hippocampus and neocortex–Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457

10 Petersen, S.E. *et al.* (1988) Positron emission tomographic studies of the cortical anatomy of single word processing. *Nature* 331, 585–589

11 Penfield, W. and Roberts, L. (1959) *Speech and Brain Mechanism*, Princeton University Press

12 Ojemann, G.A. (1979) Individual variability in cortical localization of language. *J. Neurosurg.* 50, 164–169

13 Nobre, A.C. *et al.* (1994) Word recognition in the human inferior temporal lobe. *Nature* 372, 260–263

14 Wada, J. and Rasmussen, T. (1960) Intracarotid injection of sodium amytal for the lateralization of speech dominance: experimental and clinical observations. *J. Neurosurg.* 17, 266–282

15 Pascual-Leone, A. *et al.* (1999) Transcranial magnetic stimulation: studying the brain-behaviour relationship by induction of 'virtual lesions'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1229–1238

16 Pascual-Leone, A. *et al.* (2000) Transcranial magnetic stimulation in cognitive neuroscience–virtual lesion, chronometry, and functional connectivity. *Curr. Opin. Neurobiol.* 10, 232–237

17 Marie, P. (1906a) Revision de la question de l'aphasie: La troisieme convolution frontale gauche ne joue aucun tole speciale dans la fonction du langage. *Semaine Medicale* 21, 241–247. (Repreinted in Cole, M.F. and Cole, M. eds., (1971), Pierre Marie's papers on speech disorders. New York: Hafner).

18 Marie, P. (1906b) Revision de la question de l'aphasie: Que faut-il penser des aphasies sous-corticales (aphasies pures)? *Semaine Medicale* 26, 493–500

19 Lashley, K.S. (1929) *Brain Mechanisms and Intelligence,* University of Chicago Press

20 Lashley, K.S. (1950) In search of the engram. In *Symposia for the Society for Experimental Biology, No, 4,* Cambridge University Press

21 Fodor, J.A. (1983) *The Modularity of Mind*, MIT Press

22 Vandenberghe, R. *et al.* (1996) Functional anatomy of a common semantic system for words and pictures. *Nature* 383, 254–256

23 Mummery, C.J. *et al.* (1999) Disrupted temporal lobe connections in semantic dementia. *Brain* 122, 61–73

24 Price, C.J. *et al.* (1999) Delineating necessary and sufficient neural systems with functional imaging studies of neuropsychological patients. *J. Cogn. Neurosci.* 11, 4371–4382

25 Price, C.J. and Friston, K.J. (1999) Scanning patients on tasks they can perform. *Hum. Brain Mapp.* 8, 102–108

26 Mummery, C.J. *et al.* (2000) A voxel based morphometry study of semantic dementia. The relation of temporal lobe atrophy to cognitive deficit. *Ann. Neurol.* 47, 36–45

27 Friston, K.J. *et al.* (1999) Multi-subject fMRI studies and conjunction analysis. *NeuroImage* 10, 385–396

28 Price, C.J. and Friston, K.J. (1997) Cognitive conjunctions: a new approach to brain activation experiments. *Neuroimage* 5, 261–270

29 Ashburner, J. and Friston, K.J. (2000) Voxel-based morphometry – the methods. *NeuroImage* 11, 805–821

30 Howard, D. and Patterson, K. (1992) *Pyramids and Palm Trees: A Test of Semantic Access from Pictures and Words*, Thames Valley, Bury St Edmunds

When a good fit can be bad

Mark A. Pitt and In Jae Myung

How should we select among computational models of cognition? Although it is commonplace to measure how well each model fits the data, this is insufficient. Good fits can be misleading because they can result from properties of the model that have nothing to do with it being a close approximation to the cognitive process of interest (e.g. overfitting). Selection methods are introduced that factor in these properties when measuring fit. Their success in outperforming standard goodness-of-fit measures stems from a focus on measuring the generalizability of a model's data-fitting abilities, which should be the goal of model selection.

Mark A. Pitt*
In Jae Myung
Dept of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210-1222, USA.
*e-mail: pitt.2@osu.edu

The explosion of interest in modeling cognitive processes over the past 20 years has fueled the cognitive sciences in many ways. Not only has it opened up new ways of thinking about research problems and possible solutions, but it has also enabled researchers to gain a better understanding of their theories by simulating a computational instantiation of it. Modeling is now sufficiently mainstream that one can get the impression that the models themselves are replacing the theories from which they evolved.

What has not kept pace with the advances and interest in modeling is the development of methods for evaluating and testing the models themselves. A model is not interchangeable with a theory, but only one of many possible quantitative representations of it. A thorough evaluation of a model requires methods that are sensitive to its quantitative form. Criteria used for evaluating theories [1], such as testing their performance in an experimental setting, do not speak to the quality of the choices that are made in building their quantitative counterparts (i.e. choice of parameters, how they are combined) or their ramifications. The paucity of such model selection methods is surprising given the centrality of the problem itself. What could be more fundamental than deciding between two alternative explanations of a cognitive process?

How *not* to compare models

Mathematical model are frequently tested against one another by evaluating how well each fits the data generated in an experiment or simulation. Such a test makes sense given that one criterion of model performance is that it reproduce the data. A goodness-of-fit measure (GOF; see Glossary) is invariably used to measure their adequacy in achieving this goal. What is measured is how much a model's predictions deviate from the observed data [2,3]. The model that provides the best fit (i.e. smallest deviation) is favored. The logic of this choice rests on the assumption that the model that provides the best fit to all data must be a closer approximation to the cognitive process under investigation than its competitors [4].

Such a conclusion is reasonable if measurements were made in a noise-free (i.e. errorless) system. One of the biggest challenges faced by cognitive scientists is that human and animal data are noisy. Error arises from several sources, such as the imprecision of our measurement tools, variation in participants and their performance over time. The problem of random

http://tics.trends.com 1364-6613/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S1364-6613(02)01964-2

Glossary

Complexity: the property of a model that enables it to fit diverse patterns of data; it is the flexibility of a model. Although the number of parameters in a model and its functional form can be useful for gauging its complexity, a more accurate and intuitive measure is the number of distinct probability distributions that the model can generate by varying its parameters over their entire range. Details of this 'geometric' complexity measure can be found in [a].

Functional form: the way in which the parameters (θ) and data (x) are combined in a model's equation: $y = \theta x$ and $y = \theta + x$ have the same number of parameters but different functional forms (multiplicative versus additive).

Generalizability: the ability of a model to fit all data samples generated by the same cognitive process, not just the currently observed sample (i.e. the model's *expected* GOF with respect to new data samples). Generalizability is estimated by combining a model's GOF with a measure of its complexity.

Goodness of fit (GOF): the precision with which a model fits a particular sample of observed data. The predictions of the model are compared with the observed data. The discrepancy between the two is measured in a number of ways, such as calculating the root mean squared error between them.

Minimum Description Length (MDL): a versatile measure of generalizability. MDL was developed within algorithmic coding theory in computer science [b], where the goal of model selection is to choose the model that permits the greatest compression of data in its description. Regularities in the data are assumed to imply redundancy. The more the data can be compressed by the model by extracting this redundancy, the more that is learned about the cognitive process.

Overfitting: the case where, in addition to fitting the main trends in the data, a model also fits the microvariation from this main trend at each data point. Compare the middle and right graph inserts in Fig. 1.

Parameters: variables in a model's equation that represent mental constructs or processes; they are adjusted to improve a model's fit to data. For example, in the model $y = \theta x$, θ is a parameter.

Probability density function: a function that specifies the probability of observing each outcome of a random variable given the value of a model's parameter.

References

a Myung, I.J. *et al.* (2000) Counting probability distributions: differential geometry and model selection. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11170–11175

b Li, M. and Vitanyi, P. (1997) *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag

error and the lengths researchers go to combat it are evident in the experimental and statistical methods used in the field (e.g. repeated measurement, inferential statistics). They serve to hold error in check so that variation due to the mental process of interest, what we are really interested in, will be visible in the data.

Noisy data make GOF by itself a poor method of model selection. As the simulation in Box 1 illustrates, a GOF measure such as the Root Mean Squared Error (RMSE) is insensitive to the different sources of variation in the data, whether it is random error or due to the cognitive process of interest. This could result in the selection of a model that overfits the data, which may not be the model that best approximates the cognitive process under study.

How to compare models

Because it is impossible to eliminate error from data, efforts have focused on improving model selection in other ways. The preferred solution has been to redefine the problem as one of assessing how well a model's fit to one data sample generalizes to future samples generated by that same process [5]. Generalizability (see Glossary) uses the data and information about the model itself to make a 'best-guess' estimate as to how likely it is the model could have generated the data sample in hand. In this approach, a good fit is a necessary but not sufficient condition for a model because many models are capable of fitting a dataset reasonably well. Because the set of such candidate models is potentially quite large, we can only ever infer the likelihood with which each model under consideration generated the data. In this regard, generalizability is a statistical inference problem. It is no different conceptually from estimating the replicability of an experimental result using inferential statistics or inferring the characteristics of a population from a sample.

A handful of generalizability measures have been proposed in the lasts 30 years [6]. By necessity, all include a measure of GOF to assess a model's fit to the data (see Box 2). Terms encoding information about the model itself are included to level the playing field among models so that one model, by virtue of its design choices (i.e. mathematical instantiation) does not have an inherent advantage over its competitors in fitting the data best, and thus being selected. By nullifying such model-specific properties, the model that is the best approximation to the cognitive process under study, and not simply the one that absorbs the most variation in the data, will be selected.

Measures of generalizability

Early measures of generalizability such as the Akaike information criterion (AIC) [7,8] and Bayesian information criterion (BIC) [9] addressed the most salient differences among models: the number of free parameters. As is generally well known, a model with many free parameters can provide a better fit to a data sample than a model with few parameters, even if the latter generated the data. The second term in these measures includes a count of the number of parameters (k). AIC and BIC penalize a model more as the number of parameters increases. To be selected, the model with more parameters must overcome this penalty and provide a substantially better fit than a model with fewer parameters. That is, the superior fit obtained with the extra parameters must justify the necessity of those parameters in fully capturing the cognitive process.

An equally salient but much less tangible dimension along which models also differ is in their functional form, which refers to the way in which the parameters are combined in a model's equation. More sophisticated selection methods, such as Bayesian model selection (BMS) [10] and minimum description length (MDL) [11,12], are sensitive to a model's functional form as well as to the number of parameters. Functional form is taken into account in the third term in the MDL measure. In BMS, both are hidden in the integral. (Cross Validation, although not listed, is another selection method that is thought to be sensitive to both dimensions of model complexity. It involves applying GOF in a non-standard way. [13])

The second and third terms in MDL together provide a measure of a model's complexity. Conceptually, complexity refers to that characteristic

Box 1. Why GOF alone is a poor measure of model selection

The ability of a model to fit data is a necessary condition that all models must satisfy to be taken seriously. GOF measures such as RMSE are inappropriate as model selection methods because all they do is assess fit. This myopic focus is problematic when variation in the data can be due to other factors (e.g. random sampling, individual differences) as well as the cognitive process under study.

Table I. Results of a model recovery simulation in which a GOF measure (RMSE) was used to discriminate models when the source of the error was varied.

Condition (sources of variation)	Model the data were generated from			Model fitted	
	M_A a = 0.4	M_A a = 0.6	M_B	M_A	M_B
(1) Sampling error	100	–	–	0.040 (0%)	0.029 (100%)
(2) Sampling error + individual differences	50	50	–	0.041 (0%)	0.029 (100%)
(3) Different models	–	50	50	0.075 (0%)	0.029 (100%)
(4) Sampling error	–	–	100	0.079 (0%)	0.029 (100%)

The severity of the problem is shown in Table I, which contains the results of a model recovery simulation using RMSE. Four datasets were generated from a combination of the two models (M_A and M_B), defined as follows: M_A: $y = (1+t)^{-a}$, M_B: $y = (b+ct)^{-a}$ where a, b, c > 0. Datasets were generated from each in the frequencies shown in the four conditions (rows). In the first condition, all 100 samples were generated by M_A with $a = 0.4$ and only sampling error introduced as noise. In the second condition, variation due to individual differences was also added to the data by using a different parameter value ($a = 0.6$) half of the time. In the third condition, half of the data were generated by model M_A and half by M_B. Condition four is the reverse of condition one, with the data plus sampling error being generated by M_B. Models M_A and M_B were then fitted to the data in each condition using RMSE. The mean RMSE fits, along with the percentage of time each model provided the best fit, are shown on the two right-most columns of the Table.

A good model selection method must be able to ignore irrelevant variation in the data (e.g. sampling error, individual differences that are not being modeled) and recover the model that generated the data. That is, the selection method must be capable of differentiating between the variation that the model was designed to capture and the variation due to noise. RMSE fails miserably at this task. M_B was chosen 100% of the time in all four conditions and the mean fit is substantially better than that of M_A. M_A never provided a better fit than M_B, even when some or all of the data were generated by M_A (conditions 1–3). This is why a good fit can be bad.

Readers who have some familiarity with modeling might not be surprised by the results given that M_B has two more parameters than M_A. A model with more parameters will always provide a better fit, all other things being equal [a]. The typical solution to this problem is to control for the number of parameters, but there are at least two reasons why this fix is unsatisfactory. The most obvious is that it limits model comparison to those situations in which the number of parameters is equated across models. The diversity of models in the cognitive sciences can make this a significant impediment in doing research. Less obvious although even more important is that a model's data-fitting abilities are also affected by other properties of the model, such as its functional form [b,c,d]. Unless they are taken into account by the selection method, simply equating for number of parameters will not place the models on an equal footing.

In summary, it may make perfect sense to use GOF to determine whether a given model can even pass the test of fitting a dataset reasonably well (i.e. capturing the main trends), but going beyond this and comparing such fits *between* models, although intuitive, is risky.

References

a Linhart, H. and Zucchini, W. (1986) *Model Selection*, Wiley

b Myung, I.J. *et al.* (2000) Special Issue on model selection. *J. Math. Psychol.* 44 (1–2)

c Li, S.C. *et al.* (1996) Using parameter sensitivity and interdependence to predict model scope and falsifiability. *J. Exp. Psychol. Gen.* 125, 360–369

d Myung, I.J. and Pitt, M.A. (1997) Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon. Bull. Rev.* 4, 79–95

Box 2. Model selection criteria as measures of generalizability

Listed in Table II are two GOF measures (RMSE, PVAF) and four generalizability measures (AIC, BIC, BMS, MDL). Except for BMS, in which the likelihood function is integrated over the parameter space, the measures of generalizability use the maximized log-likelihood, that is, $ln(f(y|\theta_0))$, as a GOF measure, the minus of which represents lack of fit. This fit index is combined with a measure of model complexity to yield a generalizability measure. The four generalizability criteria differ from one another in the conceptualization of model complexity. In AIC, the number of parameters (k) is the only dimension of complexity that is considered, whereas BIC also considers sample size (n). BMS and MDL go one step further and also take into account the functional form of a model's equation. In MDL, this is reflected through the third term of the criterion equation whereas in BMS it is hidden in the integral. These selection methods, except for PVAF, prescribe that the model that minimizes the given criterion should be chosen.

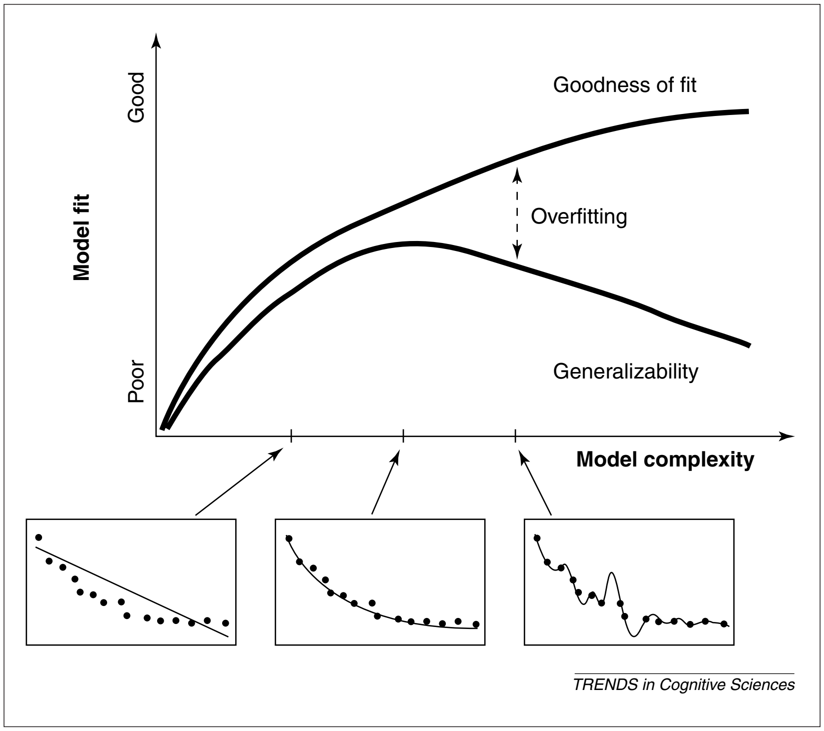
Reference

a Schervish, M.J. (1995) *The Theory of Statistics*, pp. 11–115, Springer-Verlag

Table II. Two GOF Measures, four generalizability measures, and the dimensions of complexity to which each is sensitive

Selection method	Criterion equation	Dimensions of complexity considered
Root Mean Squared Error	$RMSE = (SSE/N)^{1/2}$	None
Percent Variance Accounted For	$PVAF=100(1-SSE/SST)$	None
Akaike Information Criterion	$AIC = -2\ ln(f(y\|\theta_0)) + 2k$	Number of parameters
Bayesian Information Criterion	$BIC = -2\ ln(f(y\|\theta_0)) + k \cdot ln(n)$	Number of parameters, sample size
Bayesian Model Selection	$BMS=-ln\int f(y\|\theta)\pi(\theta)d\theta$	Number of parameters, sample size, functional form
Minimum Description Length	$MDL=-ln\ (f(y\|\theta_0)) + (k/2)ln(n/2\pi)+ln\int\sqrt{\det(I(\theta))}d\theta$	Number of parameters, sample size, functional form

In the equations above, *y* denotes observed data, θ is the model's parameter, θ_0 is the parameter value that maximizes the likelihood function $f(y|\theta)$, *k* is the number of parameters, n is the sample size, N is the number of data points fitted, SSE is the minimized sum of the squared errors between observations and predictions, SST is the sum of the squares total, $\pi(\theta)$ is the parameter prior density, $I(\theta)$ is the Fisher information matrix in mathematical statistics [a], *det* denotes the determinant of a matrix, and *ln* denotes the natural logarithm of base e.



TRENDS in Cognitive Sciences

Fig. 1. Relationship between goodness of fit and generalizability as a function of model complexity. The *y*-axis represents any fit index, where a larger value indicates a better fit (e.g. percent variance accounted for). The three smaller graphs along the *x*-axis show how fit improves as complexity increases. In the left graph, the model (represneted by the line) is not complex enough to match the complexity of the data (dots). The two are well matched in complexity in the middle graph, which is why this occurs at the peak of the generalizability function. In the right graph, the model is more complex than the data, fitting random error. It has better goodness of fit, but is overfitting the data.

Table 1. Model recovery performance (percentage of correct recoveries) for three models using three selection methods

Selection method	Model fitted	Model the data were generated from		
		M_1	M_2	M_3
PVAF	M_1	0	0	0
	M_2	38	97	30
	M_3	62	3	70
AIC	M_1	79	0	0
	M_2	9	97	30
	M_3	12	3	70
MDL	M_1	86	0	0
	M_2	1	92	8
	M_3	13	8	92

Models M_1, M_2 and M_3 were defined as follows: M_1: $y = (1+t)^{-a}$; M_2: $y = (b+t)^{-a}$; M_3: $y = (1+bt)^{-a}$. In the model equations, a, b and c are parameters that were adjusted to fit each model to the data, which were generated using the same five points, t = 0.1, 2.1, 4.1 6.1, 8.1. Each sample of five observations was sampled from the binomially probability distribution of size n = 50. One thousand samples were generated from each model and served as the data to fit. Each selection method was then used to determine which model generated each of the samples. The percentage of time each model was chosen for each dataset is shown.

of a model that makes it flexible and easily able to fit diverse patterns of data, usually by a small adjustment in one of its parameters. In Fig. 1, it is what enables the model (wavy line) in the lower right graph to provide a better fit to the data (dots) than that in the middle and left graphs. Both FUNCTIONAL FORM and the number of PARAMETERS contribute to model complexity.

How complexity is related to generalizability and GOF: the dilemma of Occam's razor

The notion of model complexity is a good vehicle with which to illustrate further the goal of generalizability and distinguish it from GOF. As demonstrated in Box 1, fit is maximized with GOF. Because additional complexity will improve fit, the two are positively related; this is depicted by the top function in Fig. 1. Generalizability, on the other hand, is not related to complexity so straightforwardly. Rather, its function follows the same trajectory as that of GOF up to a certain point, after which fit declines as complexity increases.

Why do the two functions diverge? The data being fitted have a certain degree of complexity that reflects the operation of the cognitive process. This point corresponds to the peak of the generalizability function. Any additional complexity beyond that needed to capture the underlying process (to the right of the peak) will cause the model to overfit the data by also capturing the microvariation due to random error, and thus reduce generalizability. The reason both functions overlap to the left half of the peak is that the model itself must be sufficiently complex to fit the data well. The model will underfit the data if it lacks the necessary complexity (i.e. it is too simple), as illustrated in the lower left graph in Fig. 1. The dilemma that is faced in trying to maximize generalizability should be clear: a delicate balance must be struck between sufficient complexity on the one hand and good generalizability on the other. MDL achieves this balance by choosing the model whose complexity is most justified by considering the complexity of the data relative to the complexity of the model.

Selection methods at work: the proof is in the pudding

Generalizability measures like BMS and MDL have thus far been developed for testing only those models that can be described in terms of a PROBABILITY DENSITY FUNCTION. Examples of such statistical models are models of categorization and models of information integration [14,15].

The following model-recovery tests demonstrate the relative performance of the three classes of selection methods shown in Table 1. The three models described (see Table footnote) were compared. In each simulation, 1000 datasets were generated from each model. Each selection method was then tested on its ability to determine which of the three models did in fact generate the 3000 datasets. A good selection method should be able to discern correctly the model that generated the data. The ideal outcome is one in which each model generalizes best only to data generated by itself. In the 3×3 matrices in Table 1, this corresponds to perfect recovery in the diagonal going from the upper left to the lower right. Errors (cells in the off diagonals) reveal a bias in the selection method toward either the more or less complex model.

The top matrix shows the results using PVAF, with the percentage of correct recoveries in each cell. The problem with using a GOF measure shows up in the first column of the matrix, where the data were generated by the one-parameter model M_1. M_1 never recovered its own data, with one of the two-parameter models always fitting the data best. Comparison of these data with the middle matrix shows that using AIC rectifies this problem. Because of its sensitivity to the number of parameters in a model, it does a reasonably good job of distinguishing between data generated by M_1 or M_2. By contrast, note how model recovery performance remains constant across these two matrices in columns 2 and 3. This is not surprising because AIC, like PVAF, is insensitive to functional form, which is the dimension along which M_2 and M_3 differ. Only when MDL is used is an improvement in model recovery performance observed in these columns (bottom matrix).

Why did MDL not perform perfectly, and why did it perform slightly worse than AIC and PVAF when the data came from M_2 (middle column)? Recall that, like a statistical test, model selection is an inference problem. The quality of the inference depends on the information that the selection method uses. Even though MDL makes use of all of the information available (data and the model), this does not guarantee success, but it greatly improve the chances of the inference being correct [16]. MDL will outperform selection methods such as AIC and PVAF most of the time, but neither it nor its Bayesian counterpart, BMS, are infallible.

The inferential nature of model selection makes it imperative to interpret model recovery results in the context of the data that were used in the test itself. Poor performance might not indicate a problem with the selection method, but rather a constraint on the resolving power of the selection method in discriminating which model could have generated the data. Conversely, biases in the selection method itself can masquerade as good model-recovery performance. In a recent comparison of BMS and RMSD, we demonstrated both of these outcomes [17].

Conclusion

Methods like AIC and MDL can give the impression that model selection can be automated and require minimal thought. However, choosing between competing models is no easier or less subjective than choosing between competing theories. Selection methods should be viewed as a tool that researchers can use to gain additional information about the models under consideration. These tools, however, are ignorant of other properties of the models, such as their plausibility or the quality of the data, so it is inappropriate to decide between them solely on what is learned from a test of generalizability.

Acknowledgement
Both authors were supported by NIH Grant MH57472.

References

1 Jacobs, A.M. and Grainger, J. (1994) Models of visual word recognition–sampling the state of the art. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 1311–1334
2 Smith, J.D. and Minda, J.P. (2000) Thirty categorization results in search of a model. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 3–27
3 Wichman, F.A. and Hill, N.J. (2001) The psychometric function: I. Fitting, sampling and goodness of fit. *Percept. Psychophys.* 63, 1293–1313
4 Roberts, S. and Pashler, H. (2000) How persuasive is a good fit? A comment on theory testing. *Psychol. Rev.* 107, 358–367
5 Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, (Chapters 1 and 9), Oxford University Press
6 Myung, I.J. *et al.* (2000) Special Issue on model selection. *J. Math. Psychol.* 44, 1–2
7 Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Petrox, B.N. and Caski, F.), pp. 267–281, Akademiai Kiado, Budapest
8 Burham, K.P. and Anderson, D.R. (1998) *Model Selection and Inference. A Practical Information-Theoretic Approach* (Chapter 2), Springer-Verlag
9 Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Stat.* 6, 461–464
10 Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795
11 Rissanen, J. (1996) Fisher information and stochastic complexity. *IEEE Trans. Inform. Theor.* 42, 40–47
12 Hansen, M.H. and Yu, B. (2001) Model selection and the principle of minimum description length. *J. Am. Stat. Assoc.* 96, 746–774
13 Myung, I.J. and Pitt, M.A. (in press). Model evaluation, testing and selection. In *Handbook of Cognition* (Lambert, K. and Goldstone, R. eds.), Sage
14 Nosofsky, R.M. (1986) Attention, similarity and the identification-categorization relationship. *J. Exp. Psychol. Gen.* 115, 39–57
15 Oden, G.C. and Massaro, D.W. (1978) Integration of featural information in speech perception. *Psychol. Rev.* 85, 172–191
16 Pitt, M.A. *et al.* (2002) Toward a method of selecting among computational models of cognition. *Psychol. Rev.* 109, 472–491
17 Pitt, M.A. *et al.* Flexibility versus generalizability in model selection. *Psychon. Bull. Rev.* (in press)