Journal of Mathematical Psychology 72 (2016) 183–190

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

# Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective

Wolf Vanpaemel

*Faculty of Psychology and Educational Sciences, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium*

## HIGHLIGHTS

- The adoption of the Bayes factor would serve the prototype vs exemplar debate well.
- The use of the Bayes factor affords the possibility of comparing exemplar and prototype models in a way both exemplar and prototype theorists should feel comfortable with.
- The use of the Bayes factor encourages both exemplar and prototype theorists to think carefully about what their parameters mean, and to formalize theory about these parameters in the prior.

## ARTICLE INFO

*Article history:*
Available online 15 December 2015

*Keywords:*
Category learning
Exemplar
Prototype
Response scaling
Model selection
Priors
Bayes factor

## ABSTRACT

The Bayes factor can be used to break the stalemate between prototype and exemplar theorists in category learning. Exemplar theorists do not accept prototype theorists' results, because these results are often based on a restricted version of the exemplar model, without response scaling parameter. Prototype theorists do not accept exemplar theorists' results, because these results are often based on comparing fits of models that are unbalanced in their number of parameters. Using the Bayes factor alleviates concerns about differences in model complexity between exemplar and prototype models, because it takes away the prototype theorists' fear that their model is disadvantaged, and it eliminates the need to use a constrained version of the exemplar model exemplar theorists have abandoned. Further, by virtue of its sensitivity to the prior, the adoption of the Bayes factor encourages more complete theorizing about the relevant psychological processes instantiated in the models, and increases the empirical content of the models.

© 2016 Published by Elsevier Inc.

## 1. Introduction

> The functional and quantitative aspects of the evolution of concepts are at present in about the same state as were corresponding aspects of memory when Ebbinghaus published his monograph on that subject in 1885. This is true in spite of a very considerable activity on the part of experimenters in the field of the higher mental processes within recent years.

Despite being written almost 100 years ago – this quote is the opening paragraph of the classic by Hull (1920) – the characterization of the state of the art in the study of concept learning is still painfully accurate. Hull (1920) goes on to explain why it is still poorly understood how people acquire categories from labeled exemplars:

> The problem of generalizing abstraction has itself been directly attacked a number of times, notably by Moore in 1905, by Grünbaum in 1906–07, and by Miss Fisher in 1912–13. But in every case the studies have been largely introspective in method, analytic in purpose, and qualitative in result. The functional and quantitative aspects of the problem remain almost untouched. The reason for such a condition is of course not far to seek[.] It lies in the great complexity of the process involved.

Since the time of Hull, the field has moved away from the introspective, analytic and qualitative approach mentioned in the quote. But the adoption of the extrospective, empirical and quantitative approach to the study of categorization did not provide the clear picture one could have hoped for.

One reason for this slow progress is that the field has relied too heavily on poor model selection measures. In this paper, I demonstrate how the use of the Bayes factor – the topic of the current special issue – would have advanced our understanding of category learning.

*E-mail address:* wolf.vanpaemel@ppw.kuleuven.be.
*URL:* http://ppw.kuleuven.be/okp/people/wolf_vanpaemel/.

http://dx.doi.org/10.1016/j.jmp.2015.10.006
0022-2496/© 2016 Published by Elsevier Inc.

One aspect of categorization of which the study has been decidedly empirical and quantitative in nature involves the issue of representational abstraction. For decades, psychologists have intensely tried to decide whether people learn a category by abstracting a summary from the labeled examples – as postulated by the *prototype theory*, or by remembering and using the examples as is, in their full detail – as postulated by the *exemplar theory*. Initial support for the prototype theory was provided by findings that unseen prototypical examples are sometimes categorized more easily than old examples (Posner & Keele, 1968), and by an early demonstration that the predictions of a prototype model were closer to the empirically observed data than those of an exemplar model (Reed, 1972). However, when formal modeling showed that also the exemplar theory could account for prototypicality effects (Medin & Schaffer, 1978), and that the exemplar predictions matched the data better more often than the prototype predictions (see, e.g., Nosofsky, 1992, for an overview), support for the prototype theory started to give way in favor of the exemplar theory. It seemed like a clear victory, not only for the exemplar theory, but also for an empirical, quantitative modeling approach to studying human cognition.

Suddenly, prototype theory came back from the dead. In a string of articles, J.D. Smith and his colleagues showed that in some conditions, the prototype model fits data better than the exemplar model (Minda & Smith, 2001, 2002; Smith & Minda, 2000, 2002; Smith, Murray, & Minda, 1997). However, the validity of these findings has been severely criticized by exemplar theorists, who pointed out that the exemplar model used by the prototype theorists was a restricted version that was long abandoned by exemplar theorists (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002). The more recent version used by most exemplar theorists differed from the restricted version used by most prototype theorists in that it contains an additional parameter, called the response scaling parameter.

The main appeal of the restricted version of the exemplar model is that this version has the same number of parameters as the prototype model, and thus makes model comparison relatively easy (though see Footnote 3). Comparing the more recent version of the exemplar model to the prototype model is much more difficult: if the exemplar model is found to fit the data better than the prototype model, it is impossible to decide whether the superior fit is due to the fact that the exemplar model provides a better approximation to the way people learn categories, or to the difference in number of parameters. Thus, most exemplar theorists did not accept the findings from prototype theorists, claiming that the latter rely on an abandoned model. Most prototype theorists, on the other hand, did not accept the findings from exemplar theorists, claiming that any findings of superiority in fit of the exemplar model could be due the difference in number of parameters.

The trouble is not that both sides of the debate do not accept each other's findings. The real problem is that both sides are right to do so. On the one hand, using a model to test a theorist's view that is not endorsed by this theorist is academically dishonest. On the other hand, comparing the fits of models that potentially differ widely in complexity is biased in favor of the most complex model. The arguments of both sides are both valid. As a result, the prototype vs exemplar debate has reached a stalemate, one that Hull probably could not have imagined about a century ago.

In this paper, I demonstrate how we can escape the impasse. In particular, I show that the Bayes factor affords comparing the most recent version of the exemplar model to the prototype model, without fears that the model with more parameters is favored unwarrantably over the model with fewer parameters. Thus, using the Bayes factor alleviates the major concerns of both the exemplar theorists and prototype theorists at the same time, and frees the prototype vs exemplar debate from its current stagnation. Category learning modelers can finally stop criticizing each other's modeling choices, and start focusing on tracing out the contextual and individual conditions that foster or impede representational abstraction.

## 2. Comparing prototype and exemplar models

Category learning models, such as the prototype and the exemplar model, have often been compared by evaluating their ability to capture data collected in a category learning task, in which learners are presented with labeled exemplars and are asked to predict category labels for new stimuli.

### 2.1. Category learning task

A typical category learning task involves learning a two-category structures over a small set of $m$ stimuli, varying on two dimensions, such as semicircles with an embedded radial line varying on circle size and line orientation (Nosofsky, 1989). Upon stimulus presentation, the participant is asked to classify the stimulus into one of two categories. Following each response, corrective trial-by-trial feedback is presented on trials with assigned stimuli, but not on trials with unassigned stimuli. The relevant data for modeling are the counts, $k_i$, of the number of times the $i$th stimulus is classified as belonging to Category $A$, out of the $n_i$ trials it was presented.

### 2.2. Category learning models

While several implementations of exemplar-based and prototype-based representations have been proposed, I focus on the two most dominant ones: Nosofsky's (1986) Generalized Context Model (GCM), which is an exemplar model, and the Multiplicative Prototype Model (MPM, Nosofsky, 1987; Smith & Minda, 2000). When an experimental design involves two categories, as is the case in the applications considered here, the counts are assumed to be binomially distributed, $p(k_1, \ldots, k_m|\theta) = \prod_{i=1}^{m} \binom{n_i}{k_i} p_i(\theta)^{k_i}(1-p_i(\theta))^{n_i-k_i}$, with $p_i(\theta)$ the probability that $i$th stimulus is chosen in Category $A$, given a parameter value $\theta$, as defined below.

Nosofsky (1998, p. 335) provides a succinct description of his GCM :

> According to the GCM, people represent categories by storing individual category exemplars in memory. Classification decisions are based on similarity comparisons to the stored exemplars. The exemplars are represented as points in a multidimensional psychological space. Similarity is a nonlinearly decreasing function of distance in the space, $s_{ij} = \exp(-cd_{ij}^p)$, where $c$ is an overall scaling [or sensitivity] parameter and $p$ defines the form of the similarity gradient. (The default assumptions are $p = 1$, which defines an exponential similarity gradient, and $p = 2$, which defines a Gaussian gradient. The $p = 1$ assumption is now used almost ubiquitously.…) Distances among exemplars are computed with the weighted Minkowski power model, $d_{ij} = [\sum w_m|x_{im} - x_{jm}|^r]^{1/r}$, where $x_{im}$ is the psychological value of exemplar $i$ on dimension $m$, the $w_m$ are free parameters reflecting the attention given to each dimension $m$, and the parameter $r$ defines the distance metric in the space. Furthermore, the default parameter settings in the model are that $r = 1$ when stimuli vary along highly separable dimensions (city-block metric) and $r = 2$ when stimuli vary along highly integral dimensions (Euclidean metric). Finally, in a bias-free, two-category experiment, the probability that . . . [the *ith*] item is classified in Category $A$ is given by . . . $[p_i] = SA^\gamma/(SA^\gamma + SB^\gamma)$,

> where *SA* and *SB* give the summed similarity of the item to [the] Categories A and B [exemplars], respectively, and $\gamma$ is a response-scaling parameter.

An additional central theoretical assumption of the GCM is the *attention-optimization hypothesis*, which holds that participants are inclined to distribute attention among the stimulus dimensions to optimize their performance (Nosofsky, 1986). In Nosofsky's (1998, p. 330) own words:

> [In] situations in which a single dimension is relevant for performing a classification, observers will be inclined to attend selectively to this dimension[.] . . . [I]f two dimensions are both relevant for performing a classification, then . . . one does not expect selective attention to be as extreme[.]

The attention-optimization hypothesis has resisted formalization in a non-Bayesian framework, but can be expressed in a Bayesian framework by specifying an informative prior on the attention parameter $w_m$ (Vanpaemel & Lee, 2012b). In particular, by assuming a Beta($\alpha$, $\beta$) distribution as a prior for $w_m$, and setting the values for $\alpha$ and $\beta$ depending on the category structure to be learned, it is possible to capture the expectations about attentional allocation.[1] For the sensitivity and response scaling parameters, no such theory is available. Consequently, uniform priors for $c$ and $\gamma$, using 10 as the upper bound, are assumed.

The MPM differs in two aspects from the GCM, First, *SA* does not give the summed similarity of the item to the Category A exemplars, but the similarity of the item to the Category A prototype, which is constructed by averaging all the Category A exemplars. Second, it uses $p_i = SA/(SA + SB)$ as a decision rule. Using the decision rule of the GCM would make the MPM non-identifiable, in the sense that the response scaling parameter $\gamma$ cannot be estimated separately from the sensitivity parameter $c$ (Ashby & Maddox, 1993; Nosofsky & Zaki, 2002).

### 2.3. The response scaling parameter

In its original form (Nosofsky, 1984, 1986), the GCM used the decision rule from the MPM, or alternatively, the decision rule from above with $\gamma = 1$. Nosofsky later reflected on this particular form of the decision rule as follows (Nosofsky & Johansen, 2000, p. 388):

> [W]e know of no very strong reason why one should assume $\gamma = 1$. In an early article that tested the context model [by Medin and Schaffer (1978), which the GCM generalized and inherited its original decision rule from], Medin and Smith (1981) justified use of this particular response rule simply by saying, "The best defense of the response rule is that it is a fair approximation and that it seems to work" (p. 250).

The modified decision rule, which was inspired by Ashby and Maddox (1993) and Maddox and Ashby (1993), allows a more deterministic response than the original one.[2] When $\gamma = 1$, the model assumes that learners respond by probability matching, which means that the probability of choosing an alternative reflects the relative summed similarities. When $\gamma > 1$, the model assumes that learners respond with more extreme probabilities (i.e., closer to 0 and 1, or more deterministically), following the category with the larger summed similarity. In the extreme case of $\gamma = \infty$, learners always respond with the category with the largest summed similarity. Nosofsky and Palmeri (1997) provided a direct process interpretation for the response scaling parameter, by showing it corresponds to response caution. Additional processing interpretations were provided by Navarro (2007). In the remainder of this paper, the GCM with response scaling parameter will be referred to as $GCM_c$, where $c$ stands for current, and the constrained version without response scaling parameter (or equivalently, with $\gamma = 1$) as $GCM_r$, for restricted.

## 3. Deep waters

While exemplar theorists have been quick to rely on the exemplar model with response scaling parameter (i.e., $GCM_c$) when comparing exemplar-based and prototype-based accounts, prototype theorists have been much more reluctant to use it, and often choose to rely on a version without response scaling parameter (i.e., $GCM_r$) (e.g., Homa, Proulx, & Blair, 2008; Minda & Smith, 2002; Smith et al., 1997). The key reason prototype theorists consider the response scaling parameter problematic is that it leads to "a more complex and mathematically powerful version of the exemplar model" (Smith & Minda, 2002, p. 807).

Let us examine these concerns about complexity and power in more detail. First, prototype theorists have claimed that the response scaling parameter "can enormously increase the exemplar model's mathematical complexity . . ., so that . . . [$GCM_c$] verges on being the powerful classifier commonly used by professional statisticians" (Smith, 2013, p. 315). While $GCM_c$ surely is higher in parametric complexity than its constrained version $GCM_r$, this argument has in itself very little merit. First, the number of parameters is a poor measure of model complexity, as it ignores the influence of the functional form (Lee, 2002; Myung, Balasubramanian, & Pitt, 2000; Myung & Pitt, 1997) and of the prior (Vanpaemel, 2009) on complexity. Second, building models that are low in complexity and thus high in empirical content is a goal model builders should strive to (see e.g., Vanpaemel & Lee, 2012a, for more discussion), but being complex is in itself not a bad feature of a model (neither is being used by professional statisticians). The processes the models aim to approximate are likely to be very complex, so it is unreasonable to expect a natural upper bound for the complexity of a model.

Further, prototype theorists clarified why they deemed the complexity of $GCM_c$ troubling, by stating that "the mathematical complexity of …[$GCM_c$] could give it an unfalsifiable range of predictions" (Minda & Smith, 2001, p. 794). There is a tight coupling between complexity and falsifiability. Falsifiability is of huge importance when building and evaluating models, and has been shamefully overlooked in current day modeling, as will be briefly touched upon in the discussion. However, lacking a rigorous definition and demonstration of infalsifiability, this argument is rather weak

Smith and Minda (2002, p. 809) further claim that the response scaling parameter "grants the exemplar model asymmetric power and mathematical complexity over the prototype model and makes it difficult to compare [MPM and $GCM_c$ on goodness of] fit indices". This difficulty in comparing goodness-of-fit indices is also acknowledged by exemplar theorists, who are careful to acknowledge that "[d]rawing strong conclusions [based on a superior fit of the $GCM_c$ over the MPM] is particularly difficult in view of the

[1] The expectation for the attentional allocation is hypothesis driven, but it cannot be formalized generically, as the exact prior for the attention parameter depends on the category structure to be learned. For the Angle, Criss-cross, Diagonal, and Size designs to be discussed later, the attention-optimization hypothesis was captured in, respectively, a Beta(8, 1), Beta(2, 2), Beta(2, 2) and Beta(1, 8) prior for the parameter reflecting attention to the horizontal dimension (see Vanpaemel & Lee, 2012b, for a discussion of these choices). A stronger implementation of the attention-optimization hypothesis – assuming Beta(12, 1), Beta(12, 12), Beta(12, 12) and Beta(1, 12) as a prior – leads to quantitatively similar and qualitatively identical results. Further, when the attention-optimization hypothesis was not instantiated in the model, by assuming a Beta(1, 1) or uniform prior for $w_m$, the qualitative results did not change.

[2] Note that also Nosofsky (1991) presented a modified version of the original GCM to allow deterministic responses.

fact that the ... $GCM_c$ makes use of an additional free parameter relative to the ... MPM" (Nosofsky & Johansen, 2000, p. 389).

This argument is, at least partly, valid, and touches the core of the matter. It is well known that goodness-of-fit measures, such as the sum of squared deviations between the observed and predicted performance (SSD), or the maximum likelihood, ignore the complexity of a model (e.g., Lee, 2002; Myung, 2000; Myung & Pitt, 1997; Pitt & Myung, 2002). If models that differ in complexity are evaluated on their goodness-of-fit, the most complex model tends to be favored. In an extreme case, where the simpler model is a nested version of the more complex model, the simpler model will never fit the data better than the more complex model. Thus, it is correct to point out that, in the light of the likely complexity difference between MPM and $GCM_c$, goodness-of-fit measures should be interpreted with caution.[3]

Because of the difficulty of comparing the goodness-of-fit of MPM and $GCM_c$, prototype theorists preferred to contrast the MPM to the $GCM_r$, which are balanced in the number of parameters (though see Footnote 3). An important downside of this strategy, however, is that exemplar theorists contest their findings, because these are based on a version of the exemplar that arbitrarily fixes $\gamma$ to 1 (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002; Zaki, Nosofsky, Stanton, & Cohen, 2003). Exemplar theorists, on the other hand, preferred to contrast the MPM to the $GCM_c$. A downside of this strategy, however, is that prototype theorists do not accept their findings, because the difference in number of parameters between the models makes complexity differences likely, and thus comparing goodness-of-fit meaningless.

So we are faced with a thorny situation. Not including the response scaling parameter is unacceptable for exemplar theorists, as doing so corresponds to using an arbitrary and abandoned version of their model. Including the response scaling parameter is intolerable for prototype theorists, as doing is thought to disadvantage their model. These oppositions have paralyzed the prototype vs exemplar debate. To make progress again, it seems necessary to decide which side is right and which side is wrong. However, the situation is further complicated by the realization that both arguments are correct: it is academically dishonest to rely on a model that the adversary considers arbitrary and outdated. And it is statistically unsound to compare models potentially differing widely in complexity based on their goodness-of-fit.

## 4. A bridge over deep waters

The solution is surprisingly straightforward. If it is difficult to compare models on goodness-of-fit indices, stop comparing models on goodness-of-fit indices. If one contrasts the MPM to the $GCM_c$ using a model selection measure that takes both goodness-of-fit *and* the complexity of the models into account, the reservations of each side vanish: those of the exemplar theorists, because their preferred version is considered; and those of the prototype theorists, because their fear that their model is unfairly disadvantaged is attenuated.

The solution thus involves stopping using goodness-of-fit to compare models, and instead comparing models on their generalizability — the detailed balance between goodness-of-fit and complexity (Pitt, Myung, & Zhang, 2002). One dominant measure of generalizability is the Bayes factor (Jeffreys, 1935; Kass & Raftery, 1995), which has found ample discussion and use in psychology (e.g., Dienes, 2011; Lee, 2004; Lee & Wagenmakers, 2005; Lodewyckx et al., 2011; Mulder et al., 2009; Myung & Pitt, 1997; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Vanpaemel, 2010; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009).[4]

After explaining how the Bayes factor can be used to compare models that differ in complexity, I revisit four previously published data sets, using three modeling strategies: the first is the prototype theorist's dominant strategy: comparing MPM to $GCM_r$ using a measure of goodness-of-fit. The second is the exemplar theorist's dominant strategy: comparing MPM to $GCM_c$ using a measure of goodness-of-fit. These two strategies have lead to theoretical standstill. The third is the strategy of progress: comparing MPM to $GCM_c$ using the Bayes factor.

### 4.1. Dealing with complexity: best vs average fit

The first two strategies will rely on the maximum likelihood as a measure of goodness-of-fit, whereas the third strategy will rely on the marginal likelihood as a measure of generalizability. The marginal likelihood is the core of the better known Bayes factor, which is defined as the ratio of marginal likelihoods. Both the maximum and the marginal likelihood reflect the fit of a model to data. The crucial difference is that the maximum likelihood corresponds to the *best* fit, while the marginal likelihood corresponds to the *average* fit (for a brief formal treatment, see the Appendix). This difference implies that the maximum likelihood but not the marginal likelihood ignores model complexity.

To illustrate how both approaches deal with complexity differently, consider deciding which of two players is the best in a game where a score ranges between 0 and 500 (see also Chapter 7 in Lee & Wagenmakers, 2014). Both players play five games against a computer, collecting scores of (307, 303, 294, 298, 405) and (94, 100, 105, 427, 96), respectively. Player A seems to be a consistently good player, while Player B is clearly a bad player. Both players apparently got lucky once, yielding an exceptionally high score. From a best performance perspective, Player B is the clear winner: his best score (427) is higher than his opponent's best score (405). From an average performance perspective, in contrast, Player A is the better player: his average score (321) is higher than the average score of player B (164). The best performance approach rewards Player B for a lucky extremity, whereas the average performance approach rewards Player A for his skillful consistency.

The difference between both approaches is even more striking if one of both players is given the opportunity to play five additional games. Under the best performance rule, there is nothing to loose, only to gain by playing more games. If a player gets lucky again, the score increases; if not, the score stays the same. Under the average

[3] The argument is only partly valid for at least two reasons, both related to the unfortunate practice of equating model complexity with the number of parameters. First, when counting the number of parameters, MPM is less complex than $GCM_c$. However, when the other dimensions of complexity are taken into account, there is no a priori reason to expect that this order relation still holds. Myung, Pitt, and Navarro (2007) used a model recovery study and a more sophisticated measure of model complexity to investigate the relative complexities of MPM and $GCM_c$. They found that, in two different designs, MPM was indeed less complex than $GCM_c$, providing post hoc support for the claim about asymmetric complexity. Second, as models with the same number of parameters are not necessarily equally complex, equal parametric complexity by itself does not warrant the use of goodness-of-fit as a model selection measure. Even when MPM and $GCM_r$ are compared, goodness-of-fit can be misguided as a model selection criterion. Vanpaemel and Storms (2010) demonstrated, using model recovery studies, that MPM and $GCM_r$ are similar in complexity in a range of experimental designs, lending some post hoc credence to the prototype theorist's strategy of ignoring complexity differences when comparing both models.

[4] Several other measures of generalizability have been proposed, including AIC, BIC, and cross validation (CV) (see two special issues of the *Journal of Mathematical Psychology* on model selection (Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006)). Although some of these measures have occasionally been used to compare prototype and exemplar models (e.g., Maddox & Ashby, 1993; McKinley & Nosofsky, 1995; Olsson, Wennerholm, & Lyxzèn, 2004), most comparisons relied on simple goodness-of-fit statistics, such as SSD. The Bayes factor is compared to these other measures in the discussion.

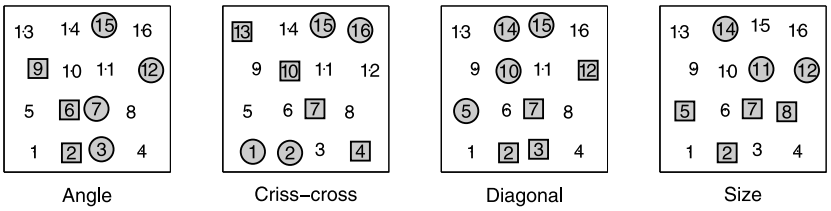


**Fig. 1.** The four category structures used by Nosofsky (1989). Each numbered location corresponds to a stimulus. A stimulus that is assigned to one of the two categories is indicated by a square (category A) or a circle (category B). The remaining stimuli are unassigned. Based on Nosofsky (1989, Figs. 1 and 3).

performance rule, there is a lot to loose for Player B. Most likely, the average score on the additional games will hover around 100, and the average score will decrease when playing more games, as the effect of the single lucky game will be attenuated. If player A can play more games, there is less to loose. Most likely, the score on the additional games will hover around 300, and the average score will be much less affected. Thus, when one of the players has more opportunities to play, it is unfair to compare them by considering the best performance, as one of them had more opportunities to get lucky. When considering the average performance, a fair comparison between players is possible, since lucky shots are less influential. It does not matter that one of the players had more opportunities to play (practice and fatigue effects aside).

To appreciate the link between gaming and model complexity, think of each player as a model, and of each score as the fit between the data and a prediction at some parameter value, which is formally given by the likelihood. Complexity can be intuitively understood as the number of (distinguishable) predictions a model makes (Myung et al., 2000; Vanpaemel, 2009), and thus corresponds to the number of games played. Under the best performance rule, making a model more complex – playing more games – is not penalized in any way. Under the average performance rule, in contrast, the number of games or the complexity is taken into account. Making a model more complex is not harmful for the model if the additional predictions yield good fits, whereas making a model more complex in a way that the additional predictions produce bad fits is penalized. Crucially, by considering the average fit, both models can be compared fairly, even if they differ in complexity.

This illustration served to highlight the difference between considering the average and the best fit. The marginal likelihood differs from the maximum likelihood in a second aspect. The former, but not the latter, is sensitive to the prior distribution over the parameter values. In the gaming illustration, a prior would correspond to a player indicating, before the start of the game, which scores should be given more weight than others when computing the average score.

### 4.2. Application

I compare the three strategies described above focusing on four previously published data sets from Nosofsky (1989). The main difference between the data sets was the category structure to be learned, as shown schematically in Fig. 1. I will refer to these data sets by using the names displayed in this figure. The data take the form of the number of times each stimulus was classified in Category A, aggregated across participants and presentations, and are provided in the original publications. Further, the data, as well as computational details, can be found in the Appendices of Vanpaemel and Storms (2010).

Table 1 provides the results of the first two strategies. Under the prototype theorist's strategy, the comparison of interest is between MPM and $GCM_r$. For each data set, the exemplar model outperforms the prototype model. The evidence, as assessed by the difference in maximum likelihood, ranges from as small as .36 (Diagonal) to as overwhelming as 183.49 (Criss-cross). The exemplar theorist's strategy involves comparing MPM and $GCM_c$. Since $GCM_r$ is a nested version of $GCM_c$, the fit of $GCM_c$ will necessarily be as good as or better than the fit of $GCM_r$. Hence, it is no surprise that the exemplar is still preferred for all data sets, and that the evidence for the exemplar model has increased. The increase in evidence ranges from .52 (Diagonal) to 16.18 (Angle). So the exemplar theorist's strategy yields more convincing evidence for the exemplar model than the prototype theorist's strategy.

**Table 1**
Goodness-of-fit measures (negative maximized log-likelihood) for MPM, $GCM_r$ and $GCM_c$.

| | MPM | $GCM_r$ | $GCM_c$ |
|---|---|---|---|
| Angle | 103.78 | 56.75 | 40.57 |
| Criss-cross | 240.19 | 56.70 | 55.61 |
| Diagonal | 49.44 | 49.08 | 48.56 |
| Size | 44.27 | 40.77 | 40.12 |

Notes: MPM = Multiplicative Prototype Model; $GCM_r$ = Generalized Context Model with response scaling parameter restricted to 1. $GCM_c$ = Generalized Context Model with response scaling parameter.

The third strategy uses the Bayes factor, which allows the comparison of MPM and $GCM_c$, despite their unbalancedness in number of parameters. To increase interpretability, a Bayes factor is reported in terms of the posterior probability of the $GCM_c$ when compared to the MPM, assuming equal priors for both models (see Eq. (A.5)), which indicates the relative support of a model given by the data and the prior belief. For the Angle, Criss-cross and Size data sets, the conclusions are identical to the conclusions based on the previous strategies. For the first two, the exemplar model is massively preferred, with a posterior probability close to 1. For the Size data set, the Bayes factor still indicates a preference for the exemplar model, with a posterior probability of about .75. For the Diagonal data set, the prototype model is slightly preferred over the exemplar model, with a posterior probability of about .39.

Thus, using the third strategy – i.e., using the exemplar model as developed by exemplar theorists, and using a model selection measure that takes complexity into account – provides *less* support for the exemplar model than when using the exemplar theorist's strategy, and – this might be surprising, especially for prototype theorists – also provides *less* support for the exemplar model than when using the prototype theorist's strategy. Ironically, by sticking to an old and largely abandoned version of the exemplar model and an old and largely useless model selection measure, prototype theorists might have missed important chances to provide evidence for their model.

## 5. Discussion

When contrasting the prototype and exemplar accounts of category learning, some prototype theorists heavily relied on the "principle of comparing models that are equivalent, balanced, and similar in assumptions and parameters" (Minda & Smith, 2001, p. 794). This principle implied that they had to use a restricted version of the exemplar model: to ensure that the exemplar model has the same number of parameters as the prototype model, they

constrained the response scaling parameter of the exemplar model to 1. Since exemplars theorists saw no valid reason for such an unusually strong commitment, they did not accept the results.

In the absence of sophisticated model selection measures that are appropriately sensitive to complexity, the above principle is probably the safest bet. By following this principle, there is at least some hope that differences in fit actually mean something (although see Footnote 3). But when model selection measures are available that balance goodness-of-fit and complexity, it is very hard to think of a good reason to rely on this principle. Provided that a model selection measure that is sensitive to complexity, such as the Bayes factor, is used, there is no need to limit model comparison to models that are balanced in number of parameters. In sum, these prototype theorists have been fighting the wrong battle. Rather than refusing to include the response scaling parameter, they should have refused to use the outdated strategy of comparing models based on their goodness-of-fit only.

### 5.1. Implications for the prototype vs exemplar debate

For the current selection of four data sets, using the Bayes factor to compare MPM and $GMC_c$ decreased the support for the exemplar model. This, however, does not imply that if both prototype and exemplar theorists would have relied on the Bayes factor to compare these models, there would have been more evidence for the prototype model. The four data sets considered here are not random samples from, and are not in any way representative for the vast collection of category learning data sets available. It is thus impossible to predict in which direction the evidence will change when considering a larger set of data sets. Most notably, all four data sets relied on the reported, aggregated data, while it has been argued that the response scaling parameter is especially important when modeling data from individual participants (McKinley & Nosofsky, 1995; Nosofsky & Johansen, 2000; Nosofsky & Palmeri, 1997).

It might seem that the prototype vs exemplar debate could be advanced by re-analyzing a larger selection of data sets using the third strategy demonstrated in this paper. There are, however, a few caveats when pursuing that direction. First, applying this approach to individual data would ignore similarities between people. For individual data, a Bayesian hierarchical latent mixture approach (see, e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014) seems better suited. This approach affords capturing both continuous and discrete individual differences, and can be used to tackle issues of model selection, such as the prototype vs exemplar debate. Second, it has been argued that a good (best or average) fit is not necessarily meaningful or persuasive. Roberts and Pashler (2000, 2002) have convincingly pointed out that a good fit is only persuasive if a bad fit was plausible, or, equivalently, if the model was falsifiable. Neither the maximum nor the marginal likelihood is informative about the plausibility of a bad fit. Penalizing complexity, as the Bayes factor does, does increase the likelihood of a plausible misfit, but does not guarantee it. Evaluating models in a way that takes the plausibility of a bad fit, and hence the falsifiability of a model, into account can be done using a combination of the prior predictive distribution and the data prior, which quantifies – before the data have been observed, and irrespective of any model – which outcomes are plausible for a certain experiment (Vanpaemel, submitted for publication). Finally, I and others have argued that the focus on exemplar and prototype models provides a rather limited window on the wealth of possible representations people might realistically entertain (e.g., Anderson, 1991; Griffiths, Canini, Sanborn, & Navarro, 2007; Love, Medin, & Gureckis, 2004; Rosseel, 2002; Vanpaemel & Storms, 2008). Therefore, a large scale re-analysis might benefit from including a larger set of representations.

### 5.2. Other measures of generalizability

The solution to the stalemate in the prototype vs exemplar debate advocated here – use an appropriate model selection measure – is conceptually a rather simple one. Several implementations of this solution other than the Bayes factor can be used, including AIC, BIC, and CV. When computationally feasible, the Bayes factor is to be preferred over these other measures of generalizability, because they fail to take all aspects of complexity into account. Both AIC and BIC ignore the contribution of functional form to complexity, and AIC, BIC and CV ignore the contribution of the prior to complexity. The Bayes factor, in contrast, is sensitive to both the functional form, through its reliance on the likelihood, and to the prior. Unlike for the Bayes factor, for AIC, BIC and CV, there is no difference between a version of the GCM that assumes an informative Beta(8, 1) prior for the attention parameter and a version of the GCM that assumes a Beta(1, 1) prior. This insensitivity to the prior is unfortunate, as both model versions clearly make different theoretical commitments: the former, but not the latter, instantiates the attention-optimization hypothesis (see Footnote 1).

The sensitivity of the Bayes factor to the prior could have been helpful for alleviating another concern some prototype theorists have raised against the response scaling parameter: the fact that the response scaling parameter lacks theoretical appeal. While the Bayes factor does not automatically solve this concern – after all, it is a model selection measure, not a theory building machine – it does push modelers in the right direction. The use of a uniform prior for this parameter clearly points to a gap in current theory on response determinism. Using the Bayes factor invites exemplar theorists to sharpen their theorizing about response determinism in order to specify an informative prior for the response scaling parameter, just like additional theorizing, in the form of the attention-optimization hypothesis, helps determining the prior for the attention parameter. The adoption of the Bayes factor not only invites but also encourages formalizing theory about response determinism in the parameter prior, since specifying theoretically informed priors in the exemplar model can increase the evidence for the exemplar model, at least if the theory accords with the data.

Without relying on the Bayes factor, there is no incentive to develop theory about response determinism, let alone formalize it in the model. Thus, using the Bayes factor not only provides a bridge over the deep waters between prototype and exemplar theorists, it also forces modelers at both sides to do their job as theory builders properly. They should not only build theories about how parameters relate to each other to give rise to behavior. A theory is only complete if it can specify which parameter values are expected to give rise to this behavior (Vanpaemel, 2010). More generally, the Bayes factor urges modelers to make their models more theoretically complete, and to increase what Popper (1959) called the *empirical content* of the models. As a higher empirical content corresponds to more knowledge about the world, the Bayes factor thus encourages modelers to increase our understanding of the world (Glöckner & Betsch, 2011; Vanpaemel & Lee, 2012a).

Some might still be wary to use the Bayes factor, exactly because of its sensitivity to the prior. Their fear is that, since widely different conclusions can be reached if the prior is changed, the prior introduces some degree of arbitrariness in the conclusions. What is often less appreciated is that many (Bayesian and non-Bayesian) model selection measures are sensitive to the likelihood: widely different conclusions can be reached if the likelihood is changed. For example, prototype theorists used the likelihood of $GCM_r$ and found evidence for the prototype model, while exemplar theorists used the likelihood of $GCM_c$ and found evidence for the exemplar model. Different likelihoods, different conclusions: the likelihood surely introduces some degree of arbitrariness in the conclusions.

It seems hypocritical to complain that model selection is sensitive to the prior, while at the same time being untroubled by the fact that model selection is sensitive to the likelihood. As the case of the response scaling parameter demonstrates, there is nothing sacrosanct about the likelihood. The likelihood is subject to similar levels of arbitrariness (or subjectivity if you like) as the prior. The likelihood of "the" exemplar model is different for prototype and exemplar theorists. So, just like the use of a prior necessitates a prior sensitivity analysis (see Footnote 1), the use of a likelihood necessitates a likelihood sensitivity analysis. As Gelman and Robert (2013, p. 4) most elegantly put it, we should all stop "strain[ing] on the gnat of the prior distribution while swallowing the camel that is the likelihood".

### 5.3. Conclusion

Adoption of the Bayes factor would serve the prototype vs exemplar debate well in two respects. First, the use of the Bayes factor affords the possibility of comparing exemplar and prototype models in a way both exemplar and prototype theorists should feel more comfortable with. Second, the use of the Bayes factor encourages both exemplar and prototype theorists to think carefully about what their parameters mean, and to formalize theory about these parameters in the prior.

## Acknowledgments

The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (OT/11/032) and by the Interuniversity Attraction Poles Programme financed by the Belgian government (IAP/P7/06).

## Appendix. Maximum vs marginal likelihood

Consider a model $M$ with parameter $\theta$ and likelihood function $p(d \mid \theta, M)$. The likelihood denotes the probability of having observed data $d$ given parameter $\theta$ and model $M$, and can be interpreted as a measure of fit, i.e., a match between data and model.

The largest possible value of the likelihood, across different parameters values, indicates the model's best possible performance:

$$p\left(d \mid \hat{\theta}, M\right) = p\left(d \mid \theta = \hat{\theta}, M\right) = \max_{\theta} p(d \mid \theta, M), \tag{A.1}$$

where

$$\hat{\theta} = \arg\max p(d \mid \theta, M). \tag{A.2}$$

The maximum likelihood $p\left(d \mid \hat{\theta}, M\right)$ is a measure of goodness-of-fit, and reflects the best possible fit of the model, given the opportunity to pick those parameter values that make a prediction that is closest to the observed data. The maximum likelihood is a local measure, in the sense that it considers a single parameter value only, and a posterior measure, in the sense that it considers a prediction made based on observed data (Vanpaemel, 2010). Technically, computing the maximum likelihood relies on maximizing the likelihood.

The average value of the likelihood, across different parameters values weighted by their prior $p(\theta \mid M)$, indicates the model's average performance:

$$p(d \mid M) = \int_{\Omega} p(d \mid \theta, M)\, p(\theta \mid M)\, \mathrm{d}\theta, \tag{A.3}$$

where $\Omega$ indicates the range of the parameter vector $\theta$. The marginal likelihood $p(d \mid M)$ is a measure of generalizability, and reflects the average fit of the model, comparing the data to a range of model predictions at different parameter values. The marginal likelihood is a global measure, in the sense that it considers all parameter values entertained by the model, and a prior measure, in the sense that it considers predictions made independent of (and thus possibly before) observed data. Technically, computing the marginal likelihood relies on integrating the likelihood, as weighted by the prior.

One attractive feature of using the marginal likelihood is that it affords a natural expression of relative model preference, in terms of the posterior probability of each model under consideration, indicating the relative plausibility of the model, based on the data and the model prior. Given two models $M_1$ and $M_2$, the posterior probability for $M_1$ follows from Bayes' rule:

$$\begin{aligned} p(M_1|d) &= \frac{p(d|M_1)p(M_1)}{p(d)} = \frac{p(d|M_1)p(M_1)}{p(d|M_1)p(M_1) + p(d|M_2)p(M_2)} \\ &= \frac{1}{1 + BF_{21}\frac{p(M_2)}{p(M_1)}}, \end{aligned} \tag{A.4}$$

where $p(M_1)$ reflects the prior probability of model $M_1$, and $BF_{21} = \frac{p(d|M_2)}{p(d|M_1)}$ is the Bayes factor comparing Model 2 to Model 1. When all models under consideration are considered being equally plausible a priori (which is often the case, though see Lee & Vanpaemel, 2008; Vanpaemel, 2011), the above equation reduces to

$$p(M_1|d) = \frac{1}{1 + BF_{21}}. \tag{A.5}$$

## References


Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Gelman, A., & Robert, C. P. (2013). "Not only defended but also applied": The perceived absurdity of Bayesian inference. *The American Statistician*, *67*, 1–5.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: shortcomings and remedies. *Judgment and Decision Making*, *6*, 711–721.

Griffiths, T. L., Canini, G. K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 323–328). Austin, TX: Cognitive Science Society.

Homa, D., Proulx, M. J., & Blair, M. (2008). The modulating influence of category size on the classification of exception patterns. *The Quarterly Journal of Experimental Psychology*, *61*, 425–443.

Hull, C. L. (1920). Quantitative aspects of evolution of concepts. *Psychological Monographs: General and Applied*, *28*.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In *Mathematical proceedings of the Cambridge philosophical society*: *Vol. 31* (pp. 203–222). Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Lee, M. D. (2002). Generating additive clustering models with minimal stochastic complexity. *Journal of Classification*, *19*, 69–85.

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, *48*, 310–321.

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*, *32*, 1403–1424.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*, 331–347.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, *53*, 49–70.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 128–148.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.
Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 241–253.
Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 775–799.
Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275–292.
Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530–546.
Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.
Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175.
Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Guest editors' introduction to special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.
Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.
Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, 1043–1050.
Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, *51*, 85–98.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
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.
Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception and Psychophysics*, *45*, 279–290.
Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27.
Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K Estes*: *Vol. 1* (pp. 149–167). Hillsdale, NJ: Lawrence Erlbaum.
Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 322–339.
Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375–402.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.
Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924–940.
Olsson, H., Wennerholm, P., & Lyxzèn, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 936–941.
Pitt, M. A., & Myung, I. J. (2002). 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 of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.
Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.
Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 392–407.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.
Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review*, *109*, 605–607.
Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178–210.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
Smith, J. D. (2013). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin & Review*, 1–20.
Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3–27.
Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 800–811.
Smith, J. D., Murray Jr, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 659.
Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. *Advances in Neural Information Processing Systems*, *22*, 1919–1927.
Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.
Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*, 106–117.
Vanpaemel, W. (2015). Complexity, data prior and the persuasiveness of a good fit: Comment on Veksler, Myers and Gluck (2015) (submitted for publication).
Vanpaemel, W., & Lee, M. D. (2012a). The Bayesian evaluation of categorization models: Comment on Wills and Pothos (2012). *Psychological Bulletin*, *138*, 1253–1258.
Vanpaemel, W., & Lee, M. D. (2012b). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.
Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732–749.
Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, *42*, 421–437.
Wagenmakers, E.-J., & Waldorp, L. (2006). Guest editors' introduction to special issue on model selection: Theoretical developments and applications. *Journal of Mathematical Psychology*, *50*, 99–100.
Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*, 752–760.
Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1160–1173.