasive. Thus, the competing theory has, implicitly, provided a data prior. However, in absence of a competing theory, there is no such implicit data prior. Thus, when a theory is evaluated in isolation, the plausibility of excluded outcomes, and the associated severity of the test, cannot simply be assumed, but rather should be explicitly established using a data prior.

Given the strong tie of the term *prior* to the Bayesian framework, one might get the false impression that a data prior is only needed for Bayesian modelers. The fact that in the worked example, I relied on a Bayesian framework for deriving core predictions might reinforce this perceived connection. However, it is false. Even when the model and the predictions are not cast in Bayesian terms, a data prior is needed to assess the severity of a test and the persuasiveness of a good fit. The approach to theory testing advocated here is not simply a Bayesian version of Veksler et al.'s (2015) proposal. It is not the case that, depending on one's infer-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ential inclination, a modeler could either use the Bayesian approach discussed in this article or Veksler et al.'s non-Bayesian one. My proposal corrects the fact that Veksler et al. focused on the possibility of a bad fit instead of on the plausibility of a bad fit, and this correction is needed regardless of the choice for a Bayesian or non-Bayesian framework.

This correction is more than cosmetic. Ignoring the data prior, like Veksler et al. (2015) do, can strongly affect inference. To see this, consider again the experimental design shown in Panel A. This example highlights that focusing on the possibility of a bad fit (like Veksler et al. (2015) do) rather than on its plausibility (as in the current proposal) makes a big difference. Considering a possible bad fit as sufficient for a good fit to be persuasive, the (eventual) good fit in the Panel A design would support the theory. Thus, it would make sense to collect empirical data and to assess whether the predictions have been confirmed. In contrast, requiring a plausible bad fit implies that the (eventual) good fit does not provide strong support for the theory. From this perspective, it is clear that collecting empirical data provides very little information, because the theory was invulnerable, if not immune, to an experimental test.

Making Predictions

In the worked example, I have relied on a Bayesian framework to derive the core predictions, using the prior predictive, but the persuasiveness of a good fit can be evaluated using predictions made outside the Bayesian framework. The most important aspect of the predictions is that they are genuine, a priori predictions, rather than postdictions that take the observed data into account.[7] The reason is that when postdictions are being used, it is virtually impossible to provide support for a theory. The fit will either be bad, or the good fit will not be persuasive.

To see what happens when the predictions are made after observing the data, let us pretend for a moment that the predictions shown in Panels B and D are postdictions, that is, based on the observed data. It is now crucial to realize that using the observed data to make the predictions influences the data prior. Using already observed data to derive the predictions means that the theory can be seen as being devised to predict the data. Per the precession of Mercury's orbit example discussed in the preceding text, this should be reflected in the data prior. Since we know and use the observed data, the data prior is a narrow interval around the observed outcomes. If the observed data points fall outside the core predictions, the good news is that the theory rules out plausible outcomes, which makes the test strong and a good fit persuasive. But the bad news is that the fit is bad. If, on the other hand, the observed data points fall inside the core predictions, the good news is that the fit is good. But the bad news is that the test is weak, and the good fit is not persuasive: Given that the plausible outcomes are close to the observed data points, all plausible outcomes will fall inside the core predictions, so no plausible outcomes are excluded. Thus, when posterior predictions are being used, the data will never provide support for the theory.

An immediate consequence of relying on Bayesian prior predictions is that the predictions are sensitive to the parameter prior (not to be confused with the data prior). Some modelers might feel uncomfortable with model inference being sensitive to the parameter prior, but there are good reasons they should not (e.g., Gill & Walker, 2005; Vanpaemel, 2010). Just like the likelihood, the parameter prior is an integral part of the model, and it provides a way of formalizing available information and theoretical assumptions. It would be nonsensical if theory testing would be sensitive to one part of a model that captures theory (the likelihood) but not to the other (the parameter prior).

Another consequence of using Bayesian prior predictions is that, per their probabilistic nature, modelers need to set a threshold to define the core predictions. In the current examples, I used 99.99% as a threshold, but I have no justification whatsoever to offer for this value. One way to understand the threshold it to consider its relation to the severity of the test. If I had included a lower threshold (e.g., the 50% core predictions), the model would have excluded more outcomes, making it much more likely that it ruled out plausible outcomes, and that good fits are persuasive. As adopting a standard for test severity that is expected to hold across different theories, domains and goals is most likely not only impossible but even undesirable, adopting a standard threshold for core predictions is misguided.

The threshold for defining core predictions is closely related to the α level in the traditional null hypothesis significance testing approach to data analysis. The basic procedure is to compute the sampling distribution of a test statistic, assuming the null hypothesis holds, and use a specified α level to determine the rejection region. If the observed test statistic falls in the rejection region, the null hypothesis is rejected. How the α level should be set is still an ongoing debate. Recently, Lakens et al. (2018) suggested that researchers should justify their choice of α, and the same reasoning seems applicable for the threshold for defining core predictions.

The arbitrariness in setting this value is also related to the arbitrariness in setting the threshold value of what constitutes a low enough value of φ for a good fit to be persuasive. If this threshold is high, good fits are likely to be persuasive. Veksler et al. (2015, p. 761) are careful to note that ϕ "may be interpreted differently based on the empirical domain, model generalizability, and other factors" and that "[i]t may well be the case, for example, that a certain class of models or a certain class of behavioral phenomena requires a more conservative or a more liberal [ϕ value] . . . than others" (p. 764). Therefore, they consider it "ill-advised to have a universal "significance" threshold for ϕ-values" (p. 764). When using the ϕ value to gauge the persuasiveness of a good fit, it is to some extent up to the modelers to decide and justify what level of severity is appropriate for the given context and goal. The exact same is true for setting the threshold for computing the core predictions: A one size fits all approach is ill-advised.

Risky Predictions

A test is only strong, and a good fit is only persuasive, if the predictions made by the theory are bold or risky. This means that a good fit should not be practically guaranteed, or, equiv-

[7] In a Bayesian framework, postdictions can be made using the posterior predictive distribution. Whereas prior predictions average over the parameter prior, posterior predictions average over the parameter posterior, which is the parameter prior updated by the observed data. Therefore, only prior predictions are genuine predictions, made independent of the data, whereas posterior predictions should be seen as postdictions.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

alently, that a bad fit should be plausible. Yet another way to put is that the theory should rule out plausible outcomes. The requirement that the theory should rule out plausible outcomes is not equivalent to the requirement that the theory should predict outcomes that are a priori implausible or unlikely (or, in Lakatos's words: novel, stunning, surprising, unexpected, or dramatic). This might be surprising, especially when one has a deterministic mindset. If a deterministic theory predicts an implausible outcome, this logically implies that the theory rules out a plausible one (i.e., the opposite of the implausible outcome). For example, when Newtonian gravity predicts the implausible outcome of the return of Halley's comet, it necessarily rules out the plausible outcome of the comet not returning. With probabilistic models, however, predicting an implausible outcome is neither necessary nor sufficient for ruling out a plausible outcome. To understand it is not necessary, consider again Panel D. For Stimulus 5, the model predicts plausible outcomes only, a high proportion of Category A responses. Yet, despite not predicting implausible outcomes, the model excludes plausible ones—a low proportion of Category A responses. To understand it is not sufficient, consider again Panel B. For Stimulus 2, the model predicts that the proportion of Category A responses can be above .5, which corresponds to an implausible outcome. However, the fact that the model predicts the implausible above .5 proportion does not automatically imply that the model rules out the plausible below .5 proportion. In fact, the model also predicts a below .5 proportion for that stimulus. Despite predicting implausible outcomes, the model does not rule out plausible outcomes.

Although precise predictions are not necessarily risky, as argued in the first part of this article, and novel predictions are not necessarily risky, as argued in the preceding text, precise predictions that are novel are very likely to be risky. On the account of being precise, the theory rules out a lot of outcomes; and on the account of making novel predictions, the theory predicts implausible outcomes; together, this means that the theory is likely to rule out plausible outcomes. In other words, a theory that makes precise predictions that could not be made in absence of that theory is likely to make risky predictions, and to face a strong test.

So How Persuasive Is a Good Fit?

My proposed strategy for theory testing is rather crude, in the sense that the outcomes are very limited: a fit is either good (inside the core predictions) or bad (outside the core predictions), and a good fit is either persuasive (if the theory rules out plausible outcomes) or meaningless (otherwise). The same holds for the strategy proposed by Veksler et al. (2015). Although they do provide a quantification of how good the fit is (using RMSE) and of how complex a model is (using ϕ), they do not provide a quantitative assessment of how persuasive a good fit is, which should rely on a principled combination of both goodness of fit and complexity. Instead, in their use cases, the conclusions involve vague labels like *unpersuasive*, *fairly persuasive*, and *persuasive*. Even Roberts and Pashler (2000) use vague conclusions only, such as *weak support*, *strong support*, *real support*, and *substantial evidence*, which is somewhat surprising, given that their article is titled *How Persuasive Is a Good Fit?* Although they do clearly answer the question of when a good fit is persuasive, they are silent about exactly how persuasive a good fit is.

The basic approach advocated in the article could be further developed into a more sophisticated version, which treats the persuasiveness of a good fit as a matter of degree rather than as a dichotomous decision. The most persuasive good fits occur when a theory rules out a lot of plausible outcomes, but at the same time strongly predicts the observed outcome. Prediction strength could, for example, be assessed by looking at the distance to the edge of the core predictions, or, more directly, by looking at the prior predictive mass assigned by the model. Recall that in Panels B and D each black bar corresponds to a probability distribution that has been flattened out. Nevertheless, some outcomes are assigned more prior mass than others. In developing such a fine grained, continuous measure, the corroboration index as proposed by Meehl (1990) could serve as a useful starting point. That being said, a crude assessment of persuasiveness seems sufficient for theory testing or model evaluation, which involves assessing the absolute adequacy of an individual theory by itself. When, however, the goal is model comparison rather than model evaluation, a more fine-grained approach seems useful.

Conclusion

To evaluate whether a good fit to empirical data should count as a success for a theory, it is inevitable for one to specify a data prior, quantifying which outcomes are plausible. Without this information, it is impossible to judge the severity of a test and the value of a good fit. Because a stand-alone measure of model complexity like the PPC or ϕ does not contain any information about the plausibility of outcomes, they are not sufficient for evaluating the persuasiveness of a good fit.

Setting a data prior requires expert knowledge about the theory and the experimental paradigm under consideration. It will seldom be straightforward, and will often be done based on limited, vague, and subjective information. But these barriers do not change the fact that the data prior provides critical information for theory testing and should not serve as excuses for testing a theory without a data prior. Although incorporating a data prior in theory testing is likely to be controversial and face resistance, will be demanding, and requires thoughtful scientific judgment, it is an indispensable step when testing a theory. Whether or not we like it, without a data prior there is simply no way of putting a theory to a strong test, and of knowing whether a good fit between data and theory is persuasive.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. http://dx.doi.org/10.1037/0033-295X.98.3.409

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General), 143,* 383–404. http://dx.doi.org/10.2307/2982063

Chandramouli, S. H., & Shiffrin, R. M. (2016). Extending Bayesian induction. *Journal of Mathematical Psychology, 72,* 38–42. http://dx.doi.org/10.1016/j.jmp.2015.10.002

Dyson, F. J. (1999). *The sun, the genome, and the Internet: Tools of scientific revolutions*. New York, NY: Oxford University Press.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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

Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological Review, 124,* 339–345. http://dx.doi.org/10.1037/rev0000057

Gill, J., & Walker, L. D. (2005). Elicited priors for Bayesian model specifications in political science research. *The Journal of Politics, 67,* 841–872. http://dx.doi.org/10.1111/j.1468-2508.2005.00342.x

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44. http://dx.doi.org/10.1037/0033-295X.99.1.22

Lakatos, I. (1978). *The methodology of scientific research programmes*. New York, NY: Cambridge University Press.

Lakatos, I., & Zahar, E. (1976). Why did Copernicus' research program supersede Ptolemy's? In R. Westman (Ed.), *The Copernican achievement* (pp. 354–383). Los Angeles: University of California Press.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., . . . others. (2018). Justify your alpha. *Nature Human Behaviour, 2,* 168–171. http://dx.doi.org/10.1038/s41562-018-0311-x

Lee, M. D. (2018). Bayesian methods in cognitive modeling. In J. Wixted & E.-J. Wagenmakers (Eds.), *The Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 5, pp. 37–84). Hoboken, NJ: Wiley.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25,* 114–127. http://dx.doi.org/10.3758/s13423-017-1238-3

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332. http://dx.doi.org/10.1037/0033-295X.111.2.309

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141. http://dx.doi.org/10.1207/s15327965pli0102_1

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. http://dx.doi.org/10.1006/jmps.1999.1283

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 11170–11175. http://dx.doi.org/10.1073/pnas.170283897

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. http://dx.doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108. http://dx.doi.org/10.1037/0278-7393.13.1.87

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113,* 57–83. http://dx.doi.org/10.1037/0033-295X.113.1.57

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. http://dx.doi.org/10.1037/0033-295X.109.3.472

Pitt, M. A., Myung, J. I., Montenegro, M., & Pooley, J. (2008). Measuring model flexibility with parameter space partitioning: An introduction and application example. *Cognitive Science, 32,* 1285–1303. http://dx.doi.org/10.1080/03640210802477534

Platt, J. R. (1964). Strong inference. *Science, 146,* 347–353. http://dx.doi.org/10.1126/science.146.3642.347

Popper, K. R. (1959). *The logic of scientific discovery*. London, UK: Hutchinson.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367. http://dx.doi.org/10.1037/0033-295X.107.2.358

Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review, 109,* 605–607. http://dx.doi.org/10.1037/0033-295X.109.3.605

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1411–1436. http://dx.doi.org/10.1037/0278-7393.24.6.1411

Taagepera, R. (2007). Predictive versus postdictive models. *European Political Science, 6,* 114–123. http://dx.doi.org/10.1057/palgrave.eps.2210120

Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 1919–1927). Red Hook, NY: Curran Associates Inc.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498. http://dx.doi.org/10.1016/j.jmp.2010.07.003

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review, 19,* 1047–1056. http://dx.doi.org/10.3758/s13423-012-0300-4

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis. *Psychological Review, 122,* 755–769.

Received October 1, 2015
Revision received July 27, 2019
Accepted July 30, 2019 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.