Psychonomic Bulletin & Review
2007, 14 (6), 1043-1050

Does response scaling cause the generalized context model to mimic a prototype model?

Jay I. Myung and Mark A. Pitt
Ohio State University, Columbus, Ohio

and

Daniel J. Navarro
University of Adelaide, Adelaide, Australia

Smith and Minda (1998, 2002) argued that the response scaling parameter γ in the exemplar-based generalized context model (GCM) makes the model unnecessarily complex and allows it to mimic the behavior of a prototype model. We evaluated this criticism in two ways. First, we estimated the complexity of the GCM with and without the γ parameter and also compared its complexity to that of a prototype model. Next, we assessed the extent to which the models mimic each other, using two experimental designs (Nosofsky & Zaki, 2002, Experiment 3; Smith & Minda, 1998, Experiment 2), chosen because these designs are thought to differ in the degree to which they can discriminate the models. The results show that γ can increase the complexity of the GCM, but this complexity does not necessarily allow mimicry. Furthermore, if statistical model selection methods such as minimum description length are adopted as the measure of model performance, the models will be highly discriminable, irrespective of design.

How do humans learn to categorize objects (e.g., dogs) that vary along multiple dimensions (e.g., size, shape, color, texture) into psychologically equivalent categories? This question has attracted a great deal of interest in cognitive science and has led to diverse conceptualizations of the cognitive processes underlying category formation and structure. In prototype theories, humans are assumed to extract commonalities across instances of a class and encode these generalizations in memory (Reed, 1972; Smith & Minda, 1998). In exemplar theories, on the other hand, humans encode each instance of the class that is encountered, thereby preserving much of the detail present in the input (Medin & Schaffer, 1978; Nosofsky, 1986).

Given such different theories and the fact that quantitative models of each type have been put forth, one might think that decisive evidence favoring one position would have been generated long ago, but in recent years, the debate in this field has actually intensified. One reason for this stems from disagreement about the proper quantitative formulation of the exemplar model, of which there are two versions—the original generalized context model (GCM; Nosofsky, 1986) and an elaborated version we refer to as GCMg (Ashby & Maddox, 1993; McKinley & Nosofsky, 1995), which includes an additional response-scaling parameter, γ.[1]

Smith and Minda (1998, 2002) expressed severe reservations about γ, questioning its validity and arguing that it makes GCMg so adept at fitting behavioral data that it becomes a "prototype in exemplar clothing" (Smith & Minda, 1998, p. 1413). They supported this claim with simulation data showing that GCMg can mimic an additive prototype model quite well. These researchers were sufficiently wary of the additional data-fitting power that γ adds to GCM that in subsequent studies (Minda & Smith, 2001, 2002) they compared multiple prototype models (PRT, additive and multiplicative versions[2]; see Minda & Smith, 2001) with only the original GCM, to ensure that the models were equated in their numbers of parameters.

Nosofsky and colleagues (Nosofsky & Zaki, 2002; Zaki, Nosofsky, Stanton, & Cohen, 2003) defended the introduction of the response-scaling parameter, arguing that, among other things, it was necessary to capture the deterministic behavior that participants exhibit early in learning, when they tend to focus on a single dimension of a stimulus. Probably most convincing in countering the claims of Smith and Minda (1998) are the results of two simulations that showed that data generated by PRT were fitted better by PRT than by GCMg. If GCMg were in fact equivalent to PRT, then its fits should always be comparable with those of PRT.

Although illustrative, the preceding evidence needs to be pursued to draw strong conclusions about whether the response-scaling parameter does or does not cause GCMg to mimic PRT. What is needed is an understanding of just how data-fitting performance changes when γ is added to GCM, with respect to PRT in particular. In this article, we used statistical model selection methods to provide this understanding. Analyses were performed that not only quantified the extent to which the data-fitting abilities of GCM increased when γ was added, but they also, impor-

J. I. Myung, myung.1@osu.edu

Copyright 2007 Psychonomic Society, Inc.

tantly, determined whether this increase was in fact due to GCMg's ability to mimic PRT. We begin by reviewing the models and describing the quantitative methods used to analyze them. This is followed by the application of these methods in two experimental designs, chosen for their abilities to discriminate the two models.

Categorization Models

For the three models, the probability of deciding that the ith stimulus S_i belongs to category A, $P(A|S_i)$, is given by a multinomial probability that is proportional to the similarity of stimulus S_i to category A. For exemplar models, the category similarity is found by summing across individual stimulus similarities, whereas for prototype models, a single idealized stimulus S_A is used. To calculate stimulus similarities, it is assumed that each stimulus is mentally represented as a point located in an m-dimensional Minkowski space, and the similarity between any two stimuli is assumed to decrease exponentially with the distance between them. Therefore, the similarity s_{ij} between the ith and jth stimuli is given by

$$s_{ij} = \exp\left[-\lambda\left(\sum_{t=1}^{T} w_t \,|\, x_{it} - x_{jt} \,|^r\right)^{1/r}\right], \tag{1}$$

where x_{it} is the coordinate value of S_i along dimension t. In this equation, w_t denotes the proportion of attention applied to the tth dimension $(\sum_{t=1}^{T} w_t = 1)$, r determines the distance metric that applies in the space, and λ denotes the steepness of the exponential decay (called the specificity parameter).

One methodological issue to keep in mind is that categorization models can vary across studies, making comparisons difficult. However, because one of our goals was to evaluate the effects of design differences on model discriminability, we held models constant across designs. This was done in two ways. First, the metric parameter r was fixed at 1 (i.e., city block distance) for all three models, since the stimulus dimensions are assumed to be perceptually separable (Garner, 1974). Second, a guessing parameter, $q (0 < q < 1)$, was introduced to each model. The role of this parameter was to assume that, with probability q, a participant chose a category at random. With these changes, the probability that the observed stimulus S_i belongs to the category A is defined as

$$\text{PRT}: \quad P(A|S_i,\theta) = \frac{q}{2} + (1-q)\left(\frac{s_{iA}}{\sum_{C} s_{iC}}\right)$$

$$\text{GCM}: \quad P(A|S_i,\theta) = \frac{q}{2} + (1-q)\left(\frac{\sum_{x\in A} s_{ix}}{\sum_{C}\sum_{y\in C} s_{iy}}\right)$$

$$\text{GCMg}: \quad P(A|S_i,\theta) = \frac{q}{2} + (1-q)\left[\frac{\left(\sum_{x\in A} s_{ix}\right)^{\gamma}}{\sum_{C}\left(\sum_{y\in C} s_{iy}\right)^{\gamma}}\right], \tag{2}$$

where the sum over C is taken over all relevant categories, γ is the response-scaling parameter, and θ represents the set of model parameters, $\theta = (w_1, \ldots, w_{T-1}, \lambda, q)$ for PRT and GCM and $\theta = (w_1, \ldots, w_{T-1}, \lambda, \gamma, q)$ for GCMg.[3]

Measuring Model Complexity and Discriminability

Quantitative models are evaluated on the basis of their fit to data, with a superior fit generally interpreted as an indication that the model is a closer approximation to the underlying categorization process. The problem with such a conclusion is that a good fit can be achieved for other reasons, such as extra parameters, which in essence provide additional flexibility (more degrees of freedom) to improve fit. This is exactly the concern Smith and Minda (1998) raised about GCMg: γ made the model so flexible it could fit a wide range of data patterns, including those fit well by PRT. In the fields of statistics and computer science, this property of a model is termed *complexity* and refers to the inherent ability of a model to fit data.

Statistical model selection methods have been developed that penalize models for extra complexity, thereby placing them on an equal footing with models having less complexity. The most sophisticated of these is the minimum description length (MDL) method (Pitt, Myung, & Zhang, 2002; Rissanen, 1996, 2001). It is composed of two main parts.[4] The first is a goodness-of-fit measure, in this case the log maximum likelihood (LML), $\ln f(x|\theta^*)$, where θ^* represents the parameters that maximize the probability $f(x|\theta)$ that the model assigns to the observed data x. The second is a complexity measure, denoted by G, that takes into account the number of free parameters (k), the sample size (N), and the functional form of the model equation through the Fisher information matrix, $I(\theta)$. (See the Appendix for details on calculating these quantities as well as a brief explanation of why complexity is measured on an interval scale.)

$$\text{MDL} = -\text{LML} + \text{G} \tag{3}$$

$$\text{LML} = \ln f(x|\theta^*) \tag{4}$$

$$\text{G} = \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int_{\Theta}\sqrt{\det I(\theta)}\, d\theta. \tag{5}$$

Note that the measures of fit and complexity in MDL are additive and thus can be computed separately, providing a straightforward means of answering questions concerning γ. By comparing the complexities of the three models, we can learn how much the response-scaling parameter increases the flexibility of GCM and how much more complex GCMg may be when compared with PRT.

Model mimicry is a slightly different issue than model complexity. In this case, the concern is that if GCMg mimics PRT well, then the two models should be difficult to discriminate. We can evaluate whether this is the case by performing a model recovery test, in which two models first generate large samples of data.[5] These same two models are then fitted to every data set, generally using a measure such as LML, although more sophisticated methods like MDL can also be used. If the two models are discriminable, then each model should fit the data it generated better than the competing model does. If GCMg mimics PRT, then GCMg should not only fit its own data better than PRT would, but it should also fit data generated by PRT better than PRT itself does. Nosofsky and Zaki (2002; Zaki et al., 2003) showed that this was not always the case, demonstrating that there are conditions in which GCMg does not mimic PRT.

We performed a set of model recovery tests to determine how discriminable the two models are. In doing so, we varied the statistical method used to recover the models in order to show that the use of advanced model selection methods can greatly improve model discriminability. In these evaluations, LML was compared with MDL and another selection method, Akaike information criterion (AIC; Akaike, 1973), whose complexity takes into account only the number of parameters, (k),[6]

$$\mathrm{AIC} = -2 \cdot \mathrm{LML} + 2k. \tag{6}$$

Additionally, we varied the sample size to illustrate how discriminability improves as sample size increases. However, the effectiveness of any change in sample size or statistical method depends on the experimental design. If the design is poor, only marginal gains in discriminability might be obtainable via either method. In contrast, the proper design can be so decisive in favor of one model that there is little need for fancy statistical selection methods or large samples. The results from a comprehensive set of recovery tests can therefore speak to the quality of an experimental design as well.

Complexity and Discriminability of Categorization Models

The contribution of γ to model complexity and the ability of GCMg to mimic PRT were evaluated using two experimental designs that Nosofsky and Zaki (2002) argued differ in model discriminability, namely those used by Smith and Minda (1998; Experiment 2) and Nosofsky and Zaki (2002; Experiment 3). Both designs employ traditional category-learning experiments in which participants are trained to classify objects into one of two categories, with learning evaluated across or after training. The difference between the designs is that whereas Smith and Minda (1998) examined categorization performance using the same set of stimuli on which the participants were trained, Nosofsky and Zaki made a key change that was intended to further differentiate the models: They had participants classify additional, never-before-seen objects in a subsequent test phase, for which the models made classification predictions very different from those for the first training phase. If the Nosofsky and Zaki design is indeed more powerful, MDL should provide little additional information beyond what can be learned using common measures of fit (e.g., LML). Should this be the case, the design is doing most of the work, making fancy statistical machinery redundant. Changes in experimental design can alter model complexity dramatically, but the consequences of this for model selection are not easily predictable because, as will be seen, complexity does not vary in a uniform or constant way, but again, depends on design.

Analysis of the Smith and Minda (1998) design. In Smith and Minda's (1998) Experiment 2, participants were presented with 14 six-letter (i.e, $T = 6$) nonsense words, such as "gafuzi," and learned to categorize 7 of these words as belonging to one category and 7 to another. The experiment included two conditions, one that used linearly separable category stimuli and another that used nonlinearly separable stimuli; for the sake of simplicity, however, only stimuli from the latter category structure were adopted in the present analyses. Smith and Minda (1998) used a standard block-sampling technique, with category feedback provided. They analyzed the data from 10 blocks of 56 trials, each of which comprised 4 repetitions of 14 stimuli, examining each block separately for evidence favoring GCM or PRT. Accordingly, the sample size in Equation 5 would be $N = 56$, obtained by multiplying the number of category stimuli ($m = 14$) by the binomial sample size ($n = 4$).

We preface the discussion of the results by noting that model complexity is measured on an interval scale. For the present discussion, the most important implications of this are that there is no absolute zero point (i.e., complexity can take on negative values), and ratios of complexity differences are meaningfully interpretable, but ratios of complexities are not (Roberts, 1979; Stevens, 1946). In calculating complexity, we assumed $0 < q < 1$, $0 < \lambda < 20$, and $0 < \gamma < 10$, the same range of values used by Smith and Minda (1998). Note that both PRT and GCM have seven free parameters consisting of five attention weights (w_ts), one specificity parameter (λ), and one guessing parameter (q), whereas GCMg has eight free parameters, including the additional response-scaling parameter (γ). The model complexities for PRT, GCM, and GCMg are G = -0.986, 0.184, and 0.305, respectively, for sample size $N = 56$.

To interpret these values properly, they must be adjusted for the number of parameters, k, and the functional form of the model. For example, for $k = 1$ and $N = 56$, the first term in Equation 5 comes to $(1/2)\ln(56/2\pi) = 1.09$. Thus, we can think of each additional parameter as contributing an increase in complexity by this amount.[7] Viewed in this light, the difference in complexity between GCMg and GCM of 0.12 ($0.305 - 0.184 = 0.12$) indicates that adding the response-scaling parameter γ increased the complexity of GCM only slightly, 11% of the expected increase due to the difference in number of parameters between the two ($0.12/1.09 = 0.11$). Turning to the comparison between GCM and PRT, we note that there is a complexity difference of 1.17, despite the fact that both have the same number of parameters. Obviously, this is due to differences in functional form, which makes GCM more complex than PRT by almost one effective parameter ($1.17/1.09 = 1.07$). Similarly, the difference of 1.29 between GCMg and PRT [$0.305 - (-0.986) = 1.29$] implies that GCMg is more complex than PRT by about one effective parameter ($1.29/1.09 = 1.18$). GCMg is therefore more capable of fitting arbitrary data sets than PRT is, so caution is required when comparing fits.

When considering model mimicry, however, it is important to recognize that GCMg's extra complexity does not necessarily imply that GCMg can mimic PRT. To assess mimicry, we performed a set of model recovery tests. We sampled 3,000 data sets from each model and fit both PRT and GCMg to all data sets for three binomial sample sizes of $n = 4$, 20, and 100 (equivalently, $N = 56$, 280, 1,400) using LML, AIC, and MDL.[8] The results are shown in Table 1. When LML was used as the selection method, PRT provided the better fit to its own data 43%–57% of the time, with higher recovery rates observed for larger sample sizes, as should be the case. Data from GCMg were almost always fit better by GCMg, rarely by PRT. The use of more powerful selection methods shows that the poor recovery rate when the PRT model was used was due to the imbalance of complexity in favor of GCMg. Model discriminability improved greatly when AIC was used (83%–89%) and even more so when MDL was used (88%–99%), due to MDL's additional correction for functional form differences.

A more precise understanding of the discriminability of the two models, one that also makes the merits of the various selection methods easier to evaluate, is to compare the magnitudes by which one model (e.g., PRT) fitted a data set better than did the other model (e.g., GCMg). To perform this analysis, the difference in LML fits ($\text{LML}_{\text{PRT}} - \text{LML}_{\text{GCMg}}$) was calculated for the data generated by PRT.[9] The same was done for the data generated by GCMg, creating two distributions of fit difference scores. These scores, obtained using Smith and Minda's (1998) design, are plotted in Figure 1A, with the GCMg distribution specified by triangles, and the PRT distribution specified by crosses.

When PRT fits the data better than GCMg does ($\text{LML}_{\text{PRT}} > \text{LML}_{\text{GCMg}}$), the data fall to the right of the dotted line (located at 0 on the x-axis), which denotes the LML-based decision criterion. Negative values indicate a better fit by GCMg. Unfortunately, only about half of the PRT distribution (43%) falls to the right of the dotted line, implying that GCMg provides superior fits to PRT data more often than PRT itself does. For the GCMg data, GCMg, not surprisingly, almost always (97%) fits the data better than does the simpler PRT model. Looking at the two distributions together, there is an asymmetry in the extent to which they overlap that hints at the mimicry that concerned Smith and Minda (1998, 2002): GCMg extends across most of the PRT distribution, but the reverse is not true.

Despite these concerns, one can see that there is an optimal decision point for discriminating the models that lies where the distributions intersect (at the abscissa of approximately −2). By correcting for differences in complexity, the decision boundaries for AIC and MDL approach this ideal location. However, the fact that the distributions overlap means that no selection method performed perfectly, a result that is also reflected in the errors shown in Table 1. In short, GCMg and PRT are in fact

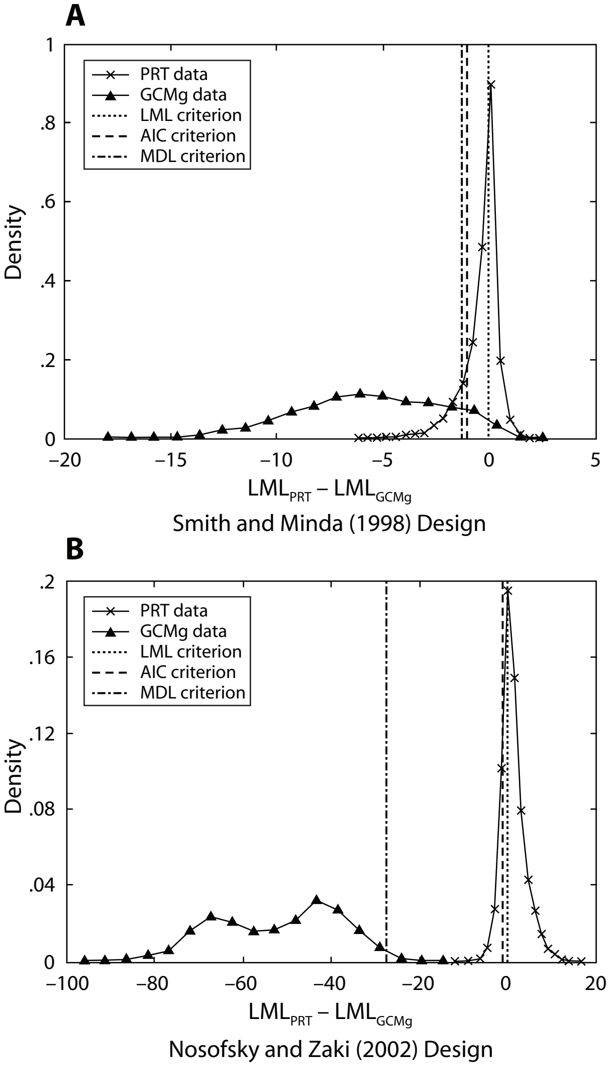

Figure 1. The inherent discriminability of PRT and GCMg. Panel A shows distributions of log maximum likelihood (LML) differences in fit the two models provide for simulated data generated from each model using the Smith and Minda (1998) design. Panel B shows similar distributions obtained using the Nosofsky and Zaki (2002) design. The binomial sample size $n = 4$ was used to obtain all four distributions, which were estimated by a kernel-smoothing method.

Table 1
Model Recovery Rates, in Percentages, of Two Categorization Models Under Three Selection Methods for Smith and Minda's (1998) Experimental Design

		Data Source					
		$n = 4$		$n = 20$		$n = 100$	
Selection Method	Model Fitted	PRT	GCMg	PRT	GCMg	PRT	GCMg
LML	PRT	43	3	45	1	57	1
	GCMg	57	97	56	99	43	99
AIC	PRT	83	10	86	3	89	1
	GCMg	17	90	14	97	11	99
MDL	PRT	88	12	96	4	99	2
	GCMg	12	88	4	96	1	98

Note—The value in each cell represents the percentage of samples in which the particular model was selected under the given selection method. Simulated data were generated by sampling across the entire parameter space according to the Jeffreys prior. This way, 3,000 parameter values were sampled for each model and binomial sample size (n).

reasonably discriminable in the Smith and Minda (1998, Experiment 2) design, but only if a selection method that controls for complexity is used. If such a method is not used, experimenters run a real risk of favoring the overly complex model, particularly when only small (i.e., realistic) samples are available.

Analysis of the Nosofsky and Zaki (2002) design. Like Smith and Minda (1998), Nosofsky and Zaki (2002) used 14 six-dimensional objects for training in Experiment 3 of their study, but the objects were cartoon bugs instead of nonsense words. The critical design change was that in the test phase, participants had to classify all 64 possible stimuli ($m = 64$), not just the 14 training items. The number of test blocks (i.e., binomial sample size) was again 4 ($n = 4$).

We estimated the complexities of the three models to be G = 17.26, 11.06, and 44.56, for PRT, GCM, and GCMg, respectively (sample size $N = nm = 256$). These values are much larger than those calculated for the Smith and Minda (1998) design, even though the models are exactly the same in both experiments. One cause of this increase in complexity value is due to the contribution of the sample size used in the design. Because there were 64 category stimuli in the present design, the sample size, N, in Equation 5 is much greater than the sample size in the Smith and Minda (1998) design ($N = 256$ vs. $N = 56$); thus, the value of the first term in Equation 5 increased. Another cause for the complexity increase comes from the second, functional form term of the complexity measure, which can also make a substantial contribution to complexity through the Fisher information matrix, $I(\theta)$. These differences clearly illustrate the profound effect experimental design can have on model complexity. It is tempting to think of model complexity as being constant, but as these data show, it is not static but rather depends on many factors, an important one of which is experimental design (see Pitt et al., 2002).

The overall effects of the contributions of functional form and number of parameters to model complexity greatly increases the complexity of GCMg relative to GCM and PRT. In fact, the introduction of the response-scaling parameter to GCM is now equivalent to adding about 31 effective parameters [$(44.56 - 11.06)/1.09 = 30.7$]. Similarly, GCMg is substantially more complex than PRT, specifically by about 25 effective parameters [$(57.44 - 17.06)/1.09 = 25.0$]. An experimental design that forces models to harness their computational power to fit data well exposes the extent to which γ can contribute to model performance. However, whether this increased complexity causes GCMg to mimic PRT to such an extent that the two are indiscriminable is another question entirely. The set of model recovery tests performed next was intended to answer this question.

The model recovery data are displayed in Table 2. The results using LML indicate that PRT provided the better fit to its own data 66%, 86%, and 93% of the time for binomial sample size $n = 4$, 20, and 100, respectively. GCMg perfectly recovered its own data across all sample sizes. With AIC, recovery improved noticeably, and with MDL, it was virtually perfect. Without a doubt, the models are highly discriminable in this experimental design. Comparison of the recovery results across Tables 1 and 2 shows the models to be more discriminable in the Nosofsky and Zaki (2002) design, as those authors argued.

Table 2
Model Recovery Rates, in Percentages, of Two Categorization Models Under Three Selection Methods for Nosofsky and Zaki's (2002) Experimental Design

		Data Source					
		$n = 4$		$n = 20$		$n = 100$	
Selection Method	Model Fitted	PRT	GCMg	PRT	GCMg	PRT	GCMg
LML	PRT	66	0	86	0	93	0
	GCMg	34	100	14	100	7	99
AIC	PRT	87	0	95	0	98	0
	GCMg	13	100	5	100	2	100
MDL	PRT	100	1	100	0	100	0
	GCMg	0	99	0	100	0	100

Note—The value in each cell represents the percentage of samples in which the particular model was selected under the given selection method. Simulated data were generated by sampling across the entire parameter space according to the Jeffreys prior. This way, 3,000 parameter values were sampled for each model and binomial sample size (n).

This improved discriminability is visible in Figure 1B, where the LML differences for the PRT and GCMg data are plotted. The distributions do not even come close to overlapping, clearly showing that the two models are entirely discriminable. Absolutely no mimicry is occurring. If this is the case, why were PRT data not fully recovered in some of the model recovery tests? The locations of the decision boundaries reveal why. For LML, the decision boundary is located well inside the PRT distribution. Selection errors are made when the fit difference is to the left of the boundary. Although selection improves with AIC, it is still prone to errors. The AIC criterion's close proximity to the LML criterion is due to the fact that the number of parameters contributed comparatively little to model complexity in this design. Only when functional form effects are neutralized, using MDL, is a more appropriate criterion found, although here, too, one might wonder whether the location of the criterion is too far to the left. The optimal location for it would appear to be in the region between distributions, not within the GCMg distribution. As the pattern of model recovery data for MDL in Tables 1 and 2 shows, the accuracy of MDL increases with sample size. In Figure 1, this translates into placement of the criterion in a location that improves discrimination. With a binomial sample size of $n = 4$, some imprecision is to be expected.

CONCLUSION

The model complexity calculations and the model recovery results across two experimental designs provide a clear and thorough understanding of how the response-scaling parameter γ affects GCM performance. Adding response scaling to the model does increase its complexity, not just by the addition of an extra degree of freedom for the model, but also by changing the way in which those degrees of freedom may be harnessed. In some contexts, such as the experimental design used by Nosofsky and Zaki (2002), the increase in model complexity is enormous. This increase in complexity allows GCMg to provide good fits to PRT data, as illustrated by the fact that the PRT distributions in Figure 1 have a sizable amount of their mass to the left of the 0 line. As a consequence, simple measures of fit such as LML perform poorly irrespective of the experimental design. However, the extent to which GCMg can mimic PRT is only partial: The GCMg and PRT distributions in Figure 1 are quite distinct, particularly in the Nosofsky and Zaki design. Accordingly, when we switch to statistical methods such as MDL that correct for the source of the problem—namely, model complexity—the apparent mimicry vanishes.

Additionally, the results of the two analyses show that the relationship between model complexity and model discriminability is not simple. Intuitively, an increase in complexity, which means an increase in data-fitting ability, should lead to greater mimicry. The current simulations show that this is not the case. If it were, the two models should be *less* discriminable in the Nosofsky and Zaki (2002) design than in the Smith and Minda (1998) design (since the complexity differences are larger in this design), when in fact the opposite is true.

How can complexity and discriminability increase together? The answer lies in considering the relationship of the models to one another and to the data. An increase in complexity is unlikely to generate an improvement in data-fitting precision across the range of all possible data patterns but rather is likely be localized to only a few patterns. In fact, a decrease in model discriminability will be found only in those cases in which the additional complexity increases the model's (e.g., GCMg's) ability to produce patterns that the competing model (PRT) fits well. This could be a very small region of the space of all possible data patterns. As for the increase in discriminability, GCMg and PRT are highly discriminable in the Nosofsky and Zaki (2002) design precisely because the data are fit exceptionally well only by GCMg. In other words, the data pattern is one that GCMg can generate but that PRT cannot.

An increase in model complexity does not guarantee an increase in model mimicry. In fact, additional complexity could be just what is needed to discriminate models but only if it further differentiates their predictions. If additional complexity does nothing more than improve fit to data that do not discriminate between models, then it is unjustified.

Finally, this study shows that statistical model selection methods can compensate for weaknesses in experimental design. To the extent that a highly informative experimental design can be found, the less need there is for model selection methods like MDL. In these situations, the differences in LML fit between models will be so large that even though substantial differences in complexity may exist between them, the contribution of complexity to the MDL value will be insignificant. When the design is weaker, reliance on MDL will be greater. In both cases, statistical model selection methods should always be used as supplementary tools in decision making and should never be the sole arbiter when evaluating competing models.

AUTHOR NOTE

All authors were supported by NIH Grant R01-MH57472. D.J.N. was also supported by Australian Research Council Grant DP-0773794 and a grant from the Office of Research at Ohio State University. We thank Nancy Briggs, Woojae Kim, Michael Lee, J. Paul Minda, Yves Rosseel, Yong Su, Eric-Jan Wagenmakers, and an anonymous reviewer for helpful comments and discussions. We also thank J. Paul Minda and Rob Nosofsky for kindly providing us with raw data sets for preliminary analyses that we conducted while preparing this article. Correspondence concerning this article should be sent to J. I. Myung, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210 (e-mail: myung.1@osu.edu).

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Casaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

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

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, **9**, 349-368.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1995). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision-bound models in large-size, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 128-148.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

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, & 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, & Cognition*, **28**, 275-292.

Myung, J. I., Forster, M. R., & Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, **44**, 1-2.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, **50**, 167-179.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, **51**, 85-98.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

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, & Cognition*, **28**, 924-940.

Olsson, H., Wennerholm, P., & Lyxzen, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 936-941.

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

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40-47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, **47**, 1712-1717.

Roberts, F. S. (1979). *Measurement theory*. New York: Addison-Wesley.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1411-1436.

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, & Cognition*, **28**, 800-811.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677-680.

Su, Y., Myung, J. I., & Pitt, M. A. (2005). Minimum description length and cognitive modeling. In P. Grunwald, J. I. Myung & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 411-433). Cambridge, MA: MIT Press.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, **48**, 28-50.

Wagenmakers, E. J., & Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, **50**, 99-100.

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, & Cognition*, **29**, 1160-1173.

NOTES

1. Navarro (2007) provided an in-depth discussion of the response-scaling parameter, including three theoretical interpretations one can attach to it: at the decision level, category similarity level, and representational structure level.

2. Only the multiplicative prototype model is evaluated in this article. We refer to it as PRT.

3. The results of this investigation do not change qualitatively if the parameter q is omitted. Conclusions drawn from this study, however, may not be generalizable to hybrid models, such as the mixed prototype model of Smith and Minda (1998).

4. There are several formulations of MDL. We used Rissanen's (1996) asymptotic approximation to the optimal normalized maximum likelihood method because it met our needs best. See Myung, Navarro, and Pitt (2006) for a simple introduction.

5. Sampling a data set x from a model is generally easy if the parameter values θ are known, since the sampling process reduces to sampling from $f(x|\theta)$ The difficult part is choosing a distribution $\pi(\theta)$ from which to sample the parameters. In this application, we used a Metropolis–Hastings algorithm (see, e.g. Gilks, Richardson, & Spiegelhalter, 1995) to sample θ from Jeffreys noninformative distribution, $\pi(\theta) \propto \sqrt{\det I(\theta)}$, which assigns equal prior probability to every distinguishable probability distribution (Balasubramanian, 1997).

6. Besides AIC and MDL, many other statistical selection methods, such as cross validation, Bayesian information criterion, and Bayes factor, to name a few, can be used for the same purpose. Olsson, Wennerholm, and Lyxzen (2004) recently used cross validation to compare prototype and exemplar models. For an in-depth discussion of various selection methods, see two special issues of the *Journal of Mathematical Psychology* on model selection (Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006).

7. This interpretation of the first term of model complexity in Equation 5 obviously requires that the sample size N be greater than $2\pi = 6.28$. Otherwise, adding each parameter would cause a decrease in complexity, which is nonsensical. Fortunately, this is not a problem in the present investigation. Because all three categorization models have at least seven parameters, the sample size N must be greater than seven in order to make the models identifiable.

8. Recall from Equation 5 that the first term of model complexity is a logarithmic function of sample size (N), whereas the second term is independent of N. As such, once the complexity of a model (e.g., $GN1$) is known for a particular sample size (e.g., $N1$), the model's complexity for any other sample size, say $N2$, is fully determined as $GN2 = GN1 + (k/2)\ln(N2/N1)$, where k is the number of model parameters.

9. This method of analyzing the whole distribution of LML differences obtained in a model recovery test is called the *landscaping technique* (Navarro, Pitt, & Myung, 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).

(*Continued on next page*)

APPENDIX

Likelihood Function

The likelihood function $f(x|\theta)$, given the data set $x = (x_{\mathrm{A1}}, x_{\mathrm{A2}}, \ldots, x_{\mathrm{Am}})$, in a two-category decision experiment, C = A or B, is given by

$$f(x\,|\,\theta) = \prod_{i=1}^{m} \frac{n!}{x_{\mathrm{A}i}!\left(n - x_{\mathrm{A}i}\right)!} P\left(\mathrm{A}\,|\,\mathrm{S}_i, \theta\right)^{x_{\mathrm{A}i}} \left[1 - P\left(\mathrm{A}\,|\,\mathrm{S}_i, \theta\right)\right]^{\left(n - x_{\mathrm{A}i}\right)}.$$

In this equation, m is the number of test stimuli, n is the binomial sample size or number of independent binary trials, $x_{\mathrm{Ai}} = \{0,1, \ldots, n\}$, $i = 1, \ldots, m$, is the observed number of category A decisions out of n trials for the ith stimulus S_i, and finally $P(\mathrm{A}|\mathrm{S}_i, \theta)$ denotes the categorization probability defined in Equation 2.

Fisher Information

The Fisher information matrix of sample size 1 ($N = 1$), $I(\theta)$, is the expected value of the second partial derivatives of the negative log likelihood of sample size 1 (see, e.g., Rissanen, 1996, Equation 7; Schervish, 1995, pp. 110-115),

$$I_{uv}(\theta) = -\frac{1}{mn} E\left[\frac{\partial^2 \ln f(x\,|\,\theta)}{\partial\theta_u \partial\theta_v}\right], (u, v = 1, \ldots, k),$$

where θ_{u} and θ_{v} correspond to the uth and vth elements of the model parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Since PRT, GCM, and GCMg are all multinomial, a standard result (Su, Myung, & Pitt, 2005) can be used to obtain the uvth element of the Fisher information matrix.

Calculation of the Complexity Measure

Calculating the complexity measure G in Equation 5 for the categorization models is reasonably simple in principle, if slightly tedious in practice. Calculating the second term of the complexity, involving the integrated Fisher information, is a simple task so long as we are able to find $I(\theta)$ for a given θ value. Once $I(\theta)$ can be calculated, all that is needed is the integration $\sqrt{\det I(\theta)}$ over the parameter range Θ. For the categorization models, these integrals are not very high dimensional, so simple Monte Carlo methods suffice. That is, we use the numerical approximation,

$$\int_{\Theta} \sqrt{\det I(\theta)}\, d\theta \approx \frac{1}{T}\left(\sum_{i=1}^{T} \sqrt{\det I(\theta^{(i)})}\right) \cdot V_{\Theta},$$

where V_{Θ} denotes the volume of the parameter space, and the $\theta^{(i)}$ values are T (=10,000, e.g.) independent samples from a uniform distribution over Θ.

The Measurement Scale of the Complexity Measure

The complexity measure G in Equation 5 is an interval scale of measurement. To understand why, consider that the complexity measure is obtained as an asymptotic approximation to the logarithm of the normalizing constant in the normalized maximum likelihood (NML) selection criterion (Rissanen, 2001):

$$\ln \int f\left(y\,|\,\theta^*\right) dy \approx \frac{k}{2} \ln\left(\frac{N}{2\pi}\right) + \ln \int_{\Theta} \sqrt{\det I(\theta)}\, d\theta,$$

for large N. The integration on the left-hand side of the equation is carried out over all possible data patterns that one could observe in an experiment. Thus the integral $\int f(y|\theta^*)dy$ represents the sum of all best fits a model can provide collectively, and the logarithm of this sum defines model complexity (see Myung, Navarro, & Pitt, 2006, for a review of NML).

Given that the maximum likelihood $f(y|\theta^*)$ is a ratio scale measurement, the integrated volume of maximum likelihoods, $\int f(y|\theta^*)dy$, is a ratio scale as well. The logarithm of ratio scale measurements satisfies the two key properties of the interval scale: rank order and equality of intervals (Roberts, 1979; Stevens, 1946). Consequently, the complexity measure G, which is essentially equal to the logarithm of the integrated volume, is also on an interval scale. Furthermore, we note that the LML is an interval scale measure as well and that it is directly additive with G, as shown in Equation 3.

(Manuscript received January 12, 2007;
revision accepted for publication April 18, 2007.)