

Model Comparisons and Model Selections Based on Generalization Criterion Methodology

Jerome R. Busemeyer

Indiana University

and

Yi-Min Wang

Purdue University

The purpose of this article is to formalize the generalization criterion method for model comparison. The method has the potential to provide powerful comparisons of complex and nonnested models that may also differ in terms of numbers of parameters. The generalization criterion differs from the better known cross-validation criterion in the following critical procedure. Although both employ a calibration stage to estimate parameters, cross-validation employs a replication sample from the *same design* for the validation stage, whereas generalization employs a *new design* for the critical stage. Two examples of the generalization criterion method are presented that demonstrate its usefulness for selecting a model based on sound scientific principles out of a set that also contains models lacking sound scientific principles that are either overly complex or oversimplified. The main advantage of the generalization criterion is its reliance on extrapolations to *new* conditions. After all, accurate *a priori* predictions to new conditions are the hallmark of a good scientific theory. © 2000 Academic Press

Rapid advancements in computing technology have facilitated the use of increasingly complex models of human behavior. Examples include production rule models of problem solving, neural network models of language development, or diffusion models of signal detection. Corresponding to this rise in model complexity, there is an increasing need for rigorous methods that can be used to compare the scientific value of complex models (see, e.g., Jacobs & Grainger, 1994). The problem of comparing and selecting models for complex systems is difficult because

This work was supported by Grants from NIMH (R01 MH55680) and NSF (SBR 9602102). The authors thank Michael Browne, Malcom Forster, Richard Golden, and In Jae Myung for their very helpful comments on this manuscript. Correspondence and reprint requests should be addressed to Jerome R. Busemeyer, Psychology Department, Indiana University, Bloomington, Indiana 47405-1301. E-mail: jbusemey@indiana.edu.



the models being compared may differ in terms of function class and parameter dimensionality. One model may fit better than another simply because it has a more flexible function form or more parameters than the other and not because it is based on better scientific principles (see Collier, 1985; Cutting, Bruno, Brady, & Moore, 1992).

Many researchers believe that the best method for comparing models is the *strong inference test* (Platt, 1964). Essentially, this method requires deriving *a priori* and *parameter-free* predictions concerning the rank ordering of performance across a set of experimental conditions, where each class of models predicts a unique rank order. Model selection is based on eliminating the classes of models that predict rank orders contrary to the observed results. In other words, this tests the qualitative properties derived from models, and no parameter estimation or quantitative model fitting is necessary.

Strong inference tests have been developed for a limited number of domains in psychology. For example, measurement theorists have developed ordinal tests for choosing among polynomial measurement classes (see Krantz & Tversky, 1971), and response time theorists have developed ordinal tests for choosing among information processing architectures (see Schweickert, Fisher, & Goldstein, 1998; Townsend, 1990).

Often the strong inference test cannot be applied to complex models because it is too difficult or impossible to derive parameter-free predictions. Even when it is possible, it is still informative to quantitatively evaluate the magnitude of the prediction errors. Scientists are rarely satisfied with a model that predicts the correct order, but makes dramatically incorrect predictions regarding magnitude.

The purpose of this article is to present a quantitative method for comparing complex models called the *generalization criterion*. Like strong inference, the generalization criterion is based on *a priori* predictions (made before observing the data) rather than *post hoc* fits (made after observing the data). However, unlike strong inference, the generalization criterion uses parameter dependent quantitative predictions from complex models. The generalization criterion has been employed informally for many years, but this article serves the purpose of formally identifying the procedures so that the method becomes generally recognized and applied.

The remainder of this article is organized as follows. First we review the scientific goals upon which model selection is based, second, we examine several existing model comparison methods with respect to these goals, then we present the generalization criterion procedure and provide some example demonstrations of this method, and we finish with a discussion of ways to broaden the applicability of the method.

I. SCIENTIFIC GOALS FOR MODEL SELECTION

There are general agreements among scientists about some of the goals of science. The primary goal is to build general theories that are founded on *sound scientific principles* which explain existing facts in a rigorous manner, and more importantly, they are useful for deriving accurate *a priori* predictions for new findings under

novel conditions. The derivation of predictions normally requires the construction of a detailed model from the general theory for the specific situation under investigation (see Anderson, 1993). While the detailed model is primarily based on key theoretical principles, the general theory fails to completely specify all of the details for every situation. Minor ad hoc assumptions must be attached to form a computational model. For example, a theory may provide the basic causal structure or information processing architecture, but specific measurement assumptions (e.g., linearity) or distribution assumptions (e.g., normality) may need to be added to make a computational model.

Therefore the model is partially based on auxiliary details or minor ad hoc assumptions that fall outside the general theory and could be wrong. So when two or more models are compared, scientists are mainly interested in determining which general theory is better, and he or she is less interested in the possibly incorrect auxiliary assumptions used to construct each model. This view of model construction implies that all models are wrong in detail. Scientists are more interested in identifying the model based on *sound scientific principles* that permit one to make new *a priori* predictions under novel conditions.

Scientists often distinguish between models based on sound scientific principles from two other extreme types of models—oversimplified and overly complex models. By definition, both of the latter types *lack* sound scientific principles, but they have advantageous mathematical or statistical properties. To illustrate these ideas in a concrete manner, consider the problem of selecting a model for the forgetting curve in human memory (see, e.g., Rubin & Wenzel, 1996; Wickens, 1998; Wixted & Ebbesen, 1991).

Oversimplified models are based on simplifying assumptions and a small number of parameters. For example, one could linearly regress percent recall on time delay to model forgetting, but this simple linear regression model violates known principles from human memory research. Nevertheless, if the sample size is small and the data are very noisy, then this linear model may fit about as well as other models based on sound scientific principles. The robustness of linear models makes them relatively effective for fitting very noisy data under small sample sizes (see Dawes & Corrigan, 1974).

Overly complex models are based on an excessively large number of parameters. For example, one could fit percent recall as a function of time delay by a high order polynomial regression model. Like the linear model, the polynomial model violates known principles from human memory research. Nevertheless, if the sample size is large and the data are not very noisy, then this model could fit better than models based on sound scientific principles. The flexibility of such models allows them to provide good *post hoc* fits without using any knowledge about the principles governing the causal relations.

For basic science, the first requirement of a model comparison method is that it selects a model based on sound scientific principles from a set that also includes oversimplified and overly complex types that lack such principles. The problem with many existing methods is that they tend to miss the model based on sound scientific principles, and instead they tend to pick out either an oversimplified model or an overly complex model that lacks these principles.

There is a second requirement that needs to be discussed, but first a few definitions are required to clarify the presentation. Consider an experiment that was designed to compare two complex models, models a versus b , where the models may be nonnested and may even differ in terms of numbers of parameters. Model a is nested within model b if model a is a special case or restricted version of model b . In other words, for a given design, any set of predictions produced by model a can also be produced by model b .

Define X as a fixed design matrix where each of the K rows is used to code the experimental design. For example, the first two columns could represent intensity and duration measurements for each stimulus, and each row could represent one of the K different stimuli presented to a subject. Define N as the sample size or number of independent replications for each row of the design. Define $S(X)$ as a K -dimensional vector of sample statistics (sample proportions, means, covariances, etc.), and

$$\mu(X) = p \lim_{N \rightarrow \infty} S(X)$$

represents the corresponding population values (population proportions, means, covariances, etc.).

The predictions of each model depend on a collection of parameters (such as the intercept, slope, and exponent of a power function) which are represented by a parameter vector θ_m for model m . The predictions are considered *a priori* when the parameters are estimated from past experiments and used to make predictions for a new experimental design. The predictions are considered *post hoc* when the parameters are estimated from the current experiment and are used to make predictions for exactly the same experimental design.

In general, define $Q_m[X, \theta_m]$ as a K -dimensional vector of predictions generated from model m ($m = a$ or $m = b$) for design X using parameters θ_m . These predictions may be either *a priori* or *post hoc*, depending on the nature of the research. Define $D\{Q, V\}$ as a nonnegative measure of the lack of fit, that is the discrepancy between some prediction vector Q and some target vector V (e.g., the sum of squared deviations between the predictions and the data or the negative log likelihood of the data given the predictions).

Figure 1 depicts the relationships between these concepts (see Linhart & Zucchini, 1986, for a related discussion). The branch labeled α represents the *population discrepancy*

$$\alpha_m = D\{Q_m[X, \theta_m], \mu(X)\}$$

between the actual population values and the model predictions using the parameters θ_m ; the branch δ represents the *sample discrepancy*

$$\delta_m = D\{Q_m[X, \theta_m], S(X)\}$$

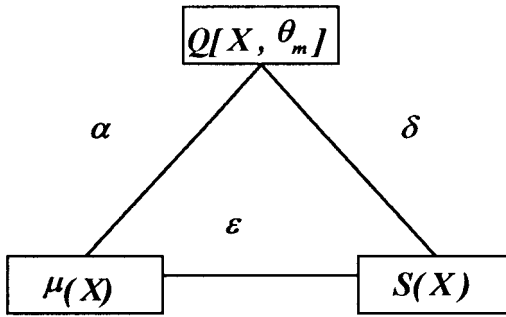


FIG. 1. Array of discrepancies.

between the actual sample statistics and the same model predictions using the same parameters θ_m ; finally, the branch labeled ε represents the *sampling error* or difference between the population values and the sample statistics.

Obviously, researchers would like to know the population discrepancy, α_m , but because of practical concerns, they are limited to observing the sample discrepancy, δ_m , which is contaminated by unwanted sampling error ε . To achieve rigorous tests of models, scientists minimize sampling error by choosing a large sample size that provides precise estimates of the sample statistics. If a model comparison method fails to achieve the goal of selecting models with sound scientific principles under the ideal conditions at the population discrepancy level, there is little reason to believe that it will be effective for this purpose under noisy conditions at the sample discrepancy level. Therefore, the second requirement of a model comparison method for scientific purposes is that it should be effective at the population discrepancy level (i.e., as $N \rightarrow \infty$).

In summary, the goal of a model selection method is to help pick out a model based on sound scientific principles from a set that includes oversimplified and overly complex models, and furthermore, this method should be most effective at very large sample sizes that closely approximate the population discrepancy level of analysis. However, if the parameters are estimated in a post hoc manner, this goal is hard to achieve in practice—typically the most complex model is selected. Fortunately, the odds of achieving these goals are greatly increased when the models are compared on the basis of a priori predictions.

II. REVIEW OF COMMONLY USED METHODS

Chi-square test. The most commonly used method of model comparison is (a) to fit each model to the entire data set using a maximum likelihood criteria, and then (b) to compare the model fits using a chi-square test based on the log likelihood ratio statistic (Wilks, 1938; see Mood & Graybill, 1963),

$$G^2(a, b) = -2 \cdot \ln[L_a(X, \theta_a)/L_b(X, \theta_b)], \quad (1a)$$

where $L_a(X, \theta_a)$ is the maximum likelihood for model a and $L_b(X, \theta_b)$ is the maximum likelihood for model b . It is convenient to rewrite the chi-square criterion

in terms of a comparison of the average log likelihood for each model (sample discrepancies),

$$G^2(a, b)/2N = (\delta_a - \delta_b), \quad (1b)$$

where

$$-\delta_m = \ln[L_m(X, \theta_m)]/N$$

is the average over the N individual log likelihood terms produced by each independent observation $i = 1, \dots, N$ for model m .

The chi-square criterion measures the increase in discrepancy produced by changing from a general (model b) to a more specific model (model a). This method is limited to comparisons among nested models (model a is nested within model b), and the nested relation between the two models implies that model b must produce a smaller discrepancy than model a . If the chi-square criterion exceeds a cutoff, then the null hypothesis (H_0) is rejected, which implies no significant model differences. Usually this means testing the null hypothesis (H_0) that one or more parameters in the more complex model (e.g., model b) are equal to some fixed values (e.g., zero) when the model is fit to the population values. This limitation to nested model comparisons is a serious drawback, because scientists are generally more interested in comparisons among qualitatively different, nonnested models. To use this test, one must also assume that the true data generating process is nested within the most complex model.

It is well known that this method tends to pick the oversimplified model (fails to reject H_0) with small sample sizes that suffer from a lack of statistical power, and it tends to pick the overly complex model (rejects H_0) in large sample sizes that enjoy extremely high statistical power (see Cudeck & Browne, 1983). The latter tendency is due to the fact that the additional flexibility provided by the more complex model produces a better fit, $\delta_b - \delta_a < 0$, and this improvement is magnified by the sample size N [see Eq. (1b)].

AIC. The Akaike information criterion (AIC; Akaike, 1973; also see Rust, Simester, Brodie, & Nilikant, 1995, for a review of this and other related criteria) is rapidly becoming more popular among scientists, primarily because it permits comparisons between nonnested models that may also differ in terms of numbers of free parameters. (A closely related method is the Bayesian information criterion, also known as the BIC; Schwarz, 1978.) The purpose of the AIC is to select a model that produces the smallest *expected* discrepancy, where the expectation is taken across the population of replications generated by a fixed design (see Bozdogen, 2000). A complex model may give the smallest discrepancy for the particular replication of the design to which it was fit, but it may give a larger expected discrepancy, averaged across many replications of the design.

The AIC uses the G^2 discrepancy measure in Eq. (1a), but it adds a penalty factor that is proportional to the difference in the number of parameters between two models,

$$\text{AIC}(a, b) = G^2(a, b) + 2p, \quad (2a)$$

or alternatively

$$\text{AIC}(a, b)/2N = (\delta_a - \delta_b) + p/N, \quad (2b)$$

where p is the number of free parameters in θ_a minus the number of free parameters in θ_b .

Note, that the penalty term is relatively more important for small sample sizes, which increases the tendency to select simpler models. However, as the sample size increases ($N \rightarrow \infty$), then the AIC produces the same selection as the chi-square criterion. In short, similar to the chi-square method, it tends toward overly complex models with large sample sizes (for an example, see Browne, 2000).

Cross-validation criterion. Mosier (1951) presented the first clear presentation of the cross-validation criterion, and subsequently it has been elaborated by Mosteller & Tukey (1977) and Stone (1974) for regression, and Cudeck and Browne (1983) for covariance structure analysis. A recent overview of this method is provided by Camstra and Boomsma (1992).

The essential idea is to randomly divide the total *sample* of size N into two sub-samples of sizes N_1 and N_2 ($N_1 + N_2 = N$), producing two statistically independent vectors of sample statistics: $S_1(X)$ and $S_2(X)$. During the first *calibration stage*, best fitting parameter estimates $\theta_a(S_1)$ for model a are obtained by from $S_1(X)$ by selecting parameters that minimize the discrepancy

$$D\{Q_a[X, \theta_a(S_1)], S_1(X)\}$$

and similarly best fitting parameter estimates $\theta_b(S_1)$ for model b are obtained from $S_1(X)$ by selecting parameters that minimize

$$D\{Q_b[X, \theta_b(S_1)], S_1(X)\}.$$

Any legitimate discrepancy measure could be used to define D in the above calculations, including for example, the negative average log likelihood or a weighted sum of squared prediction errors. Then during the second *validation stage*, the previously estimated parameters from stage 1 are used again to compare the two models in terms of their predictive performance using the second independent sample:

$$cv = D\{Q_a[X, \theta_a(S_1)], S_2(X)\} - D\{Q_b[X, \theta_b(S_1)], S_2(X)\}.$$

If $cv > 0$, then choose model b , and if $cv < 0$, then choose model a .

This process can be repeated many times by using different divisions of the calibration and test samples to produce a distribution of cv statistics, from which one can compute the mean and standard deviation of the cv statistics (see Browne, 2000).

Like the AIC, the usefulness of cross-validation is limited to small sample sizes. For large sample sizes, cross-validation provides little or no additional information over a direct comparison of models using only the calibration stage. This is true for

the following simple reason: As $N \rightarrow \infty$, the sample statistics produced by the calibration sample converge to the same population values as the sample statistics produced by the validation sample, and consequently, the discrepancy measures for the calibration and validation stages will converge to identical magnitudes. Therefore, if the more complex model provides a lower discrepancy for a large calibration sample, it is almost guaranteed to produce a lower discrepancy for a large validation sample because the large sample statistics produced by the two samples are almost identical. Formally,

$$p \lim_{N \rightarrow \infty} S_1(X) = \mu(X) = p \lim_{N \rightarrow \infty} S_2(X)$$

and

$$p \lim_{N \rightarrow \infty} \theta_m(S_1) = \theta_m(\mu),$$

therefore

$$\begin{aligned} p \lim_{N \rightarrow \infty} D\{Q_m[X, \theta_m(S_1)], S_1(X)\} \\ &= D\{Q_m[X, \theta_m(\mu)], \mu(X)\} \\ &= p \lim_{N \rightarrow \infty} D\{Q_m[X, \theta_m(S_1)], S_2(X)\}. \end{aligned}$$

In short, there is only a calibration stage and *no* validation stage for large samples. For the purpose of selecting models based on sound scientific principles, the larger the sample the better are the conditions for comparing models, but cross-validation is not very useful in this ideal case. In fact, it has been shown that the cross-validation criterion is asymptotically equivalent to the AIC (Stone, 1977; Browne & Cudeck, 1989). Like the AIC, cross-validation tends to pick the more complex model in large samples (for an example, see Browne, 2000).

III. GENERALIZATION CRITERION

Mosier (1951) suggested an alternative method for comparing models called the *validity generalization* criterion that has received much less attention. Unlike cross-validation, which takes two samples from exactly the same design, the key idea is to employ two completely different designs. More specifically, the total *design* X is divided into two subdesigns X_1 and X_2 producing two statistically independent vectors of sample statistics: $S_1(X_1)$ and $S_2(X_2)$. During the calibration stage, the parameter estimates θ_a for model a are obtained by minimizing the discrepancy using the first design X_1 ,

$$D\{Q_a[X_1, \theta_a(S_1)], S_1(X_1)\},$$

and the parameter estimates θ_b for model b are also obtained by minimizing

$$D\{Q_b[X_1, \theta_b(S_1)], S_1(X_1)\}.$$

Then during the second generalization stage, the previously estimated parameters obtained from design X_1 are used again to compare predictive accuracy for the second design X_2 :

$$g = D\{Q_a[X_2, \theta_a(S_1)], S_2(X_2)\} - D\{Q_b[X_2, \theta_b(S_1)], S_2(X_2)\}.$$

If $g > 0$, then select model b , and $g < 0$ then select model a . Note that the model comparisons for the second stage are based on *a priori* predictions concerning *new* experimental conditions. Essentially, this tests the models ability to accurately interpolate and extrapolate, which is one of the major goals of a general scientific theory.

The generalization criterion method consists of the following steps:

1. Partition the complete design X into two subdesigns, a calibration design X_1 and a generalization design X_2 . The calibration design must provide precise estimates of the model parameters, and the generalization design must include conditions that distinguish the predictions of models being compared.

2. Using the calibration design, estimate the parameters for each model $\theta_m(X_1)$ that minimize a discrepancy D between the model predictions and the calibration sample statistics $S(X_1)$. (E.g., D can be defined as the negative log likelihood or weighted sum of squared errors.)

3. Check the standard errors of the parameter estimates to ensure that they are estimated with sufficient precision. (E.g., each parameter makes a statistically significant contribution to the fit in step 2. If the precision is inadequate, then redo step 1.)

4. Using the same parameter estimates from step 2, compute the new predictions for each model for the generalization design, $Q_m[X_2, \theta(X_1)]$. Check the predictions of each model to ensure they are distinguishable. (E.g., the discrepancy D between models must be nontrivial.) If discriminability is inadequate, then redo step 1.

5. Compute the discrepancy D between the predictions from step 4 and the sample statistics observed from the second design, $S(X_2)$, for each model.

6. Select the model that produces the smallest discrepancy in step 5 as the best in the set of models being compared.

Note that the generalization criterion can be used to compare any set of models, nested or nonnested, and they may even differ in terms of the number of parameters, because the selection is based on the *a priori* predictions computed from each model. The use of *a priori* predictions puts the models on equal footing in the generalization stage, despite the fact that the more complex model has an unfair advantage in the calibration stage.

IV. EXAMPLES

Two examples of the generalization criterion are provided below. The first is a simulation based on a simple hypothetical design and hypothetical data. This hypothetical case illustrates the utility of the generalization criterion in the ideal

case when the population parameters are known, so that sampling error does not confuse the basic issues. The second is a simulation based on a more complex published design and real data. This second example employs sample data that contain sampling error.

EXAMPLE 1. For the hypothetical design, imagine a preference study in which consumers decided whether or not to purchase a product. The utility of a product is represented by an independent variable, x , and the complete design contains 500 levels of product utilities represented by the transpose of the matrix

$$X = [0.01, 0.02, \dots, x_k, \dots, 4.99, 5.0].$$

The complete design is divided into two parts:

$$X_2 = [0.01, 0.02, 0.03, \dots, 3.50]$$

is the calibration design containing the first 350 levels; and

$$X_2 = [3.51, 3.52, \dots, 5.0]$$

is the generalization design containing the last 150 extrapolation levels.

The dependent variable is the relative frequency of purchases of a product with utility x_k denoted $S(x_k)$. The population proportions, $\mu(x_k) = p \lim_{N \rightarrow \infty} S(x_k)$, were generated by an S-shaped preference function bounded between zero and one:

$$\mu(x) = 1 - g(x, 1)/g(x, \infty)$$

$$g(x, y) = \int_0^y e^{-t(x-1)} dt.$$

Four models were compared using this design:

$$Q_t[x, \theta_t] = \mu(x) \quad (\text{true})$$

$$Q_s[x, \theta_s] = \theta_{s0} + \theta_{s1} \cdot x \quad (\text{simple})$$

$$Q_c[x, \theta_c] = \sum_{j=0, q} \theta_{cj} \cdot x^j \quad (\text{complex})$$

$$Q_p[x, \theta_p] = [1 + \exp(-\theta_{p1} - \theta_{p2} \cdot x)]^{-1} \quad (\text{principled})$$

Model *true* is the true data generator. Model *simple* is interpreted as the oversimplified model, model *complex* (with $q=3$) is interpreted as the overly complex model, and both of these models lack sound scientific principles about choice probability functions. For example, the polynomial model permits nonmonotonic relations between utility and choice probability, despite the fact that they are known to be increasing monotone. Also, both models *simple* and *complex* permit predictions that lie outside of the unit interval. Model *principled* is interpreted as the model based on sound scientific principles because it incorporates many of the

TABLE 1
Parameter Estimates (Standard Errors) for Each Model

<i>Principle</i>	<i>Complex (q = 3)</i>	<i>Simple</i>
0.56664 (0.1441)	0.6425 (0.0326)	0.5511 (0.0214)
1.7001 (0.1764)	0.3356 (0.0440)	0.2780 (0.0139)
	-0.0567 (0.0158)	
	-0.0219 (0.0176)	

important principles about the relation between utility and choice probability (S-shaped between zero and one), although it is clearly wrong in detail in this example because model *true* is the true model.

The negative of the mean log likelihood produced by each model was used to measure the population discrepancy between the model predictions and the population proportions for each design:

$$\alpha_m = - \sum_{k=1}^K \mu(x_k) \ln(Q_m[x_k, \theta_m]) + (1 - \mu(x_k)) \ln(1 - Q_m[x_k, \theta_m]). \quad (3)$$

The parameters for each model were obtained by minimizing the population discrepancy, α_m , between the predicted and population proportions using the calibration design. Table 1 shows the parameter estimates and their standard errors (based on the inverse of the Hessian matrix). Note that the standard errors for the parameters are reasonably small so that the parameter estimates are reasonably precise.

Then these same parameters were used to make *a priori* predictions for the generalization design. Table 2 provides a summary of the difference, $(\alpha_m - \alpha_t)$, between the negative mean log likelihood produced by model *m* and model *true*. The second column shows the comparisons based on the calibration design (first 350 levels) and the last column shows the comparisons based on the generalization design (last 150 levels).

First note that the overly complex model produces the minimum population discrepancy for the calibration design (based on fitted values rather than *a priori*

TABLE 2
Mean Log Likelihood Ratios for Hypothetical Example

Model	Design	
	Calibration (X_1)	Generalization (X_2)
<i>Principle</i>	1.61	3.63827
<i>Complex</i>	0.21	9.56196
<i>Simple</i>	15.85	^a

^a Undefined because all predicted values exceeded 1.0.

predictions). More important, the principled model produces the minimum population discrepancy for the generalization design. The reason for this reversal is the following fact: when the overly complex model is fit to the calibration design, it selects a preference function that is nonmonotonic in the upper extrapolation region.

The models were compared again using an alternative way to divide the complete design into calibration and generalization designs.

$$X_1 = [1.51, 1.52, \dots, 2.50]$$

is the calibration design containing the 100 intermediate levels; and

$$X_2 = [0.01, 0.02, \dots, 1.50] \cup [2.51, 2.52, \dots, 5.0]$$

is the generalization design containing $150 + 250 = 400$ extrapolation levels. In addition, the overly complex model was defined by a fifth order rather than a third order polynomial. Furthermore, the standard method of comparing models was also performed by using the parameters that minimized the population discrepancy between predicted and population proportions using *all* the levels from the complete design.

Table 3 provides a summary of the difference, $(\alpha_m - \alpha_t)$, between the negative mean log likelihood produced by model m and model *true*. The second column shows the comparisons based on the calibration design (middle 100 levels), the third column shows the comparisons based on the generalization design (extrapolation levels), and the last column shows the comparisons based on the complete design (all 500 levels).

Once again, the overly complex model produces the minimum population discrepancy for the calibration design, and it also produces the minimum for the complete design (both are based on post hoc fitted values rather than *a priori* predictions). More important, the principled model produces the minimum population discrepancy for the generalization design. As before, the reason for this reversal is that when the overly complex model is fit to the calibration design, it selects a preference function that is nonmonotonic in the upper extrapolation region. This reflects a general problem with the overly complex model—it fails to incorporate theoretical principles that impose appropriate constraints on the predictions.

TABLE 3
Mean Log Likelihood Ratios for Hypothetical Example

Model	Design		
	Calibration (X_1)	Generalization (X_2)	Complete (X)
<i>Principle</i>	0.000181	3.63827	1.74772
<i>Complex</i>	0.000000	9.56196	0.00564
<i>Simple</i>	0.015205	95.70174	31.49282

Flexibility and lack of appropriate constraints produce better fits at the cost of poor extrapolation performance.

EXAMPLE 2. The second example is based on a model comparison that used real data reported by Busemeyer and Townsend (1993). College students were asked to choose between two risky courses of action described by a payoff matrix of the form illustrated below:

	State of nature	
Action	1	2
A	g_A	$-c_A$
B	$-c_B$	g_B

For example, if action A is chosen and state 1 is observed, then a gain equal to g_A dollars would be earned; but if action A is chosen and state 2 is observed, then a cost equal to c_A dollars would be lost. The complete design included a total of 29 conditions formed by manipulating the state probabilities (π and $1 - \pi$) and the payoff magnitudes (g_A, c_A, g_B, c_B). This total design was divided into two parts: the calibration design consisted of 17 conditions where the payoffs varied from small to medium magnitudes; and the generalization design consisted of 12 conditions where the payoffs were relatively large in magnitude (see Tables 6 and 7 in Busemeyer & Townsend, 1993, for details). The dependent variable was the proportion that action A was chosen over action B.

The true model underlying the *real* data is unknown. However, the model used to generate the *simulated* data was constructed as follows. First, the general form of the simulation model was derived from *decision field theory*, which has successfully explained a large number of empirical facts found in the risky decision making literature (see Busemeyer & Townsend, 1993). Second, the parameters of the simulation model were obtained by selecting parameter estimates that maximized the likelihood of the real data.

This produced the following model for generating the *simulated* population proportions (denoted μ as a function of the known variables g_A, c_A, g_B, c_B , and π),

$$\mu(g_A, c_A, g_B, c_B, \pi) = \{1 + \exp[-(2.86)(d/v)]\}^{-1}, \quad (\text{true})$$

where d is the difference in expected utilities for the two actions,

$$d = \pi \cdot [u(g_A) - u(-c_B)] + (1 - \pi) \cdot [u(-c_A) - u(g_B)], \quad (4a)$$

and v is the variance of the utility difference,

$$v = \pi \cdot [u(g_A) - u(-c_B) - d]^2 + (1 - \pi) \cdot [u(-c_A) - u(g_B) - d]^2, \quad (4b)$$

and the utility function was defined by a two piece power function

$$u(g) = (0.545) g^{(0.803)} \text{ for gains, } (g > 0) \quad (4c)$$

$$u(-c) = -(0.348) c^{(0.922)} \text{ for losses, } (-c < 0). \quad (4d)$$

The four constants used to define the utility functions in Eqs. (4c), (4d) were selected to maximize the likelihood of the experimentally observed sample proportions.

The variance term (v) in Eq. (4a) is essential for explaining violations of *strong stochastic transitivity*¹ that often occur in choice experiments (see Busemeyer & Townsend, 1993). The basic idea is that the discriminability of a difference in utility between two actions is moderated by the variance of the utility difference. When the variance is small, discriminability is high, producing choice probabilities closer to zero or one; but when the variance is large, discriminability is low, producing choice probabilities closer to 0.50.

Three probabilistic models of risky choice are compared using simulated sample proportions generated by the simulation model [Eq. (4)]. Model *simple* is interpreted as an oversimplified model, because it has only one free parameter and it omits the variance term (v) in Eq. (4a):

$$Q_a[(g_A, c_A, g_B, c_B, \pi), \theta_a] = \{1 + \exp[-d]\}^{-1} \quad (\text{simple})$$

$$d = \pi \cdot [u(g_A) - u(-c_B)] + (1 - \pi) \cdot [u(-c_A) - u(g_B)]$$

$$u(g) = g^{\theta_a} \text{ for gains } (g > 0),$$

$$u(-c) = -c^{\theta_a} \text{ for losses, } (-c < 0).$$

The one free parameter of the oversimplified model is used to define the utility function.

Model *complex* is interpreted as the overly complex model, because it has six free parameters but it also omits the variance term (v):

$$Q_b[(g_A, c_A, g_B, c_B, \pi), \theta_b] = \{1 + \exp[-d]\}^{-1} \quad (\text{complex})$$

$$d = s(\pi) \cdot [u(g_A) - u(-c_B)] + (1 - s(\pi)) \cdot [u(-c_A) - u(g_B)]$$

$$u(g) = \theta_{1b} g^{\theta_{2b}} \text{ for gains } (g > 0),$$

$$u(-c) = -\theta_{3b} c^{\theta_{4b}} \text{ for losses } (-c < 0),$$

$$s(\pi) = \theta_{5b} \pi^{\theta_{6b}}.$$

Four of the six parameters of the overly complex model are used to define the utility function, and the other two are used to define a subjective probability function.

¹ Strong stochastic transitivity is defined by the following property. Define A, B, and C as three arbitrary actions, and let $\Pr(X, Y)$ denote the probability of choosing X over Y . If $\Pr(A, B) \geq 0.5$, and $\Pr(B, C) \geq 0.5$, then $\Pr(A, C) \geq \max[\Pr(A, B), \Pr(B, C)]$.

Model *principled* is interpreted as the principled model, because it contains the critical variance term:

$$\begin{aligned}
 Q_c[g_A, c_A, g_B, c_B, \pi, \theta_c] &= \{1 + \exp[-(d/v)]\}^{-1} && (\textit{principled}) \\
 d &= \pi \cdot [g_A - (-c_B)] + (1 - \pi) \cdot [(-c_A) - g_B] \\
 v &= \theta_{1c} + \theta_{2c} \cdot \{ \pi \cdot [g_A - (-c_B) - d]^2 + (1 - \pi) \cdot [(-c_A) - g_B - d]^2 \}.
 \end{aligned}$$

The principled model incorrectly assumes a linear utility function and it has only two free parameters associated with the variance term.

The sample proportions used in this analysis were simulated using the following procedure. Recall that the complete design consists of 29 payoff conditions. The true model [Eq. (4)] was used to generate simulated sample proportions for these 29 conditions. One simulated replication of this design produces a vector of 29 sample proportions, $[S_{1,r}, \dots, S_{k,r}, \dots, S_{29,r}]$, where $S_{k,r}$ is the sample proportion for the k th payoff condition obtained from the r th simulated replication. Each replication was produced by randomly sampling from a binomial distribution with a sample size equal to $N = 2000$ for each payoff condition.² A total of 55 replications of the complete design was generated, producing 55 vectors of sample proportions. The models were fit separately to each vector of simulated sample proportions, producing 55 discrepancy indices per model, and these 55 discrepancy indices were averaged to produce a mean discrepancy for each model.

Maximum likelihood estimates of the parameters were obtained using the calibration design for each model and replication by minimizing the sample discrepancy

$$\delta_{m,r} = - \sum_{k=1}^K S_{k,r} \ln(Q_m[x_k, \theta_m]) + (1 - S_{k,r}) \ln(1 - Q_m[x_k, \theta_m]). \quad (5)$$

The arithmetic average over 55 replications was used as a single estimate of the sample discrepancy, δ_m , for each model. A G^2 statistic was computed for each model using Eq. (1b):

$$G^2 = 2 \cdot N \cdot (\delta_m - \delta_t).$$

AIC indices were then obtained for each model from Eq. (2a).

Table 4 shows the results of the model comparisons for the calibration and the generalization designs. The second column shows the AIC values for the calibration phase, and the last column shows the G^2 produced by the generalization design. As can be seen in this table, the AIC favors the overly complex model in the calibration phase, but the principled model performs an order of magnitude better than the overly complex model in the generalization design. In fact, the overly complex

² This matches the sample size for the real sample proportions used in the real model comparison by Busemeyer and Townsend (1993). Simulated data were used rather than real data for two reasons—one is that the true model can be defined for simulated data and the second is that simulated replications of the design are possible with simulated data.

TABLE 4
Results of Model Comparison for Risky Decision Experiments

	Design	
	Calibration	Generalization
Models <i>simple vs true</i>	AIC = 840.8 (24.1)	$G^2 = 6420$ (45.1)
Models <i>complex vs true</i>	AIC = 44.8 (17.0)	$G^2 = 74,216$ (3333.1)
Models <i>principled vs true</i>	AIC = 176.8 (17.0)	$G^2 = 272$ (17.0)

Note. Numbers in parentheses are the standard errors of the cell means based on 55 replications.

model performs more poorly than the oversimplified model in the generalization design.

Both the oversimplified and the overly complex models fail in the generalization design because they lack the principle needed to explain violations of strong stochastic transitivity. In sum, the generalization design was critical for selecting the principled model, which is the only model that is based on the correct principles (but wrong in detail).

V. CONCLUDING COMMENTS

The purpose of this article is to formalize the generalization criterion method for model comparison. The method has the potential to provide powerful comparisons of complex and nonnested models that may also differ in terms of numbers of parameters. The generalization criterion differs from the better known cross-validation criterion in the following critical procedure. Although both employ a calibration stage to estimate parameters, cross-validation employs a replication sample from the *same design* for the validation stage, whereas generalization employs an entirely *new design* for the critical generalization stage. The main advantage of the latter is the emphasis placed on extrapolations to *new* experimental conditions in the generalization design. Accurate *a priori* predictions to new conditions are the hallmark of a good scientific theory.

The generalization criterion serves a different purpose than the crossvalidation criterion.³ The generalization criterion is useful with large sample sizes to select a model based on sound scientific principles (but wrong details) from a set of models that also contain *oversimplified* models or *overly complex* models (where both of the latter are lacking sound scientific principles). The cross-validation criterion is useful with smaller sample sizes to select the model that yields the smallest discrepancy expected from using this model again on another replication. Therefore, the two criteria should be viewed as complementary and useful for different purposes, rather than as competitors for the same purpose. In fact, cross-validation (or AIC) could be used first to decide which models to include in a subsequent generalization criterion comparison.

³ The AIC can be interpreted as a single sample estimate of the cross-validation criterion, and therefore it serves the same purpose as cross-validation.

The generalization criterion places very demanding requirements on the researcher to accomplish two things simultaneously: one is to select a *comprehensive* calibration design that provides *precise* estimates of the parameters (measured by their standard errors), and the second is to provide a *diagnostic* generalization design that discriminates among competing models. If either of these two design issues are treated inadequately, then the generalization criterion method will fail to work properly.

The generalization criterion also requires an important theoretical assumption—parameter invariance across calibration and generalization design conditions. Good scientific theories usually satisfy the property of parameter stability, so that new parameters do not need to be estimated for every new situation. An implicit assumption of the generalization criterion method is the existence of continuous mappings from experimental variables to causal theoretical variables. For example, the generalization criterion method works well for regression models, which provide a continuous map from predictor variables to dependent variables; but this method does not work well for analysis of variance models, which allow arbitrary mappings and require new parameter estimates with each addition of a new experimental condition.

It is possible to employ weaker variations of the generalization criterion. Suppose two complex models, *a* versus *b*, are compared, and *p* parameters are estimated for model *a* and *q* parameters are estimated for model *b* during the calibration stage; but then *r* additional new parameters must be estimated for both models for the generalization design. The generalization criterion may still be effective as long as the *same* number of parameters are estimated in the generalization design. For example, the causal parameters of two different structural equation models may be estimated from one set of indicators during the calibration phase. Then these same causal parameters are used to make predictions for a new set of indicators in the generalization stage. But a common set of measurement model parameters must be estimated for both models in the generalization stage.

Another variation of the generalization criterion is to use one set of dependent variables or measures to estimate the parameters of each model during the calibration stage and then use these same parameters to make predictions for another set of dependent variables in the generalization stage. This type of generalization criterion may still be effective as long as the measures used for calibration and generalization are not too highly correlated. For example, in another application reported in Busemeyer & Townsend (1993), parameters of choice models were estimated from the choice probabilities, and then these same parameters were used to make predictions for choice response time.

A third variation of the generalization criterion is to *randomly* partition the complete design containing *K* conditions into two subdesigns, a calibration subdesign with K_1 conditions and a generalization subdesign with K_2 conditions. The random division could be repeated many times to produce a distribution of generalization criteria for each model. This corresponds to the bootstrapping procedure used with cross-sample validation (Efron & Gong, 1983). However, a partition that is skillfully crafted by a clever researcher will generally produce a more diagnostic generalization design than a random partition.

A fourth variation is to compute Bayesian predictions for the generalization criterion from each model. First, the posterior distribution, $f_m[\theta_m | S(X_1)]$, over the parameter space can be estimated for each model from the data obtained in the calibration stage. Second, the mean of the predictions

$$E[Q_m | f_m] = \int Q_m[X_2, \theta_m] f[\theta_m | S(X_1)] d\theta_m$$

can then be inserted into

$$\delta_m = D[E[Q_m | f_m], S(X_1)]$$

to compute the sample discrepancy for each model during the generalization stage.

Alternatively, the expectation of the discrepancy can be directly computed:

$$E[\delta_m | f_m] = \int D[Q_m[X_2, \theta_m], S(X_2)] f[\theta_m | S(X_1)] d\theta_m.$$

This Bayesian generalization criterion could complement the standard Bayesian approach to model comparison (see Myung & Pitt, 1997, for example) which makes use of the complete design.

Finally, the use of the generalization criterion does not preclude the use of other model comparison methods. After performing a generalization criterion comparison, one can always recombine all of the conditions, and reevaluate the global fit of each model using AIC or BIC using the complete design. The latter has the advantage of providing improved parameter estimates. The essential point is to plan ahead and include a strong generalization design within the complete design.

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory*. Budapest: Adadeiai Kiado.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Bozdogen, H. (2000). Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology*, **44**, 62–91.
- Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, **24**, 445–455.
- Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, **100**, 432–460.
- Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods and Research*, **21**, 89–115.
- Collier, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception and Psychophysics*, **38**, 476–481.
- Cudeck, R., & Browne, M. W. (1983). Cross validation of covariance structures. *Multivariate Behavioral Research*, **18**, 147–167.

- Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364–381.
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, **81**, 95–106.
- Efron, B., & Gong, G. (1983). A leisurely look at the Bootstrap, the Jackknife, and cross-validation. *American Statistician*, **37**, 36–48.
- Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition—sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, **20**, 1311–1334.
- Krantz, D. H., & Tversky, A. (1971). Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, **78**, 151–169.
- Linhart, H., & Zucchini, W. (1986). *Model selection*. New York: Wiley.
- Mood, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics*. New York: McGraw–Hill.
- Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, **11**, 5–11.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison–Wesley.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, **4**, 79–95.
- Platt, J. R. (1964). Strong inference. *Science*, **146**, 347–353.
- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, **103**, 734–760.
- Rust, R. T., Simester, D., Brodie, R. J., & Nilikant, V. (1995). Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. *Management Science*, **41**, 322–333.
- Schwarz, J. H. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
- Schweickert, R., Fisher, D. L., & Goldstein, W. M. (1998). Latent task networks: Structural and quantitative analysis of networks of cognitive processes. Manuscript submitted for publication.
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–133.
- Stone, M. (1977). On asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B*, **39**, 44–47.
- Townsend, J. T. (1990). Serial versus parallel processing: sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, **1**, 46–54.
- Wickens, T. D. (1998). On the form of the retention function: Comments on Ruben and Wenzel. *Psychological Review*, **105**, 379–386.
- Wixted, J. T., & Ebbeson, (1991). On the form of forgetting. *Psychological Science*, **2**, 409–415.

Received: November 21, 1997; revised: July 29, 1998