HHS Public Access

Author manuscript

Behav Processes. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

Behav Processes. 2017 November ; 144: 20–32. doi:10.1016/j.beproc.2017.08.004.

Methods of Comparing Associative Models and an Application to Retrospective Revaluation

James E. Witnauer[1], **Ryan Hutchings**[1], and **Ralph R. Miller**[2]

[1]State University of New York at Brockport

[2]State University of New York at Binghamton

Abstract

Contemporary theories of associative learning are increasingly complex, which necessitates the use of computational methods to reveal predictions of these models. We argue that comparisons across multiple models in terms of goodness of fit to empirical data from experiments often reveal more about the actual mechanisms of learning and behavior than do simulations of only a single model. Such comparisons are best made when the values of free parameters are discovered through some optimization procedure based on the specific data being fit (e.g., hill climbing), so that the comparisons hinge on the psychological mechanisms assumed by each model rather than being biased by using parameters that differ in quality across models with respect to the data being fit. Statistics like the Bayesian information criterion facilitate comparisons among models that have different numbers of free parameters. These issues are examined using retrospective revaluation data.

Keywords

mathematical models of learning; Pavlovian conditioning; associative learning; Bayesian information criterion; free parameters; retrospective revaluation

1. Introduction: Models and computational simulations

Acceptance or rejection of quantitative models of associative learning should be informed by how well a model fits the available data *relative* to alternative models. We propose that tests of associative models should be based on formal *comparative* model selection procedures and statistics. For example, human category learning researchers who have made claims

Mailing Address: Ralph R. Miller, Department of Psychology, SUNY – Binghamton, Binghamton, NY 13902-6000, USA, TEL: (607) 777-2291, FAX: (607) 777-4890, rmiller@binghamton.edu.

The simulators described in this paper are available online as supplemental material to this paper. The supplemental material contains a set of commented scripts that documents the procedures used in the present simulations. Additional materials can also be downloaded by pasting the following link into the navigation bar of a web browser: http://courses.brockport.edu/~jwitnaue/WitnauerEtAl(BPSubmission).zip Those additional materials include the complete results of the present simulations, including the best-fitting parameters discovered by each iteration of the hillclimbing algorithm.

about the merits of exemplar- and prototype-based models of category judgments for years have bolstered their assertions with such simulations. For them, it is not sufficient to show that a model simply provides a good fit to data from empirical procedures. Instead, their modeling procedures usually include comparisons between models (e.g., Minda & Smith, 2001; Nosofsky & Zaki, 2002). Similarly, a few researchers interested in animal learning have in fact adopted these techniques (e.g., Cheung, Neisewander, & Sanabria, 2012; Simen, Balci, de Souza, Cohen, & Holmes, 2011). However, this approach seems to be largely absent from the basic Pavlovian literature (e.g., see the 2012 special issue of *Learning & Behavior*). To be sure, modelers disagree concerning best practices for such comparisons (e.g., Olsson, Wennerholm, & Lyxzen, 2004) and the present article does not aim to resolve those disagreements because such comparisons are almost absent from the associative learning literature. Instead, we merely propose that claims about associative models could be better assessed if they were considered in the framework of a comparative model selection method that includes both an optimization procedure for identifying best-fitting parameters (e.g., hill climbing) and a statistical analysis of simulation results that controls for differing numbers of free parameters across models.

A mathematical model of behavior allows researchers to precisely instantiate a theory as a system of equations. Computational models facilitate assessment of theories by producing quantitative predictions that can be scaled to the response measurements used in experimental procedures (e.g., choice behavior, response rate, or response magnitude). Such computer-aided simulations are necessary because contemporary models of learning and behavior are quite complex. Thus, the predictions of an associative model as well as the parameter dependence of it are frequently difficult to ascertain without conducting simulations. The complexity of most learning models is often necessitated by seemingly discrepant empirical findings. For example, when multiple cues are reinforced in compound, competition among the cues is often observed (e.g., overshadowing, Pavlov, 1927; blocking, Kamin, 1968); however, facilitation can occur in select circumstances (e.g., potentiation; Rusiniak, Hankins, Garcia, & Brett, 1979). And when multiple cues are presented in compound without reinforcement and one of those cues is separately reinforced, either excitatory behavioral control (second-order conditioning) or inhibitory behavioral control (Pavlovian conditioned inhibition) by the nonreinforced cue can result (Pavlov, 1927). Further complicating any theoretical analysis of the empirical literature is the fact that attempts to replicate even widely cited empirical effects have not been universally successful (e.g., Maes et al., 2016).

Early models of associative learning (e.g., Rescorla & Wagner, 1972) explained only a relatively small category of associative effects, centrally cue competition and conditioned inhibition. Such models are so simple that their predictions can often be derived either analytically or intuitively. For example, the Rescorla-Wagner model assumes that animals learn and respond based on a $n \times 1$ array of associations to a given unconditioned stimulus (US), where n equals the number of cues. Anticipation of the blocking effect is easily derived from the system of equations in the Rescorla-Wagner model. Assuming that Phase 1 training was asymptotic (i.e., $V_A = \lambda$), the change in associative strength for B (ΔV_B) in the experimental group on the first trial of Phase 2 is

$$\Delta V_B = \alpha_B * \beta_{US} * (\lambda - [V_A + V_B]) = \alpha_B * \beta_{US} * (\lambda - \lambda) = 0 \quad (1)$$

where α_B is the associability of B, β_{US} is the associability of the US, and λ is the maximum associative strength supportable by the US. Thus, the prediction is that an unsurprising US will fail to produce learning with respect to B. Computational modeling is not necessary in this situation because the predictions of the Rescorla-Wagner model can readily be arrived at 'intuitively.' Occasionally, an elegant closed-form solution to a model can be derived analytically (e.g., Yamaguchi, 2006). However, the assumptions of these closed solutions are not always consistent with the empirical procedures used in experimental tests of those models. For example, predictions based on some closed solutions to the Rescorla-Wagner model assume that training has reached asymptote and that all trial types are presented in random order (Danks, 2003). Thus, closed solutions to a model can result in a weakened connection between the assumptions of the model and the empirical procedures that the model is intended to simulate. This is not a weakness of the Rescorla-Wagner model, but it is weakness in the use of closed solutions to derive predictions from the Rescorla-Wagner model.

Moreover, early, more intuitive models of associative learning, such as the Rescorla-Wagner model, fail to explain behavior in many well-documented ways. For example, the Rescorla-Wagner model fails to explain latent inhibition (Lubow & Moore, 1959), spontaneous recovery from extinction (Pavlov, 1929), retrospective revaluation (Kaufman & Bolles, 1981), taste-odor potentiation (Rusiniak, Hankins, Garcia, & Brett, 1979), and backward blocking (Shanks, 1985; for reviews, see Miller, Barnett, & Grahame, 1995; Witnauer, Urcelay, & Miller, 2014). Additionally, the Rescorla-Wagner model does not explain cue facilitation effects such as potentiation or second-order conditioning. In blocking situations, the model fails to explain a number of different observations. Blocking is sometimes reduced with posttraining extinction of the blocking cue (i.e., retrospective revaluation; e.g., Blaisdell, Gunther, & Miller, 1999). Blocking does not occur after massed elemental trials (Wheeler & Miller, 2007) or with a long retention interval between compound training and testing (Piñeno, Urushihara, & Miller, 2005). Blocking also depends on the number of compound trials; the effect is reduced with large numbers of compound training trials (Azorlosa & Cicala, 1986). Moreover, blocking fails to occur when the target stimulus is trained in compound with two previously independent excitors (Witnauer, Urcelay, & Miller, 2008). Thus, contemporary models of associative learning have been revised. For example, revisions of the Rescorla-Wagner model have been developed to explain retrospective revaluation (e.g., Van Hamme & Wasserman, 1994) and second-order conditioning (e.g., Piñeno, 2007). However, these revised models are necessarily more complex than their predecessors and their predictions must be derived using computational (instead of analytical or intuitive) methods.

Contemporary associative theories often consider the role of within-compound associations (e.g., Van Hamme & Wasserman; Stout & Miller, 2007), microelements (Ludvig, Sutton, & Kehoe, 2008), or both (Wagner, 1981). Simulations of these models become computationally intensive when they are applied to experiments with multiple cues, many trials, or an

approximation of real time representation of events within a trial. In contrast to the simplicity of the Rescorla-Wagner model, the Van Hamme and Wasserman (1994) model's dependence on within-compound associations implicitly assumes that, given a single US, animals learn and respond based on an n x n+1 array of associations, where n is the number of cue elements. It is difficult for modelers to derive the predictions of this model in experimental situations (which have increased in complexity, along with the number of models that they attempt to differentiate) without computational methods. For example, Rescorla's (2000) illuminating observation that excitors gain less behavioral control than inhibitors, when they are compounded and paired with a US, is based on a procedure that included six different stimuli (including the training context) and an outcome. In this situation, the Rescorla-Wagner (1972) model assumes that the critical contents of learning are limited to six different associations (i.e., a six-element vector representing the associative strength of each cue with the US). In contrast, Van Hamme and Wasserman's extension of the Rescorla-Wagner model assumes that animals learn and respond based on a complete unidirectional associative matrix where each row represents a 'sending cue' and each column represents a 'receiving stimulus'. Thus, the Van Hamme and Wasserman model assumes that the critical contents of learning includes up to 42 (6 x 7) different associations in Rescorla's (2000) experiments.

Another problem with many conventional empirical papers in the study of associative learning is that they often use null hypothesis significance testing as the primary quantitative technique for making theoretical inferences based on behavioral data. This approach is of limited value because comparisons between observed effects and null hypotheses usually fail to quantitatively assess the specific theoretical prediction in question. That is, using this technique, hypotheses are dichotomously either not supported or viewed as viable without assessing the degree to which the observed behavior conformed to predictions of the model. Rodgers (2010) reviewed the problems with null hypothesis testing and argued that comparative computational modeling is emerging as an important family of techniques for overcoming this and other related problems. In reality, this movement is far from new as in some areas of psychology (e.g., category learning) researchers have been using these alternative techniques for decades (e.g., Kruschke, 1992). In this context, simulations are important because they introduce quantitative rigor to the difficult task of relating data to specific theories.

Intuitive and analytical methods of relating a model to the null hypothesis are inadequate as they usually assess only ordinal differences. Computational models allow researchers to derive precise, numerical predictions that can be quantitatively compared to observed behavior. Most researchers in associative learning are well aware of the existence of these methods, but rarely employ them. Moreover, when they do, they are usually applied to a single model rather than a comparison across models (e.g., *Learning & Behavior*'s 2012 Special Issue on Computational Models of Classical Conditioning). However, in practice theoretically driven empirical research often does not actually employ computational methods at all (e.g., Melchers, Lachnit, & Shanks, 2004), which is problematic because investigators might err concerning what a given model predicts. The purpose of this paper is to highlight the merits of quantitative comparative model testing and provide an example. There are several different general computational methods available to researchers, some of

which more readily permit direct comparisons across models than others. These methods differ with respect to how free parameters are selected. Free parameters can be optimized *post hoc* or selected *a priori*. Centrally, we argue that parameters should be optimized on a *post hoc* basis when simulations are used to compare models with respect to goodness of fit to observed behavior.

2. Model comparison and selection

The scientific method depends on selective, empirically-based rejection of some theories *relative* to others. Computational methods can inform this process. When one mathematical model fits a behavioral data set better than another, we learn something about the theoretical processes underlying that behavior. This process is made more difficult by the fact that models are intentionally simplified systems. For example, the Rescorla-Wagner model makes many simplifying assumptions (which are ideally constant across models being compared) concerning perception, attention, memory, and decision making. If the Rescorla-Wagner model fails to fit a data set, is that failure caused by the model's critical theoretical ideas (error detection and reduction) or is it caused by the model's simplifying assumptions? We propose that model comparison can assist researchers in answering this question. Comparing two (or more) models with similar simplifying assumptions and different critical theoretical principles allows a potentially illuminating computational comparison of principles. In the same way that an empirical experiment compares experimental and control conditions that are well-matched on some variables and different on (ideally) one target variable, a 'computational experiment' compares models that are well-matched in their simplifying assumptions and different in their theoretical assumptions. In this situation, a distinct superiority of one model over another is more likely to reflect a problematic theoretical assumption than a problematic simplifying assumption, at least with respect to the data set at hand. Importantly, the goal of this strategy is not to select a 'winner;' this strategy does not confirm a single model or class of models. The strategy will at best only identify models that might be rejected on the basis of data, and then only with respect to the procedures used to generate the data set in question.

Ultimately, simulations of a data set using a single model can at best tell us only the amount of variance between conditions that the model can explain. Should we be content with 90%, 95%, or only 99%? Such criteria do not allow us to either falsify or fail to falsify a model, and in fact that is not conceptually possible by simulating only one model. The important question is how well a model performs relative to alternative models. We view comparative modeling as a more conservative strategy than simulating a single model because it can identify experiments that are well-fit by two or more competing models. Consider the blocking phenomenon. All contemporary associative models predict blocking (which is not surprising because a model that failed to would not be entertained). Thus, a 'successful' simulation of blocking using a contemporary model (e.g., Stout & Miller, 2007; Harris, 2006) would be uninformative in choosing among models. That is, if simulations of a contemporary model were compared to Rescorla-Wagner model simulations, both models would explain the basic blocking effect and we would be forced to accept the Rescorla-Wagner model based on its greater simplicity (i.e., far fewer free parameters). Thus, failures to falsify are only compelling in the context of less successful alternative models. Otherwise,

we must always accept the simplest model (i.e., formally, model selection reduces to selecting the simplest 'successful' model). Notably, this points to the need for Pavlovian researchers to continually update the empirical benchmarks against which a model's predictions are compared. Such revisions to empirical benchmarks should be informed by formal modeling procedures as those procedures are best able to reveal the empirical effects that have the potential to differentiate among models. We have used retrospective revaluation (e.g., Kaufman & Bolles, 1981) and counteraction effects (Wheeler & Miller, 2008) to differentiate among models (e.g., Witnauer & Miller, 2011; Witnauer, Urcelay, & Miller, 2014). However, formal, simulation-based *comparisons between models* are almost absent from the animal learning literature (e.g., Kutlu & Schmajuk, 2012). Formal model comparison is the most appropriate strategy for differentiating among models when multiple models are able to explain some empirical benchmark. For example, most associative models predict blocking but formal model comparisons have revealed that only some associative models are able to accurately predict the degree of blocking (usually incomplete) that is observed in most experiments (e.g., Witnauer et al., 2014). Moreover, even when reasonable empirical criteria seem to differentiate between models, researchers can underestimate the flexibility of models, which might lead to erroneous rejection of a potentially viable model.

Likewise, the falsification of a model depends on the context in which the falsification occurs. All mathematical associative models are, by definition, simplified implementations of some theory or system of principles that instantiate focal theoretical ideas. They are not meant to serve as complete descriptions of an animal's cognitive and behavioral system. Thus, all models are inherently inaccurate. Associative models often oversimplify mechanisms of attention, memory, and perception that are known to contribute to behavior. The update equation (i.e., $V^{n+1} = V^n + \Delta V^n$) that is ubiquitous in associative models deliberately oversimplifies the storage of information in memory in that it implies no retention of prior associative states. Comparisons across models are usually most illuminating when the models being compared subscribe to similar simplifying assumptions.

The capacity of a model to fit empirical data is most informative when it is compared to another model's fit to the same data in the same manner that a group mean is most informative when it is compared to the means of other groups that have been differentially treated. In these situations, researchers compare models based on goodness of fit to data, weighted by simplicity (usually captured by the number of free parameters). The enormous potential of this strategy was well-described by Rodgers (2010). He argued that simulations could be designed in a manner similar to the way that experiments are designed. Comparisons between fits of models can be used to evaluate contrasting theoretical principles. Like comparisons between experimental conditions with tightly controlled differences in treatment, comparisons between models with clearly defined differences and similarities in structure can inform concerning the mechanisms of behavior. Witnauer and Miller (2011) attempted this with their treatment of two models of retrospective revaluation. They compared the original Van Hamme and Wasserman (1994) model and the Within-Compound model (a distinct variant of the Van Hamme & Wasserman model) that differ systematically in their treatments of absent cues. Both models assume that representations of absent cues are activated (i.e., retrieved) by presentation of their associates. The original Van

Hamme and Wasserman model assumes that the salience of an absent cue is a constant (negative) fraction of the salience of the same cue when it is presented. In contrast, the Within-Compound model assumes that the salience of an absent cue is proportional to the amount of activation of the absent cue achieved through the within-compound association. The two models share the important psychological assumption of Van Hamme and Wasserman that the associative strength of absent cues can be modified, which is a significant departure from the original Rescorla-Wagner model on which they are based. Across many simulations, comparisons between these models revealed that the Within-Compound model provided a better fit to experimental results, which suggests that the strength of the within-compound association mediates the strength of retrospective revaluation. Thus, using computational modeling to fit theoretical models to empirical data can illuminate the mechanisms underlying effects.

3. Parameter selection and quantifying procedural variables

In some situations, the values of parameters are strongly constrained by previously established theoretical principles or empirical findings. For example, in Hodgkin and Huxley's (1952) biophysical model of the membrane of a neuron, the values of many of the parameters were already known from independent prior research. The value of the sodium reversal potential is calculated based on, among other things, the temperature of the experimental apparatus and the intracellular and extracellular concentrations of sodium ions. Unfortunately, such precision is often difficult in associative modeling because the parameters are rarely closely linked to readily measurable physical variables. Even salience parameters, which are related to the physical intensity of a stimulus, are not perfectly related to any single physical dimension of a stimulus. For example, the salience of an acoustic stimulus is determined by many physical dimensions of the stimulus (e.g., amplitude, frequency, timbre, and duration) as well as state variables of the subject (e.g., arousal), all of which often vary from one experimental preparation to another. Similar problems exist in other sensory modalities. Thus, associative models often use free parameters to represent variables like salience, which are (by definition) permitted to vary in a way that can influence the predictions of the model. Free parameters also assist in quantifying abstract parameters like associability that are often superficially viewed as being closely linked to variables like salience, but are clearly far removed from being determined entirely by readily measurable aspects of the physical stimulus. Ideally, one should work toward developing models that produce parameter- and procedure-invariant predictions, but only to the extent that the data are invariant. Including more data in the optimization procedure is one way to constrain a model's predictions and work towards parameter invariance. The values of some parameters are estimable only by procedures that fit the model to empirical data. Researchers in associative learning often omit such procedures and, instead, intuitively select parameters and then simulate multiple data sets with a single model without changing parameters. If the observed pattern of behavior resembles the pattern of results predicted by the model, researchers argue that the model successfully predicted the data. This is sometimes misleading for at least two reasons. The first concerns how procedural variables are quantified, and the second concerns the selection of parameters. Here we sequentially address these two issues.

There are many experimenter-controlled procedural variables that must be quantified before a computational model can generate predictions, and how this quantification is implemented can have enormous influence on the model's goodness of fit. Moreover, the consequences of these decisions can differ markedly across models. Two examples of such variables are how to treat intertrial intervals, and how many trials to assume have occurred. Intertrial intervals can be ignored, each can be treated as one context extinction trial independent of the length of the intertrial interval, or even as a number of context extinction trials with the number of trials being a direct function of the length of the intertrial interval which is possibly measured in terms of the duration of trials in which punctate stimuli (i.e., conventional CSs and USs) are presented. The literature reveals instances of each of these three strategies.

With respect to number of trials, one might think that matching number of trials in a simulation to the actual number of trials that were administered in the experiment being simulated would be the only acceptable option, but modelers often deviate from this. For example, Kutlu and Schmajuk (2012) used the SLGK model (presented in Kutlu & Schmajuk) to simulate Blaisdell et al.'s (1999) demonstration of recovery from blocking. In Blaisdell et al.'s critical experiment, rats received Phase 1 training consisting of 12 elemental (A+ or D+) conditioning trials. In Phase 2, compound training consisted of 4 AX+ trials for all subjects. Thus, X was either potentially blocked by A (when Phase 1 consisted of A+ trials) or overshadowed by A (when Phase 1 consisted of D+ trials). In Phase 3, rats received either 800 A-alone trials or a treatment designed to control for exposure to the experimental apparatus. Blaisdell et al. observed less responding when compound training was preceded by A+ trials (i.e., blocking) and, critically, they observed a recovery from blocking when Phase 2 training was followed by A-alone trials. In Kutlu and Schmajuk's simulations of Blaisdell et al., they used the same set of parameters that they had used from prior simulations of other data sets despite the parameter values being initially derived from eye blink conditioning data. Kutlu and Schmajuk's simulations involved 40 Phase 1 trials and 20 Phase 2 trials. Thus, both the absolute numbers of trials and the ratio of different trial types were distorted compared to the empirical work. To our knowledge, researchers have not yet attempted to replicate Blaisdell et al.'s recovery from blocking effect in the type of eye blink conditioning procedure that provided the basis for Kutlu and Schmajuk's choice of free parameters. Thus, it remains possible that the SLGK model erroneously predicts recovery from blocking in a procedure that does not actually replicate the effect as it was observed in a conditioned suppression procedure. Clearly, this is neither an *a priori* technique nor a rigorous method for comparing the SLGK model's predictions to the results of experiments. Moreover, nothing was said about how the number of trials in each phase was decided. This is especially problematic in cue competition situations because competition among cues often wanes with extensive compound training (e.g., Stout, Arcediano, Escobar, & Miller, 2003).

In addition to translating empirical procedures (such as number of trials and characterization of intertrial intervals) to simulations, modelers must select values for free parameters. Ideally, modelers would describe the predictions of a model across parameter space. For example, grid search algorithms involve repeating simulations across an equally spaced and large subset of parameter space. However, most contemporary associative models have too many free parameters to allow for a sufficiently fine-grained parsing of parameter space.

Thus, values for free parameters are usually either arbitrarily set (i.e., based on intuition) or selected based on optimization procedures. Intuitively selected parameters are usually initially used to assess the general functioning of the model. But with more mature models, parameters are almost always determined by some sort of procedure that optimizes the fit of the model to some data set. In practice, the method for selecting the free parameters is seldom reported (e.g., most of the models discussed in the 2012 *Learning & Behavior* Special Issue on Computational Models of Classical Conditioning). However, unbiased comparisons among models require that the free parameters of all models being compared be equally optimal across models. Most predictions are parameter dependent even in simple models (e.g., Van Hamme & Wasserman, 1994). When arbitrary values are used for the free parameters of a model, differences in the fits of the models to the data can be attributed either to differences across the models in the optimality of the parameters with respect to the data set in question or to the psychological principles reflected in the equations. Of course, it is the differences in psychological principles in which we are interested, but differences in optimality of parameters across models can seriously confound such comparisons. Parameter optimization (e.g., through hill climbing) using the precise data set in question reduces this problem because, in principle, the technique finds the best (or near best) set of parameters for each of the models; thus, there should be no systematic difference in the optimality of the parameters of the different models. Notably, the optimality of the parameters discovered by parameter optimization depends on the nature of the optimization problem and the optimization algorithm. Some optimization problems are susceptible to issues like local minima and plateaus, which make it less likely that the parameters discovered are optimal. However, generally if one model provides a poorer fit to data than another model given optimal parameters for both models, the simulations reveal something about the psychological constructs captured by the models.

In principle, one can use parameter optimization to conduct *post hoc* or *a priori* simulations. An *a priori* approach uses hill climbing (or other optimization techniques) to find the best-fitting parameters for data set 1 (the optimized data set). Those best-fitting parameters are then used to make *a priori* predictions concerning a second experiment (so-called *cross validation*). The predictions are then compared to the results of the second experiment (the target data set). The biggest problem with this strategy is that it, like arbitrary parameter selection, is unlikely to produce equally optimal parameters with respect to the target data set for each model being compared. Thus, if the parameters derived from fitting to the *optimized* data set fail to allow a model to explain the key results in the *target* data set, one cannot unarguably conclude that the model fails to explain the target data set or, more importantly, that the model cannot explain both the optimized and target data sets. That is, other parameters may better fit the target data set without appreciably impairing the fit to the optimized data set. Consequently, this strategy is useful only when claims are being assessed about the predictive potential of a model *given these specific parameters*. To be sure, cross-validation is essential in applied predictive modeling, where the aim is to develop a statistical or predictive model *and* a set of parameters that will generalize to new data. However, the value of this procedure in experimental modeling is limited because the predictive potential *of the model* is not exhaustively tested by using *a priori* best-fitting parameters. There is no reason to expect *a priori* that a single model would or would not be

selectively penalized by the use of cross-validation procedures. However, cross-validation procedures limit the flexibility of the model in fitting an empirical effect so a difference between models with respect to their performance in cross-validation tests does not mean that a model is unable to explain an effect. In our view, the theoretical basis of the model is best tested using *post hoc* parameters. Obviously post hoc parameter selection greatly improves the fit of each model, but not necessarily equally so across models. However, in so doing, models can be compared on a level playing field of parameters that are optimal for each model. In this case, differences in the goodness of fit of each model to the data set of interest are not biased by the specific values of free parameters. Thus, differences between models in goodness of fit can be assumed to arise uniquely from the models' positing different psychological constructs.

Post hoc parameter selection uses hill climbing to find the best-fitting parameters for the target data set rather than using parameters that were discovered by a hill climbing procedure applied to a different data set. The use of this former technique is ubiquitous in most areas of human cognitive science. In fact, some of the most important computational modeling work in category learning has involved *post hoc* parameter selection (e.g., Kruschke, 1992). This method provides several advantages relative to the *a priori* and arbitrary methods of parameter selection. The most meaningful simulation results are falsifications. That is, the scientific method does not allow for theories or models to be confirmed. Instead, scientific advancements are achieved through falsification of theories or models. In our view, the most compelling falsifications of a model occur when the values of free parameters in the model are presumably optimal for the target data set. A bad fit of a model to the data set (relative to other models) implies that it is impossible for the model to explain an experimental result through any subsequent improvement in parameters. Moreover, the use of optimal parameters facilitates comparisons between models. In addition, the *a priori* parameter selection method fundamentally relies on a between-experiment comparison, which assumes that the procedures used to produce the optimized data set are comparable to those used to produce the target data set, or at least that procedural differences have no impact on parameters. *Post hoc* parameter selection eliminates this assumption. In principle, there are two types of parameters, those that reflect psychological processes alone and those that hinge at least in part on specific procedures used in an experiment. However, at the current stage of development of models of learning, it is doubtful that even those parameters that in principle reflect purely psychological processes are truly independent of procedures. Obviously *post hoc* parameter selection will result in better fits for each model, but in so doing it will minimize bias across models that might otherwise exists.

4. Penalizing for Number of Parameters and Criteria for Selecting Models

Parameter selection techniques like hill climbing identify only the best-fitting parameters. Thus, these procedures are but a first step in comparing models. After the best-fitting parameters are identified, statistical comparisons between the fits of models are needed. The simplest statistic describing the fit of a model is SSE, which is the sum of the squared deviations between a model's numerical predictions and empirical observations, i.e.,

$$\mathrm{SSE}=\sum(\mathrm{Predicted}-\ \mathrm{Observed})^2 \quad (2)$$

SSE is often (although not always) the variable that is optimized in hill climbing algorithms. However, comparing SSE values is problematic when the models used to produce those values differ in their numbers of free parameters. All other things being equal, more free parameters allows models to achieve a better fit to data. Thus, statistical comparisons between models need to quantify the fit of each model's predictions to the target empirical observations taking into consideration the number of free parameters that each model uses to fit those observations. The Bayesian Information Criterion (BIC) is a statistic that can be used to compare models with different numbers of free parameters. The BIC assumes that:

$$\mathrm{BIC}=n*log\,(\mathrm{SSE}/n)+P*log\,(n) \quad (3)$$

where n represents the number of data points that were simulated and P is the number of free parameters used to fit the data (Waekliem, 2004). Effectively, $n*log\,(\mathrm{SSE}/n)$ represents a badness-of-fit measurement and the far-right term ($P*log\,[n]$) represents a penalty for model complexity. The Akaike Information Criterion (AIC) is similar to the BIC, but with slightly different statistical assumptions. The AIC assumes that

$$\mathrm{AIC}=n*log\,(\mathrm{SSE}/n)+2P \quad (4)$$

BIC and AIC statistics are useful because they penalize models for each free parameter, thereby compensating for the better fit that comes from using more free parameters independent of any relationship to underlying psychological processes. For example, in a Fourier analysis a complex waveform can be better fit using more component frequencies, all other things being equal. Similar principles apply to the number of free parameters in a model. BIC and AIC statistics often agree on the ordinal rankings of models. The similarities between equations 3 and 4 might cause one to conclude that BIC and AIC are interchangeable. However, BIC and AIC were developed to select models with different attributes in idealized situations (e.g., based on Monte Carlo simulations).

The AIC statistic emerges from statistics that summarize the distance between a target model and the true model (i.e., the psychological process as it exists in nature). Unfortunately, the true model is usually unknown in psychological research; thus, it is impossible to compute the distance between a target model and the true model. The AIC statistic is an unbiased estimator of the distance between a candidate model and an unknown true model that permits comparisons between models. The candidate model that produces the lowest AIC value corresponds to the best model in the set of models being compared. Importantly, the AIC statistic cannot indicate that a model is good in an absolute sense – it can only be used in model comparisons. The AIC statistic assumes that $2P$ is an unbiased estimator of the loss of information incurred by adding free parameters. Unfortunately, this

assumption is true only in select situations. For example, when P is large relative to n, $2P$ is systematically less than the actual loss of information and researchers should use the corrected AIC (i.e., AIC_C), which converges with simple AIC for models with small numbers of parameters and large numbers of observations but otherwise applies a greater penalty for model complexity. BIC is guaranteed to select the true model when it is included in the set of candidate models as $n \rightarrow \infty$. In fact, BIC has the peculiar property of applying a greater penalty for free parameters with increases in the number of observations [i.e., P * log(n)]. A detailed discussion of the differences between BIC and AIC is beyond the scope of this paper (see Vriez, 2012, for a discussion).

In situations in which there is disagreement between the BIC and AIC, it is best to report both statistics. Both are measures of fit for each model that 'correct' for differences across models in the number of free parameters (albeit through different functions) and hence prevent overfitting by adding free parameters. Lower values of BIC and AIC indicate a better score for the model. While it is possible to compare models based solely on their BIC/AIC scores, in our view it is best to use the BIC/AIC statistics to supplement an analysis of a model's ordinal predictions. For example, in a data set consisting of conventional Pavlovian acquisition, blocking, overshadowing, and backward blocking effects, one would observe only a small difference between the BIC/AIC values produced by the Rescorla-Wagner (1972) model and Van Hamme and Wasserman's (1994) model of retrospective revaluation. However, close inspection of the ordinal differences in these models' predictions would reveal that only the Van Hamme and Wasserman model succeeds in explaining the ordinal relationship constituting backward blocking (e.g., Shanks, 1985). Thus, the BIC/AIC statistics should be used in conjunction with an evaluation of the ordinal differences in the models' predictions. In addition to such qualitative comparisons between predictions and observations, it is possible to quantitatively compare fits based on AIC scores. There are several ways to accomplish this (Burnham & Anderson, 2002), but the simplest statistic for comparing AIC values is ΔAIC, which is the difference between the AIC score of a model (i) and the lowest AIC score in a set of candidate models. Thus, $\Delta AIC_i = AIC_i - AIC_{min}$. Higher values of ΔAIC indicate less support for a model (see Cheung, Neisewander, & Sanabria, 2012, for an application).

For meaningful fits of each model to data, there obviously needs to be a basis for expecting that a single set of parameters is applicable across the data set in question. For example, all of the data should have been collected using the same measurements, stimuli, apparatus, and species. For the same reason that it is inappropriate to compare data points from different papers in an ANOVA, it is potentially misleading to pool data points from different papers in calculating SSE for a hill climbing procedure. We take the middle-ground and assume that most experiments within a paper (provided they are conducted in the same preparation) are sufficiently comparable to be included in the same hill climbing exploration.

Research in associative learning frequently tests the predictions of theories. Researchers often interpret the results of an experiment as falsifying one or more models and failing to falsify a different set of models. In actuality, a model can only explain or fail to explain the results of an experiment better than a different model. In principle, there are four possible outcomes of model testing: (1) a researcher *claims that a model fails* to explain the data

better than or equal to other models and the *model actually fails* to explain the data better than or equal to other models, (2) a researcher *claims that a model explains* the data better than or equal to other models and the *model actually explains* the data better than or equal to other models, (3) a researcher *claims that a model fails* to explain the data better than or equal to other models but the *model actually explains* the data better than or equal to other models (Type 1 error), and (4) a researcher *claims that a model explains* the data better than or equal to other models but the *model actually fails* to explain the data better than or equal to other models (Type 2 error). Thus, model testing, like null hypothesis statistical testing, creates the opportunity for two types of errors. Type 1 errors are reduced (relative to research that omits simulations or uses a fixed set of parameters) by using hill climbing to allow a model optimal flexibility in explaining the data. Type 2 errors are reduced by basing comparisons on a statistic (e.g., BIC or AIC) that penalizes models for using more parameters, which could result in the rejection of an unduly complex model in favor of simpler models. In our view, Type 1 errors in model testing (3; rejecting a model when the model actually explains the data) are more egregious than Type 2 errors because revisions of models and, more generally, scientific progress depends critically on the falsification of models. Actual simulations of models are seldom reported in conventional empirical papers that were aimed at testing theories. Instead, researchers often intuit their way through the predictions of a model and relate them to empirical data. However, in our experience, this technique frequently underestimates the flexibility of the existing computational models, thereby increasing the likelihood of a Type 1 error. In the following exemplar simulations, we apply these procedures to data concerned with retrospective revaluation.

5. An application to retrospective revaluation

After training a target cue (X) in compound with a companion cue (A), associative inflation or deflation of A often results in changes in the response potential of the absent target cue (Kaufman & Bolles, 1981; for a review, see Miller & Witnauer, 2016). These retrospective revaluation effects are problematic for the original Rescorla-Wagner (1972) model (RWM), which does not allow for learning about absent cues. However, Wasserman and his collaborators (Van Hamme & Wasserman, 1994; Wasserman & Castro, 2005) proposed that retrospective revaluation could be explained by the RWM if it is modified so that the eligibility (aka associability) of an absent cue or CS to enter into association with the outcome is negative. In this framework, learning about an absent cue is based on mechanisms similar to those involved in learning about an omitted outcome when it is expected (e.g., extinction). Perhaps the simplest instantiation of this view is a model that uses separate values for the associabilities of present cues and absent cues (e.g., Danks, 2003). Despite the simplicity of this model, previous simulations found a better fit to data using a slightly more complex set of rules for modeling the eligibility of absent cues to change their associative values (Witnauer & Miller, 2011). Specifically, a model that weights the associability of an absent cue by the degree to which that cue is activated through within-compound associations provided a better fit to retrospective revaluation effects than a model that represents the associability of absent cues by a constant (negative) fraction of the associability of a present cue. The present simulations further explored this issue.

5.1 Models

The defining equations of the models that were compared in the present simulations are listed in Table 1. We selected three relatively simple models of cue interaction and used them to simulate experiments on retrospective revaluation. The models were good candidates for comparison because they have similar numbers of free parameters and identical assumptions about stimulus representation (elemental), time (trial-wise), and the relationship between associative strength and performance. These models differ only with respect to their treatment of absent cues. The RWM assumes null learning about absent cues. The Within-Compound Model [WCM] and Conjoint Error Model [CEM] are different implementations of the Van Hamme and Wasserman (1994) model in that they assume that absent cues can be eligible to undergo associative changes. In principle, one could assume that the eligibility of an absent cue is a constant negative fraction of the cue's associability when it is physically presented. Indeed, Wasserman and his colleagues (e.g., Wasserman & Castro, 2005) championed the idea that separate values for α be used on cue present ($0 < \alpha_{present} < 1$) and cue absent ($-1 < \alpha_{absent} < 0$) trials. However, this model fails to explain the relationship between the strength of the within-compound association between the absent target and revalued cues and the degree of retrospective revaluation (Witnauer & Miller, 2011). Two alternative models assume that (1) within-compound associations might contribute to retrospective revaluation by controlling the degree to which the representation of an absent cue is activated by previous companions or (2) within-compound associations might increase cue expectancy and, consequently, surprisingness of an omitted cue. In the present simulations, the WCM assumes that the associability of an absent cue is negatively related to its degree of activation through within-compound associations (see Witnauer & Miller, 2011). In contrast, the CEM assumes that learning about an absent cue occurs when the omission of the cue was unexpected. According to this view, associative changes are driven by the extent to which the US (or outcome) and the CS (or cue) are both unexpected. The CEM explains retrospective revaluation by assuming that the association between an absent target stimulus and the outcome changes because animals expect (but do not receive) the target stimulus based on the delivery of cues that were previously trained in compound with the target. For example, after an AX+ overshadowing treatment, associative deflation of A results in greater behavioral control by X than does deflation of an irrelevant control stimulus (recovery from overshadowing; e.g., Kaufman & Bolles, 1981; Matzel, Schachtman, & Miller, 1985).

Applied to recovery from overshadowing, all of the models in Table 1 assume that Phase 1 AX+ trials establish X-Outcome, A-Outcome, X-A, and A-X associations. In Phase 2, A- trials produce unexpected omissions of both the outcome and X, and decrease in the strengths of the A-Outcome and A-X associations. The RWM assumes that both the A-X association and the unexpected omission of X are inconsequential with respect to both the X-Outcome association and subsequent behavioral control by X. The WCM assumes that activation of X through the A-X within-compound association results in the associability of X being negative and that the unexpected omission of the outcome results in strengthening of the X-Outcome association. Importantly, both the RWM and WCM assume that antecedent events are processed differently from outcomes. The CEM model rejects the view that antecedent events (i.e., cues) are fundamentally different from subsequent events (i.e.,

outcomes) with respect to learning and information processing. Instead, it assumes that learning is determined by the degree to which the antecedent and the outcome are contiguous and conjointly unexpected. This is not a large departure from other acquisition-focused models of retrospective revaluation (e.g., Witnauer & Miller, 2011) because the degree of CS expectation is based exclusively on within-compound associations. Indeed, the CEM captures the relatively old idea that both CSs and USs compete for a limited resource (e.g., Wagner, 1981; Dickinson & Burke, 1996), which we instantiated in a trial-wise, elemental model. We anticipated that the CEM would be able to explain several cue interaction phenomena that elude both the RWM and the WCM model.

The hill climbing algorithm used in the present simulations (the interior-point fmincon algorithm in the Matlab Optimization Toolbox) minimize the sum of the squared differences between the predictions of a model and the observed group means in an experiment. Thus, the predictions and observations need to be on the same scale. In principle, there are many ways to scale either the models' predictions or the observations so that the sum of squared error can be computed. The results of the simulations reported here assumed that the scaled predictions by the model are proportionally related to responding ($V_{X\text{-}O}$ when only X is tested), thus the following scaling equation was used:

$$\text{Prediction} = V_{X\text{-}O} * \text{Scaling} \quad (5)$$

where Scaling is a free parameter that could take any nonnegative value. In simulations of lick suppression experiments, $\log_{10}$ of 5 was added to $V_{X\text{-}O}$ * Scaling because the lick suppression experiments that were simulated measured the amount of time required by animals to complete 5 cumulative seconds of drinking in the presence of a fear CS. Thus, the minimum of the scale was $\log_{10}$ of 5. Notably, more sophisticated scaling functions exist, especially in the interval timing and choice behavior literatures (e.g., Brackney & Sanabria, 2015; Daniels & Sanabria, 2017; Kruschke, 2001).

5.2 Simulated experiments

We selected experiments to simulate that replicated basic retrospective revaluation effects or were claimed to challenge the revised Rescorla-Wagner model (e.g., Van Hamme & Wasserman, 1994). We structured the simulation procedures after the empirical procedures as closely as possible. For example, the models received the same number of trials as animals in the actual experiments. Moreover, the counterbalancing schemes used in empirical procedures were simulated by forcing counterbalanced cues to be represented by the same values for salience. The context was included as a cue in Pavlovian procedures that included a context change and one context extinction trial was interposed between each pair of trials except during extensive posttraining extinction treatments (e.g., McConnell et al., 2009). When multiple test cues were assessed in counterbalanced order, we disabled learning during the test phase so that test-order effects would not influence the predictions of the models. Of course, predicted test-order effects could be compared to empirical data. Unfortunately, the series of experiments that we simulated did not report the results from

separate counterbalanced test orders. Thus, it is impossible to compare the models' predictions to data in our specific application.

5.2.1. Simulation 1: Wasserman and Castro (2005)—Retrospective revaluation seems to depend on retrieval of information about absent cues during posttraining inflation or deflation of a companion cue. To test this view, Wasserman and Castro (2005; Table 2) conducted an experiment in which participants experienced recovery from overshadowing and backward blocking procedures, and were either informed or uninformed about the absence of the target cue during posttraining inflation and deflation. Their results revealed that retrospective revaluation is stronger when subjects are explicitly informed that the target cue is absent. In Simulation 1, we fit the RWM, WCM, and CEM to their results. Importantly, we assumed that the free parameter controlling the eligibility of an absent cue to undergo associative changes ($k2$) would be represented by different values between Group Informed and Group Uninformed. Inspection of the predictions by these models (Table 2) suggests that both the WCM and CEM are able to explain this result. Moreover, both the WCM and CEM provided best fitting predictions with values for $k2$ that were greater in Group Informed than in Group Uninformed. Notably, the CEM provided a better fit to the data than the WCM (Table 3), although both models were able to explain the critical ordinal differences in ratings observed by Wasserman and Castro.

5.2.2. Simulation 2: De Houwer & Beckers (2002)—The purpose of Simulation 2 was to compare the fits of the WCM and the CEM to data concerning first- and second-order recovery from overshadowing and backward blocking effects. In De Houwer and Beckers' experiment, participants in a human contingency learning procedure received pairings of weapon cues with explosion outcomes (see Table 4). All participants received AB+ trials in Phase 1 and BC+ trials in Phase 2. After Phase 2, participants rated the effectiveness of weapon cue elements A, B, and C and two filler cues. In Phase 3, participants received either A+ trials or A- trials. After Phase 3, all elements were retested. A+ trials decreased participants' ratings of the effectiveness of weapon B relative to the test of B that occurred after Phase 2; thus, A+ trials resulted in first-order backward blocking of B. Similarly, A- trials increased behavioral control by B relative to the previous test. Importantly, the opposite pattern of responding was observed in tests of cue C. A+ trials increased ratings of C's effectiveness and A- trials decreased ratings of C's effectiveness relative to the test after Phase 2. In Simulation 2, we expected the WCM to explain both the first-order retrospective revaluation effects (i.e., changes in ratings to B) and the second-order retrospective revaluation effects (i.e., changes in ratings to C). In previous simulations (Witnauer & Miller, 2011), the WCM was able to fit other instances of second-order retrospective revaluation by assuming that BC+ trials establish an inhibitory association between A and C. Thus, the WCM predicts a positive value for the associability of C based on the product of negative values of both $\Sigma V_{i\text{-A}}$ and $k2$ during A+ or A- trials.

Inspection of the predictions by the models reveals that both the CEM and WCM predicted the overall pattern of results observed by De Houwer and Beckers (2002). Specifically, both models predicted that A+ (A-) trials should decrease (increase) ratings of target cue B and increase (decrease) ratings of target cue C. However, neither model provided a good fit to all

of De Houwer and Beckers' results. Specifically, both models failed to predict the nearly equivalent ratings of A, B, and C that were observed after Phase 2. Predicted ratings to B were greater than predicted ratings of C or A because B was trained in both Phase 1 and Phase 2. Moreover, predicted ratings of A were greater than predicted ratings of C because B should have blocked learning about C and B should have only overshadowed learning about A. The SSE, BIC, and AIC values in Table 3 indicate that the WCM provided a better statistical fit to the data than the CEM based mostly on the fits of the models to data from the first test session. The WCM (SSE = 3375.40) and the CEM (SSE = 3963.00) provide similar fits to the data from second test of A, B, and C.

5.2.3. Simulation 3: Grahame et al. (1994; Experiments 1 and 3)—Latent inhibition (LI) occurs when repeated nonreinforced pre-training exposure to a CS reduces subsequently trained behavioral control by the CS (Lubow & Moore, 1959). Several associative theories assume that LI is caused by acquisition of either a CS-context (e.g., Miller & Matzel, 1988) or context-CS (Wagner, 1981) association during nonreinforced presentations of the CS in the training context. In Grahame et al.'s (1994) Experiment 1, extinction of the training context after CS-US pairings reduced the strength of the LI effect. In a lick suppression procedure, groups of rats received either 120 nonreinforced presentations of the CS (X-) or a control procedure in the context of the subsequent CS-US pairings (X+). Orthogonally, subjects received extensive extinction of the context or a handling control after the CS-US pairings. Behavioral control by the CS was weaker in the group that received LI treatment and no context extinction than in the group that received control treatments before and after CS-US pairings. Importantly, the LI effect was attenuated when context extinction was delivered after CS-US pairings. Moreover, in Experiment 3, extinction of the CS-context association by presentation of the CS in a context distinctly different from that of preexposures prior to CS-US pairings decreased the strength of the LI effect.

Grahame et al.'s (1994) Experiments 1 and 3 and the results of Simulation 3 are summarized in Table 5. Unsurprisingly, the RWM provided a poor fit to Grahame et al.'s observations; the model explained only the difference in suppression to the target CS between Test 1 and Test 2 caused by extinction of the target CS on the first test trial. Both the WCM and CEM assume that CS-context associations can influence learning about a target CS; thus, we were centrally interested in whether Simulation 3 would reveal any difference between these two models. Inspection of Table 5 reveals that the CEM but not the WCM was able to explain the central results of Grahame et al.'s experiments, including the basic LI effect (lower predicted responding to X in the group that received preexposure to X), the increased behavioral control by X after extinction of training context extinction in the LI condition, and the increased behavioral control by X observed after extinction of the X-context association caused by presentations of X in a new context after preexposure to X. Thus, the CEM provided a better fit to the empirical data than the WCM and RWM (Table 3).

5.2.4. Simulation 4: McConnell et al. (2010; Experiments 2 and 3)—Simulation 2 revealed that the WCM and CEM explain second-order retrospective revaluation in human contingency learning (see also Witnauer & Miller, 2011). The purpose of Simulation 4 was

to replicate this with a different procedure (lick suppression). Thus, we fit the RWM, WCM, and CEM to the results of McConnell et al.'s (2010) Experiments 2 and 3, which are summarized in Table 6. McConnell et al. replicated the observation that extinction of a target cue (X) in compound with a conditioned inhibitor (B) results in less behavioral control by X at test than if B is a neutral stimulus. Moreover, in Experiment 3, they demonstrated that deflation of the training excitor (A) previously used to train conditioned inhibition of B after extinction X results in a reduction in the protection from extinction effect relative to a group that received deflation of an irrelevant excitor (D). Notably, McConnell et al. embedded their experiment in a sensory preconditioning procedure; thus, the target cue and the other treatment cues were not directly paired with the US. Instead, target and treatment cues were paired or unpaired with a surrogate outcome (represented by "+" in Table 6), which was later paired with a footshock US. Because none of the models simulated are able to explain sensory preconditioning, the strength of the predicted response was assumed to be proportional to the strength of the associations between the cue and the surrogate outcome rather than the association between the cue and the US. The best-fitting predictions of the RWM, WCM, and CEM are reported in Table 6. Both the WCM and the CEM were able to explain McConnell et al.'s central results (extinction, protection from extinction, and reduced protection after deflation of the training excitor). The results summarized in Table 3 confirm this impression. Specifically, both the WCM and, to a lesser extent, the CEM provided better statistical fits to the data than the RWM both with (SSE) and with (BIC and AIC) applying a statistical penalty for additional free parameters.

5.2.5. Simulation 5: McConnell et al., (2009)—The LI effect, like the extinction effect, can be attenuated when preexposure to X occurs in compound with a conditioned inhibitor prior to X+ trials. McConnell et al. (2009) identified the mechanisms underlying this effect by conducting the lick suppression experiments summarized in Table 7. In addition to observing both the basic LI effect and protection from LI effect (Experiment 1), McConnell et al. (Experiment 2) observed a reduction in the strength of the protection effect as a result of associative deflation of the conditioned inhibitor's training excitor. In their critical groups, subjects received in Phase 1 A+ / AB- trials designed to establish B as a conditioned inhibitor. In Phase 2, subjects received 48 XB- trials, which resulted in stronger suppression to X than was observed after 48 X-. In Phase 3, subjects received extensive extinction of either A (the training excitor) or C (an irrelevant control CS). Extensive extinction of A reduced the protection from LI effect (i.e., increased suppression) relative to extensive extinction of C. Inspection of Table 7 reveals that the models differed widely with respect to their fits to the data. The RWM explained only the increased behavioral control by X observed after X was pretrained in the presence of a conditioned inhibitor. The mechanism underlying this prediction is similar to the mechanism that the model uses to predict superconditioning. The WCM predicted the effect of extinction of A on behavioral control by X; however, it failed to predict the basic LI effect. Lastly, the CEM explained almost all aspects of McConnell et al.'s data. In fact, the largest difference between McConnell et al.'s (2009) observations and the CEM's predictions was in the group that received preexposures to the target stimulus in compound with a neutral stimulus. The CEM model predicted a larger protection effect by a neutral stimulus than was observed by McConnell et al. Based on the CEM being able to explain both protection from LI effect and the reduced protection

from LI achieved by extinction of A, the CEM provided a better fit to the data than the WCM and RWM.

5.3. Discussion of simulations

These simulations demonstrate that basic retrospective revaluation can be well explained by models that assume some role for within-compound associations in learning about absent cues. That is, slightly modified versions of RWM explain several phenomena (e.g., Lubow & Moore, 1951; Kaufman & Bolles, 1981) that are widely cited as incompatible with the RWM. Within the RWM framework, retrospective revaluation might occur because a treatment cue (e.g., an overshadowing cue in a recovery from overshadowing procedure) activates a representation of the absent target stimulus through a within-compound association during posttraining associative inflation or deflation (WCM). Alternatively, retrospective revaluation might be caused by the target cue being unexpectedly (based on within-compound associations) omitted during posttraining associative deflation or inflation (CEM). In Simulations 2 and 4, both the CEM and WCM were able to predict all of the critical differences between experimental conditions, although the WCM provided a slightly better fit with respect to SSE and BIC/AIC. However, Simulations 1, 3, and 5 revealed that the WCM explanation of retrospective revaluation is either incomplete or, more likely, wrong. The CEM seems to be a better model than the WCM, at least as applied to retrospective revaluation, because the CEM's mechanism for explaining retrospective revaluation allows the model to also explain latent inhibition. That is, without any additional free parameters beyond those of the WCM, the CEM is able to explain both retrospective revaluation (including first- and second-order) and latent inhibition. More importantly, the higher-order retrospective revaluation effects observed by De Houwer and Beckers (2002) and McConnell et al. (2009; 2010) can be explained by variants of the Van Hamme & Wasserman (1994) model (CEM and WCM), even though the associative literature universally claimed that the Van Hamme and Wasserman model could not explain higher-order retrospective revaluation. Thus, the present simulations highlight the importance of using actual simulations to support theoretical claims.

6. Limitations

This paper does not answer all questions concerning how to weigh the merits of different theories. Our central point concerns the selection of parameters when contrasting competing theories with respect to a specific data set. The result of such a comparison is not to completely reject the less adequate theory, but simply to acknowledge that it fares less well with the specific data set. We argue that *post hoc* selection of optimal parameters for each model levels the playing field by circumventing differences in goodness of fit that arise from differences across models in the appropriateness of free parameters set *a priori*. This is highly likely when the a priori parameters of the different models were initially selected based on different data sets for each model in question. Obviously, the use of *post hoc* parameters is a limitation because the predictions derived by simulation are more likely to be parameter dependent. However, this is not undesirable if those predictions are confirmed empirically and the procedural parameters that determine whether one or another effect is observed correspond to appropriate (or at least testable) model parameters. Additionally, this

does not mean that the model per se (the computational link between theory and data) is *post hoc*. For example, using best-fitting parameters in simulations of the Rescorla-Wagner (1972) model does not change the fact that the model is over forty years old and its formulation preceded the initial observation of many phenomena to which it is applied today. Thus, there is a very strong *a priori* component retained even when one uses the hill climbing procedure that we recommend when different models are being compared.

In the present simulations, we used a hill climbing algorithm (Matlab's fmincon algorithm) that minimized SSE; specifically, we used the least-squared error (LSE) method of parameter estimation. Importantly, other methods for parameter estimation might be a better match for some simulations (see Myung, 2003, for a comparison). LSE assumes that prediction errors are random, normally distributed, and independent (Burnham & Anderson, 2001), which is difficult to confirm. In contrast, MLE and Bayesian methods are less restrictive in their assumptions. Moreover, they reveal the likelihood of a combination of parameter (given a model and data), which is more informative than the SSE value uncovered by the LSE method. Bayesian methods assume that parameter estimation presents an inverse projection problem wherein the method aims to identify the probability of the model (M), given the data (D). Bayes theorem dictates that the $p(M|D) = [p(D|M) * p(M)] / p(D)$. Of course, this method requires that assumptions be made about the prior distribution of parameters (Lee, 2008), which formalizes $p(M)$. However, to our knowledge these prior distributions are unknown for the models used in the present simulations. One could make a relatively atheoretical assumption about the prior distribution (e.g., uniformity); however, other distributions seem reasonable (e.g., log normal; Cheung et al., 2012). Indeed, this issue is not trivial because model evaluation can depend critically on specific assumptions about the prior distribution (e.g., Vanpaemel, 2010), although such a simulation result would cast doubt on the predictive value of the model in question. That is, ideally a model's predictions are relatively invariant with respect to the prior. While interesting and certainly worthy of investigation, this issue is outside of the scope of the present article. In procedures where the probability density function of observations can be easily specified (e.g., in fitting binomially distributed response data) and where the target models make predictions about the distributions of responses within a condition, the maximum likelihood estimation (MLE) method is better than the LSE method (e.g., Myung, 2003). Both the Bayesian and MLE methods are better than LSE because they provide an estimate of the most likely values of parameters given the data and not just the values of parameters that result in the lowest value for SSE. Indeed, one can appreciate the advantages of this approach by considering the results of the present simulations. Inspection of Table 3 indicates that best-fitting values of some free parameters were equal to unity in some simulations. Of course, this does not mean that those parameters are the most likely given the data. Instead, one can only conclude that those values provided the best fit with respect to SSE. Similarly, in Simulation 1 Groups Informed and Uninformed were simulated using different values for $k1$. The difference in best-fitting values for $k1$ was consistent with the psychological intuition that informing subjects about an absent cue should increase the cue's eligibility during retrospective revaluation. However, the LSE method is not well-suited to reveal information about the size of that difference because the method does not inform concerning the likelihoods of those values. Instead, LSE is useful only in comparing the fits of models to group means when the

above mentioned assumptions are satisfied. In addition, Bayesian methods provide the distinct advantage of being able to replace null hypothesis significance tests for determining the effect of an experimental treatment (e.g., informing subjects about an absent cue) on a model's free parameters (e.g., Kruschke, 2011).

A second potential criticism of this approach is that the method of controlling for number of free parameters (i.e., using the BIC/AIC) does not fully control for model complexity as a determinant model fit. This is potentially problematic as, all other things equal, some models are decidedly more complex than others. The BIC/AIC's penalty for more free parameters assumes that all free parameters are equally important in determining the fit of model to data, whereas one could add free parameters to a model that do not appreciably improve the model's fit. Importantly, *post hoc* parameter selection increases the flexibility of models, reducing the likelihood of falsification. However, we view this as a more conservative strategy than *a priori* parameter selection because *a priori* selection lends itself to erroneous falsification. Importantly, most of the papers in the 2012 *Learning & Behavior* Special Issue on Computational Models of Conditioning focused on the successes of models, instead of failures of models or differences in the failures of different models. For example, Kutlu and Schmajuk (2012) evaluated their model based on irreversible successes, which are those successful predictions that are relatively parameter independent (i.e., the model makes those predictions across a wide range of parameter values). Across 87 different cases, their model predicted effects or correlations that were similar to 94% of the observed effects or correlations that were simulated. However, their use of a fixed set of parameters causes us to wonder whether the model might explain 100% of the data with a better set of parameters. Of course, a model that explains 100% of the data might fail to make falsifiable predictions; thus, good fits to data are useful only in the context of comparisons between falsifiable models. More importantly, we can only wonder whether their model is more or less successful than competing associative models. Successfully explaining experimental results is, in our view, less important than comparing models.

Not all data sets have equal potential to differentiate among models. Surely the relative success or (more importantly) failure of one model compared to another model will depend in part on the selection of the empirical data to which the models are fit. There are several approaches to this problem, all of which are compatible with our proposed strategy of model selection and treatment of free parameters. First, theoretical papers often seek to apply a model to a wide range of previously published data sets (usually group means), which are often selected because they challenge some other model or category of models (e.g., see the 2012 special issue of *Learning & Behavior*). Second, empirical papers sometimes include simulations of the new data. We see value in both of these strategies. Fitting previously published group means, as was done in the present example, ignores individual differences in behavior and leads to potentially misleading conclusions. The often-cited graded learning curve is a prominent example of exactly this in the animal learning literature. Specifically, while averaged data often reveal gradual changes in behavior across trials, inspection of the response patterns of individual subjects often reveals discontinuities in behavior (e.g., Donner & Hardy, 2015; Friedman, Massaro, Kitzis, & Cohen, 1995; Gallistel, Fairhurst, & Balsam, 2004). Thus, ideally modelers would fit models to the behavior of individuals rather than fitting models to group means. Model comparison, in conjunction with *post hoc*

parameter selection, can be constructively used in both situations. In principle, computational methods can be used to support any comparison between theories. But it is not obvious that doing so will be illuminating when the models being compared differ radically in intent; for example, McLaren and Mackintosh's (2000) model addresses perception and attention, whereas Stout and Miller's (2007) SOCR model addresses retrieval and response generation. However, researchers in other areas of cognitive psychology (e.g., category learning) have sometimes constructively compared computational implementations of highly dissimilar models (e.g., Johansen & Kruschke, 2005). Moreover, there are some demonstrations that under select conditions comparisons of highly dissimilar associative models can be informative (e.g., Denton & Kruschke, 2006; Perales & Shanks, 2003).

Post hoc parameter selection and computational methods in principle could be used to compare any two or more models. But not all comparisons will prove equally meaningful or informative. In general, models being compared should be well matched in their simplifying assumptions and ideally different in only one of their target theoretical assumptions. In practice, response rules are an area where models differ widely with respect to simplifying assumptions. For example, the Rescorla-Wagner (1972) model assumes a monotonic mapping of X-US associative strength onto the response potential of X. In contrast, Pineno's (2007) revision of the Rescorla-Wagner model makes more sophisticated assumptions about the link between learning and performance. Hence, computational comparison of these two models is not apt to be illuminating. In our view, the inappropriateness of computational comparisons between models mismatched in simplifying assumptions is a failure in choice of models to be compared, rather than a weakness in our suggested procedures for comparing models.

As with selection of data sets, selecting which quantitative models to compare often follows the rules of thumb commonly used by researchers when they compare theories in empirical papers. That is, models are selected that were designed to address the variable of central interest, such as 'attention.' However, some investigators have suggested more formal guidelines for identifying models that should be compared. For example, Broadbent (1958) proposed an adaptive procedure for identifying models. According to this view, model comparisons should initially focus on simple models that represent widely different classes of models, whereas subsequent comparisons should tend toward progressively more complex models. Our approach is compatible with this approach (albeit silent on this particular issue) in that we are concerned not with which models to compare, but how to compare them.

One can conceive of situations in which *post hoc* parameter selection is inappropriate. Specifically, if consistent parameter values are well-established through multiple parametric experiments across a broad range of preparations and subsequent fitting procedures, then those parameter values might be used to predict (*a priori*) the results of subsequent experiments. Moreover, as we previously noted, if the models in question use parameters that have been determined from previously established theoretical principles and empirical findings external to the model in question (e.g., the Hodgkin & Huxley, 1952, model), then the *post hoc* optimizing of parameters that we suggest would be inappropriate. However, in our view these situations seldom apply to research in associative learning. More commonly,

post hoc parameter selection improves the extent to which models are matched in the appropriateness of their parameter values when fitting them to a common data set.

7. Conclusions

In summary, simulations involving comparisons between models are inherently more illuminating than simulations of a single model. Moreover, some simulation procedures are distinctly better than others in permitting researchers to compare associative models and complement our use of traditional inferential statistics (Rodgers, 2010). These procedures allow us to quantify and compare degrees of fit of different models to specific data sets, and to arrive at a more rigorous connection between data and theory. Model selection requires that the models be matched with respect to both their simplifying assumptions and the optimality of their free parameters, which can be achieved through optimization techniques such as hill climbing. Statistics like BIC/AIC facilitate comparisons among models with different numbers of free parameters. We acknowledge that there are limitations to the inferences that can be supported by this approach but, like other scientific methods, it is informative when applied with caution. Moreover, this approach to selecting parameters seems to permit less biased comparisons among models.

Acknowledgments

Preparation of this manuscript was supported by NIH grant MH 033881. We are grateful to Robert Perez, Cody Polack, and Julia Soares for comments on a preliminary version of the paper.

References

Azorlosa JL, Cicala GA. Blocking of conditioned suppression with 1 or 10 compound trials. Animal Learning & Behavior. 1986; 14:163–167.

Blaisdell AP, Gunther LM, Miller RR. Recovery from blocking achieved by extinguishing the blocking CS. Animal Learning & Behavior. 1999; 27:63–76.

Broadbent, DE. Perception and communication. Elmsford, NY, US: Pergamon Press; 1958.

Brackney RJ, Sanabria F. The distribution of response bout lengths and its sensitivity to differential reinforcement. Journal of the Experimental Analysis of Behavior. 2015; 104:167–185. [PubMed: 26377437]

Burnham, KP., Anderson, DR. Model selection and multimodel inference: A practical information-theoretic approach. 2. Secaucus, NJ: Springer; 2002.

Cheung TH, Neisewander JL, Sanabria F. Extinction under a behavioral microscope: Isolating the sources of operant response rate. Behavioural Processes. 2012; 90:111–123. [PubMed: 22425782]

Daniels CW, Sanabria F. Interval timing under a microscope: Dissociating motivational and timing processes in fixed-interval performance. Learning & Behavior. 2017; 45:29–48. [PubMed: 27443193]

Donner Y, Hardy JL. Piecewise power laws in individual learning curves. Psychonomic Bulletin & Review. 2015; 22:1308–1319. [PubMed: 25711183]

Danks D. Equilibria of the Rescorla-Wagner model. Journal of Mathematical Psychology. 2003; 47:109–121.

De Houwer J, Beckers T. Second-order backward blocking and unovershadowing in human causal learning. Experimental Psychology. 2002; 49:27–33. [PubMed: 11975146]

Denton SE, Kruschke JK. Attention and salience in associative blocking. Learning & Behavior. 2006; 34:285–304. [PubMed: 17089596]

Dickinson A, Burke J. Within-compound associations mediate the retrospective revaluation of causality judgements. Quarterly Journal of Experimental Psychology. 1996; 49B:60–80.

Friedman D, Massaro DW, Kitzis SN, Cohen MM. A comparison of learning models. Journal of Mathematical Psychology. 1995; 39:164–178.

Gallistel CR, Fairhurst S, Balsam P. The learning curve: implications of a quantitative analysis. Proceedings of the National Academy of Sciences. 2004; 101:13124–13131.

Grahame NJ, Barnet RC, Gunther LM, Miller RR. Latent inhibition as a performance deficit resulting from CS-context associations. Animal Learning & Behavior. 1994; 22:395–408.

Harris JA. Elemental representations of stimuli in associative learning. Psychological Review. 2006; 113:584–605. [PubMed: 16802882]

Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in the nerve. Journal of Physiology. 1952; 117:500–544. [PubMed: 12991237]

Johansen MK, Kruschke JK. Category Representation for Classification and Feature Inference. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31:1433–1458.

Kamin, LJ. “Attention-like” processes in classical conditioning. In: Jones, MR., editor. Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation. Miami, FL: University of Miami Press; 1968. p. 9-31.

Kaufman MA, Bolles RC. A nonassociative aspect of overshadowing. Bulletin of the Psychonomic Society. 1981; 18:318–320.

Kruschke JK. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review. 1992; 99:22–44. [PubMed: 1546117]

Kruschke JK. Toward a unified model of attention in associative learning. Journal of Mathematical Psychology. 2001; 45:812–863.

Kruschke JK. Bayesian assessment of null values via parameter estimation and model comparison. Perspectives on Psychological Science. 2011; 6:299–312. [PubMed: 26168520]

Kutlu MG, Schmajuk NA. Solving Pavlov’s puzzle: Attentional, associative, and flexible configural mechanisms in classical conditioning. Learning & Behavior. 2012; 40:269–291. [PubMed: 22927001]

Larkin MJW, Aitken MRF, Dickinson A. Retrospective revaluation of causal judgments under positive and negative contingencies. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1998; 24:1331–1352.

Lee MD. Three case studies in the Bayesian analysis of cognitive models. Psychonomic Bulletin & Review. 2008; 15:1–15. [PubMed: 18605474]

Lubow RE, Moore AU. Latent inhibition: The effect of nonreinforced pre-exposure to the conditional stimulus. Journal of Comparative and Physiological Psychology. 1959; 52:415–419. [PubMed: 14418647]

Ludvig EA, Sutton RS, Kehoe EJ. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Computation. 2008; 20:3034–3054. [PubMed: 18624657]

Maes E, Boddez Y, Alfei JM, Krypotos AM, D’Hooge R, De Houwer J, Beckers T. The elusive nature of the blocking effect: 15 failures to replicate. Journal of Experimental Psychology: General. 2016; 145:49–71.

Matzel LD, Schachtman TR, Miller RR. Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. Learning and Motivation. 1985; 16:398–412.

McConnell BL, Miller RR. Protection from extinction provided by a conditioned inhibitor. Learning & Behavior. 2010; 38:68–79. [PubMed: 20065350]

McConnell BL, Wheeler DS, Urcelay GP, Miller RR. Protection from latent inhibition provided by a conditioned inhibitor. Journal of Experimental Psychology: Animal Behavior Processes. 2009; 35:498–508. [PubMed: 19839702]

McLaren IPL, Mackintosh NJ. An elemental model of associative learning: I. Latent inhibition and perceptual learning. Animal Learning & Behavior. 2000; 28:211–246.

Melchers KG, Lachnit H, Shanks DR. Within-compound associations in retrospective revaluation and in direct learning: A challenge for comparator theory. Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology. 2004; 57B:25–53.

Miller RR, Barnet RC, Grahame NJ. Assessment of the Rescorla-Wagner model. Psychological Bulletin. 1995; 117:363–386. [PubMed: 7777644]

Miller, RR., Matzel, LD. The comparator hypothesis: A response rule for the expression of associations. In: Bower, GH., editor. The psychology of learning and motivation: Advances in research and theory. Vol. 22. San Diego, CA, US: Academic Press; 1988. p. 51-92.

Miller RR, Witnauer JE. Retrospective revaluation: The phenomenon and its theoretical implications. Behavioural Processes. 2016; 123:15–25. [PubMed: 26342855]

Minda JP, Smith JD. Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2001; 27:775–799.

Myung IJ. Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology. 2003; 47:90–100.

Nosofsky RM, Zaki SR. Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2002; 28:924–940.

Olsson H, Wennerholm P, Lyxzèn U. Exemplars, prototypes, and the flexibility of classification models. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004; 30:936–941.

Pavlov, IP. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford, England: Oxford Univ. Press; 1927.

Perales JC, Shanks DR. Normative and descriptive accounts of the influence of power and contingency on causal judgement. Quarterly Journal of Experimental Psychology A: Human Experimental Psychology. 2003; 56A:977–1007.

Pineño O. A response rule for positive and negative stimulus interaction in associative learning and performance. Psychonomic Bulletin & Review. 2007; 14:1115–1124. [PubMed: 18229484]

Pineño O, Urushihara K, Miller RR. Spontaneous recovery from forward and backward blocking. Journal of Experimental Psychology: Animal Behavior Processes. 2005; 31:172–183. [PubMed: 15839774]

Rescorla, RA., Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH., Prokasy, WF., editors. Classical Conditioning: II. Current theory and research. New York, NY: Appleton-Century Crofts; 1972. p. 64-99.

Rescorla RA. Associative changes in excitors and inhibitors differ when they are conditioned in compound. Journal of Experimental Psychology: Animal Behavior Processes. 2000; 26:428–438. [PubMed: 11056883]

Rodgers JL. The epistemology of mathematical and statistical modeling: A quiet methodological revolution. American Psychologist. 2010; 65:1–12. [PubMed: 20063905]

Rusiniak KW, Hankins WG, Garcia J, Brett LP. Flavor-illness aversions: Potentiation of odor by taste in rats. Behavioral & Neural Biology. 1979; 25:1–17. [PubMed: 454334]

Shanks DR. Forward and backward blocking human contingency judgment. Quarterly Journal of Experimental Psychology B. 1985; 37B:1–21.

Simon P, Balci F, de Souza L, Cohen JD, Holmes P. A model of interval timing by neural integration. Journal of Neuroscience. 2011; 31:9238–9253. [PubMed: 21697374]

Stout S, Arcediano F, Escobar M, Miller RR. Overshadowing as a function of trial number: Dynamics of first- and second-order comparator effects. Learning & Behavior. 2003; 31:85–97. [PubMed: 18450071]

Stout SC, Miller RR. Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. Psychological Review. 2007; 114:759–783. [PubMed: 17638505]

Vanpaemel W. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology. 2010; 54:491–498.

Van Hamme LJ, Wasserman EA. Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. Learning and Motivation. 1994; 25:127–151.

Vrieze SI. Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Psychological Methods. 2012; 17:228–273. [PubMed: 22309957]

Waekliem DL. Introduction to the special issue on model selection. Sociological Research Methods. 2004; 33:167.

Wagner, AR. SOP: A model of automatic memory processing in animal behavior. In: Spear, NE., Miller, RR., editors. Information processing in animals: Memory mechanisms. Vol. 85. Hillsdale, NJ: Erlbaum; 1981. p. 5-44.

Wasserman EA, Castro L. Surprise and change: Variations in the strength of present and absent cues in causal learning. Learning & Behavior. 2005; 33:131–146. [PubMed: 16075834]

Wheeler DS, Miller RR. Contrasting reduced overshadowing and blocking. Journal of Experimental Psychology: Animal Behavior Processes. 2007; 33:349–359. [PubMed: 17620032]

Wheeler DS, Miller RR. Determinants of cue interactions. Behavioural Processes. 2008; 78:191–203. [PubMed: 18355987]

Witnauer JE, Miller RR. The role of within-compound associations in learning about absent cues. Learning & Behavior. 2011; 39:146–162. [PubMed: 21264569]

Witnauer JE, Urcelay GP, Miller RR. Reduced blocking as a result of increasing the number of blocking cues. Psychonomic Bulletin & Review. 2008; 15:651–655. [PubMed: 18567269]

Witnauer JE, Urcelay GP, Miller RR. The error in total error reduction. Neurobiology of Learning & Memory. 2014; 108:119–135. [PubMed: 23891930]

Yamaguchi M. Complete solution of the Rescorla-Wagner model for relative validity. Behavioural Processes. 2006; 71:70–73. [PubMed: 16274937]

Highlights

- Computational methods should be used to derive predictions from associative models.
- Optimized free parameters for each model are required for fair comparisons.
- An unbiased method of selecting free parameters for each model is described.
- Statistics like the Bayesian information criterion aid comparison of models by penalizing for number of parameters.

Table 1

Summary of models.

Model	Acquisition rule	Absent Beta	Absent Alpha
RWM	$\Delta V_{Stim1\text{-}Stim2} = \alpha_{Stim1} * \beta_{Stim2} * (\lambda - \Sigma V_{i\text{-}Stim2})$	$\beta_{Stim2} = S_{Stim2} * k1$	Null
WCM	$\Delta V_{Stim1\text{-}Stim2} = \alpha_{Stim1} * \beta_{Stim2} * (\lambda - \Sigma V_{i\text{-}Stim2})$	$\beta_{Stim2} = S_{Stim2} * k1$	$\alpha_{Stim1} = k2 * S_{Stim1} * \Sigma V_{i\text{-}Stim1}$
CEM	$\Delta V_{Stim1\text{-}Stim2} = \beta_{Stim1} * (\lambda_i - \Sigma V_{i\text{-}Stim1}) * \beta_{Stim2} * (\lambda_j - \Sigma V_{i\text{-}Stim2})$	$\beta = S * k$	NA

Note: $0 < k1 < 1$ and $-1 < k2 < 0$. By convention, when a stimulus was presented (omitted) $\lambda = 1$ (0). S = salience ($0 < S < 1$). RWM = Rescorla-Wagner (1972) model. WCM = Within Compound Model (as described in Witnauer & Miller, 2011). CEM = Conjoint Error Model. NA = Not applicable. All models assumed that $V^n = \Delta V^n + V^{n+1}$. In simulations of CEM, cues and outcomes were represented by separate values of *k*. Thus, *k1* modeled the reduced associability of absent outcomes and *k2* represented the reduced associability of absent cues.

Table 2

Wasserman & Castro (2005) results and best-fitting predictions.

	Test after Phase 2				
Group	**Cue**	**Obs**	**RWM**	**WCM**	**CEM**
Uninformed	A	8.77	10.32	10.37	8.99
	B	5.95	5.65	5.68	6.27
	C	1.08	1.83	1.90	1.55
	D	1.08	0.00	0.00	0.00
	W	**5.31**	**5.65**	**4.82**	**5.30**
	X	**6.24**	**5.65**	**5.68**	**6.27**
	Y	**6.97**	**5.65**	**6.37**	**6.54**
	Z	2.31	0.00	0.00	0.00
Informed	A	8.52	10.32	10.37	8.99
	B	5.02	5.65	5.68	6.27
	C	2.10	1.83	1.90	1.55
	D	1.84	0.00	0.00	0.00
	W	**2.89**	**5.65**	**3.49**	**3.62**
	X	**4.71**	**5.65**	**5.68**	**6.27**
	Y	**6.76**	**5.65**	**7.44**	**6.56**
	Z	1.89	0.00	0.00	0.00

Note: In a human contingency learning task, all subjects received 30 AW+, BX+, CY+, and DZ- trials in Phase 1 and 30 A+, C-, and D- in Phase 2. In Group Informed, subjects were informed that the target cue was absent during elemental trials. In Group Uninformed, information about the absent target cue was omitted during Phase 2 elemental training. In simulations of this experiment, each model used a separate absent beta parameter for the two groups. In addition to the data above, models were fit to the mean ratings of cues after Phase 1. Obs = observed food cue's effectiveness. RWM = Rescorla-Wagner (1972) model. WCM = Within Compound Model (as described in Witnauer & Miller, 2011). CEM = Conjoint Error Model. Cells with fonts in bold font represent critical comparisons for testing retrospective revaluation effects (i.e., W vs. X represents backward blocking and Y vs. X represents recovery from overshadowing).

Table 3

Results of All Simulations

Sim	Model	SSE	BIC	AIC	Scaling	*k*1	*k*2	Outcome Salience	Ctx Salience	Cue1 Salience	Cue 2 Salience	Cue 3 Salience
1	RWM	46.01	28.33	23.62	12.56	1.00	NA	0.21	NA	0.17	NA	NA
	WCM	36.71	29.27	22.20	12.72	1.00	−0.39 *	0.16	NA	0.22	NA	NA
	CEM	25.79	20.79	13.73	8.99	0.05	0.37 *	0.99	NA	0.99	NA	NA
2	RWM	6913.70	102.83	99.99	82.22	0.69	NA	0.97	NA	0.97	NA	NA
	WCM	3375.40	94.78	91.24	80.93	0.53	1.00	1.00	NA	0.70	NA	NA
	CEM	3963.00	97.19	93.65	82.21	0.23	1.00	1.00	NA	0.77	NA	NA
3	RWM	0.99	−17.47	−19.89	1.64	0.65	NA	0.59	0.54	0.58	NA	NA
	WCM	0.98	−15.12	−18.03	1.19	0.33	−0.99	0.69	0.99	0.99	NA	NA
	CEM	0.35	−27.50	−30.41	2.03	0.42	0.99	0.44	0.07	0.53	NA	NA
4	RWM	0.20	−7.82	−6.36	21.07	0.04	NA	0.56	0.77	0.06	0.41	0.09
	WCM	0.06	−13.74	−12.07	2.83	0.78	−1.00	0.99	1.00	0.33	0.03	1.00
	CEM	0.09	−10.26	−8.59	1.82	0.93	0.45	0.99	1.00	1.00	0.05	0.13
5	RWM	1.14	−1.04	−1.60	1.16	0.83	NA	0.89	0.73	0.64	0.77	0.76
	WCM	1.04	0.33	−0.31	1.65	0.99	−0.81	0.22	0.96	0.53	0.99	0.99
	CEM	0.06	−22.88	−23.51	21.89	0.10	0.99	0.11	0.60	1.00	0.04	0.22

Note: *k*1 represents reduced activation of an absent outcome. RWM = Rescorla-Wagner (1972) model. WCM = Within Compound Model (as described in Witnauer & Miller, 2011). CEM = Conjoint Error Model. For WCM and CEM *k*2 controls activation and eligibility, respectively, of an absent cue. In principle, simulations of the CEM could use *k*1 for both cues and outcomes. In practice, the addition of a second parameter for cues (*k*2) often allowed the model to provide a better fit based on BIC and AIC statistics. This reflects the trial-wise nature of the model being a poor match to the serial nature of a typical cue-outcome or CS-US pairing.

* *k*2 values for Simulation 2 represent the best-fitting values for the *k2* parameter in Group Uninformed (see Table 3 for details). The best-fitting values for Group Informed were 1.00 for both WCM and CEM, which captures the psychological intuition that informing subjects about the absent target would increase activation (WCM) or error (CEM). Notably in Simulations 1 and 2, SSE was based on only one set of predictions for both groups because the models anticipated no difference between the two groups prior to any experimental treatment (see Table 4). In Simulation 4, Cue 1 salience was used for both Cue A and Cue D, Cue 2 Salience was used for both Cue B and Cue C, and Cue 3 salience was used for X. In Simulation 4, Ctx Salience was used to represent the salience of two different contexts (inhibition training and target training), Cue 1 Salience was used for both A and C, Cue 2 Salience was used for B, and Cue 3 salience was used for X and Y.

Table 4

De Houwer and Beckers (2002) results and best-fitting predictions.

			Obs-1			RWM-1			WCM-1			CEM-1				Obs-2			RWM-2			WCM-2			CEM-2		
Group	**Ph 1**	**Ph 2**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**Ph 3**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Inf	AB+	BC+	44	50	44	30	50	19	17	61	20	10	62	21	A+	83	6	82	82	50	19	81	41	58	82	35	48
Def															A-	0	63	23	0	50	19	0	65	13	1	63	19

Note: Inf = inflate; Def = deflate. Letters A, B, and C represent weapon cues in a human contingency learning procedure. + (-) represents the occurrence (omission) of an explosion outcome in a simulated shooter task. Obs = observed ratings of the effectiveness of a weapon cue. RWM = predictions by the Rescorla-Wagner model, WCM = predictions by the Within-Compound model, and CEM = predictions by the Conjoint Error model.. In columns summarizing the first test (e.g., Obs-1), values reflect the observed mean and predicted values for the two groups.

Table 5

Grahame et al. (1994) results and best-fitting predictions.

Exp	Group	Ph1	Ph2	Ph3	Obs	RWM	WCM	CEM
1	No LI–No Ext	-	4X+	-	2.13	1.70	1.68	1.93
	No LI–Ctx Ext	-	4X+	Ctx-	1.83	1.70	1.74	1.97
	LI–No Ext	120 X-	4X+	-	1.50	1.70	1.68	1.45
	LI–Ctx Ext	120 X-	4X+	Ctx-	1.77	1.70	1.74	1.56
3	LI	120 X- (A)	120 Y- (B)	4X+	1.09	1.70	1.68	1.45
	LI–CS Ext	120 X- (A)	120 X- (B)	4X+	1.96	1.70	1.68	1.83
	Acq	- (A)	120 Y- (B)	4X+	1.82	1.70	1.68	1.93
	CS Ext Ctrl	120 X- (A)	120 X- (A)	4X+	1.46	1.70	1.68	1.45

Note: LI = Latent Inhibition, Acq = Acquisition, Ext = Extinction, Ctx = Context. X and Y were counterbalanced CSs in a conditioned lick suppression procedure with rats as subjects. Obs = observed lick suppression (log s) to X at test. RWM = predictions by the Rescorla Wagner model, WCM = predictions by the Within-Compound model, and CEM = predictions by the Conjoint Error model. Letters A and B represent the contexts of preexposures. In addition to the data presented above, we simulated Grahame et al.'s second test of the target stimulus in Experiment 1. The observations and predictions from the second test agreed with those from the first test; thus, we omitted them. Phase 3 was always conducted in Ctx A and testing of X was always conducted in an associatively neutral context. To simulate extinction of the context, we programmed 480 context-alone trials to occur after training. Notably, we did not conduct simulations of Experiment 1a because it was conducted in a leverpress suppression procedure. Neither did we simulate Experiment 2 because it was concerned with the role of the test context, which is ignored by all of the models that were simulated.

Table 6

McConnell et al. (2010; Experiments 2 and 3) results and best-fitting predictions.

Experiment	Group	Phase 1	Phase 2	Phase 3	Phase 4	Obs	RWM	WCM	CEM
1	Protect	48 A+ / 84 AB-	12 X+	36 BX- / 12 B- / 12 X-	NA	1.56	1.50	1.50	1.38
	Ctrl			36 CX- / 12 B- / 12 X-		0.85	0.93	0.78	0.85
	NoExt			60 B-		1.75	1.76	1.75	1.72
	Ext			48 X- / 12 B-		0.89	0.83	0.68	0.70
2	Protect Ext A	48 A+ / 84 AB- / 48 D-		36 BX- / 12 B- / 12	198 A-	1.01	1.34	1.01	1.16
	Protect Ctrl			X-	198 D-	1.62	1.34	1.67	1.70

Note: Ctrl = control; Ext = extinction. A-D and X were cues in a Pavlovian lick suppression procedure. A was counterbalanced with D, and B was counterbalanced with C. + represents the occurrence of a surrogate outcome that was paired with a footshock US in a session before testing. Thus, experimental manipulations were embedded in a sensory preconditioning procedure (see text for details). Obs = observed suppression (log s) to X at test. RWM = predictions by the Rescorla Wagner model, WCM = predictions by the Within-Compound model, and CEM = predictions by the Conjoint Error model.

Table 7

Summary of McConnell et al.'s (2009) results and best-fitting predictions

Experiment	Group	Phase 1	Phase 2	Phase 3	Phase 4	Obs	RWM	WCM	CEM
1	Acq	48 A+ / 84 AB-	48 Y-	3 X+	NA	2.01	1.63	1.59	1.98
	LI	48 A+ / 84 AB-	48 X-			0.99	1.63	1.59	1.10
	CmpdLI	48 C+ / 84 AB-	48 BX-			1.20	1.63	1.60	1.59
	InhibLI	48 A+ / 84 AB-	48 BX-			1.99	1.75	1.79	1.81
2	LI-ExtA	48 A+ / 48 C+ / 84 AB-	48 BX-	200 A-	3 X+	1.47	1.75	1.56	1.49
	LI-ExtC	48 A+ / 48 C+ / 84 AB-	48 BX-	200 C-		1.78	1.75	1.83	1.76
	NoLI-ExtA	48 A+ / 48 C+ / 84 AB-	48 B-	200 A-		2.06	1.63	1.82	1.99
	NoLI-ExtC	48 A+ / 48 C+ / 84 AB-	48 B-	200 C-		1.90	1.63	1.74	1.85

Note: Numbers preceding letters represent numbers of trials. Acq = acquisition control; LI = latent inhibition; CmpdLI = compounded with previously nonreinforced cue during latent inhibition treatment; InhibLI = compounded with conditioned inhibitor during latent inhibition treatment; Ext = extinction during Phase 3. A was counterbalanced with C, and X was counterbalanced with Y. + represents the occurrence of a footshock in a lick suppression procedure. Obs = observed values (log s) in a lick suppression test of X. RWM = predictions of lick suppression to X by the Rescorla Wagner model, WCM = predictions by the Within-Compound model, and CEM = predictions by the Conjoint Error model. Slashes denote interspersed trials. NA = not applicable.