2003

Equilibria of the Rescorla-Wagner Model

David Danks
University of West Florida

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Mathematical Psychology 47 (2003) 109–121
Journal of Mathematical Psychology
http://www.elsevier.com/locate/jmp

Equilibria of the Rescorla–Wagner model

David Danks*

Institute for Human & Machine Cognition, University of West Florida, 40 S. Alcaniz St., Pensacola, FL 32501, USA

Received 30 October 2000; revised 17 January 2002

Abstract

The Rescorla–Wagner model has been a leading theory of animal causal induction for nearly 30 years, and human causal induction for the past 15 years. Recent theories (especially Psychol. Rev. 104 (1997) 367) have provided alternative explanations of how people draw causal conclusions from covariational data. However, theoretical attempts to compare the Rescorla–Wagner model with more recent models have been hampered by the fact that the Rescorla–Wagner model is an algorithmic theory, while the more recent theories are all computational. This paper provides a detailed derivation of the long-run behavior of the Rescorla–Wagner model under a wide range of parameters and experimental setups, so that the model can be compared with computational theories. It also shows that the model agrees with competing theories on a wider range of cases than had previously been thought. The paper concludes by showing how recently suggested modifications of the Rescorla–Wagner model impact the long-run behavior of the model.

Keywords: Rescorla–Wagner model; Causal learning

1. Introduction

Rescorla and Wagner (1972) presented a model to explain some unusual results in animal learning. The Rescorla–Wagner model (henceforth, the R–W model) has since become influential in a wide range of areas of cognitive psychology (for an overview of its impact, see Siegel & Allan, 1996). In particular, a number of researchers have argued that the R–W model can account for much of the data on causal learning in humans (Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Lober & Shanks, 2000; Shanks, 1995; Shanks & Dickinson, 1987). Modifications of the R–W model have also been suggested to help account for some experimental data that seemingly conflict with it (see Miller, Barnet, and Grahame (1995) for an overview of the strengths and weaknesses of the R–W model, and see, e.g., Tassoni (1995) or Van Hamme and Wasserman (1994) for suggested modifications).

The R–W model describes the step-by-step changes in an individual's judgments of causal strength as that individual sees more cases. For example, an individual might want to know whether a certain kind of fertilizer causes plants to bloom more often. So, the individual might apply the fertilizer to some plants, and then determine which plants bloomed. The R–W model describes how the individual's judgment of the causal efficacy of the fertilizer changes as she sees each plant (with or without fertilizer; blooming or not blooming) individually. The R–W model is entirely iterative; nothing is said about the long-run behavior.

For several reasons, it would be useful to know how the R–W model behaves in the long run. One reason is that there are competing theories (e.g. Cheng, 1997) that describe *only* the long-run behavior. Hence, to compare the theories, we must either give an iterative instantiation of Cheng's theory (of which there are infinitely many), or else provide the (uniquely determined) long-run behavior of the R–W model. A second reason for wanting to know the long-run predictions of the R–W model is that the model accurately predicts some aspects of acquisition performance, which has led some researchers to argue that (at least part of) human causal learning must be based on something like the R–W model. However, if the long-run predictions of the model are incorrect, then we no longer have reason to

*Fax: +850-202-4440.
E-mail address: ddanks@ai.uwf.edu (D. Danks).

0022-2496/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0022-2496(02)00016-0

think that the R–W mechanism is responsible for our short-run acquisition performance.[1]

There has been prior theoretical work to determine the long-run behavior of the R–W model. One strand of research has focused on simplified experimental designs. Specifically, if we have only one binary cue and one binary effect, then Chapman and Robbins (1990) showed (using a similar method to that used in the first part of this paper) that the R–W model computes the probabilistic contrast $(P(E|C) - P(E|\sim C))$. Wasserman, Elek, Chatlosh, and Baker (1993) then extended this derivation to allow for a wider range of parameter values (but still with the same experimental design).

Another strand of work has centered on the behavior of the R–W model with a restricted set of parameters. Sutton and Barto (1981), Melz, Cheng, Holyoak, and Waldmann (1993), and Cheng (1997) used the equivalence, under certain parameter constraints (given in the following section), of the R–W model with the Widrow-Hoff rule (from Widrow and Hoff, 1960) to derive the model's long-run behavior under those constraints. Using the same parameter constraints, Gluck and Bower (1988, 1990) demonstrated the equivalence of the R–W model to a particular adaptive network. They then derived equations for the long-run behavior of an adaptive network similar (but not identical) to the R–W model using matrix notation similar to that used in the second part of this paper. In neither case, however, did they derive the equilibria for arbitrary parameter values.

We also might hope to use previous work on the delta rule, which is closely related to the R–W model. In particular, Stone (1986) provides a characterization of the long-run behavior of the delta rule in terms of the input–output correlations and the pseudo-inverse of the inter-input correlations. However, the R–W model is not simply a restricted form of the delta rule, nor vice versa. The R–W model allows for the learning rates to vary depending on the particular observed case; the delta rule allows for multi-dimensional and many-valued inputs and output. In addition, suggested modifications of the R–W model (considered in the later parts of this paper) move even further from the delta rule. Therefore, we cannot easily make use of Stone's results about the long-run behavior of the delta rule.

None of the previous theoretical work was completely general; it all made assumptions either about the experimental design or else about the parameter values.[2] In this paper, I show how to determine the long-run behavior of the R–W model outside of both types of constraints. The methods used in this paper will also generalize so that we can describe the long-run behavior of recently suggested modifications of the R–W model.

A claim in Lober and Shanks (2000) demonstrates the problems that result from not having a fully general theory of the asymptotic behavior of the R–W model. They argue that the R–W model is able to account for a specific phenomenon (the base-rate effect) if a parameter constraint is removed (and they mistakenly claim that Cheng's (1997) criticisms of the R–W model ignored this possibility). While their analysis is correct, they make the more general claim that "the unrestricted [R–W] model is able to explain many phenomena that are inconsistent with the restricted model" (Lober & Shanks, 2000, p. 198). However, Lober and Shanks' analytic solution for the long-run behavior of the unrestricted R–W model applies only to a restricted experimental design (one binary cue), and so we have no basis for evaluating their broader claim about the unrestricted R–W model.

Since there is no fully general analysis, when a researcher wants to know the long-run behavior of the R–W model for a particular experimental design that does not accord with the Widrow–Hoff restriction on parameters, she must run a simulation to determine the value to which the model (roughly) stabilizes. For example, Wasserman, Kao, Van Hamme, Katagiri, and Young (1996) used simulations to investigate cue overshadowing in the R–W model (and one other theory, which we will not consider here). Because their initial simulations ran only for 96 iterations, and seemed to give unusual results, they reran the simulations and "determined that 800 trials provided ample opportunity for the models to reach a stable point" (p. 259). Their decision, while ultimately correct, was nevertheless arbitrary. It is entirely possible (depending largely on the parameters) that the model still had not settled down after 800 iterations. The methods described in this paper eliminate the need for this kind of arbitrary decision about the stopping point.

This paper provides a straightforward algorithm for determining, for arbitrary parameter values and experimental setup, the "stable points" of the R–W model. It turns out that there is not always a unique stable point, and I further show how to determine when there are multiple stable points, and, for certain constraints, which stable point the R–W model tends towards. Finally, I apply this method to the modified R–W models suggested by recent experimental work.

2. The Rescorla–Wagner model

The R–W model applies to situations in which we have a well-defined set of logically independ-

[1] Thanks to Patricia Cheng for bringing this second reason to my attention.

[2] There is one other strand of work, best exemplified by Yamaguchi (1999). In that work, the R–W model is expressed as a set of difference or differential equations, and then Mathematica is used to determine the limit value of the equations. However, to allow for different parameter values (or a different experimental design), we must start the whole process over. Therefore, results obtained using this method are not readily generalized.

ent[3] cues $\{C_0, \ldots, C_n\}$ corresponding to the potential causes, and a well-defined outcome O. By convention, C_0 is assumed to be the experimental context (e.g., the room in which the experiment is performed). The other cues and the outcome are the events we are testing. For example, in a traditional animal conditioning experiment, we might have $C_1 =$ a tone at 60 Hz; $C_2 =$ 0.5 s light flash; and $O =$ 0.1 s shock. An experiment to test human causal learning might use $C_1 =$ fertilizer 1; $C_2 =$ fertilizer 2; $O =$ tall plants. Since it greatly simplifies the presentation of the results in this paper, we will assume that all cues and outcomes are binary variables that, for each stage in an experiment, are either present or absent (true or false).[4] Furthermore, V_i designates the "associative strength" of the outcome with the ith cue. The associative strength is (roughly) the strength attributed by the organism to the cue-outcome connection.

Given this framework, the R–W model says that the associative strength of cue C_i with the outcome after time t is $V_i^{t+1} = V_i^t + \Delta V_i^t$, where ΔV_i^t is given by:

$$\Delta V_i^t = \begin{cases} 0 & \text{if the cue does not appear at time } t, \\ \alpha_i\beta_1\left(\lambda - \sum_{\text{Cue } C_j \text{ is present at time } t} V_j\right) & \text{if the cue appears and the outcome is present at time } t, \\ \alpha_i\beta_2\left(0 - \sum_{\text{Cue } C_j \text{ is present at time } t} V_j\right) & \text{if the cue appears and the outcome is absent at time } t. \end{cases} \quad (1)$$

α_i is a measure of the salience of the particular cue. In other words, it allows us to take into consideration the fact that some cues (e.g., a shock) might be more obvious to the organism than others (e.g., a quiet tone). This unitless parameter only has meaning relative to the other cue saliences. β_1 and β_2 represent the salience of the particular situations in which the outcome does or does not occur. λ is the maximum level of associative strength possible, and in animal experiments, is usually thought to be proportional to the intensity of the outcome.

The Widrow–Hoff rule is simply Eq. (1) with all of the α_i's equal, $\lambda = 1$, and $\beta_1 = \beta_2$. Widrow and Hoff (1960) showed that, in the long run, we can minimize the sum of the squared error for the pattern frequencies in the data by updating the V_i's using the Widrow–Hoff rule. Cheng (1997) used Widrow and Hoff's result to derive some conditions under which the R–W model (with appropriate parameters) computes conditional contrasts. However, the method used by Widrow and Hoff for determining the long-run behavior does not readily generalize to other parameters permitted by the R–W model.[5]

The notion of an *asymptote* for the R–W model is not, in general, well defined. If there are combinations of cues for which the outcome is not deterministic (i.e., always occurring or not occurring), then changes in associative strength need not tend to zero as we accumulate more data. Hence, there is not really a sense in which the associative strength of a cue in the R–W model "converges" to a *particular* value.

Therefore, rather than talking about asymptotes, I will focus on finding equilibria for the R–W model. We can think about the R–W model as "settling down" when the expected change for each of the associative strengths is zero, even though the actual change will almost certainly be non-zero. I define a vector of associative strengths $\boldsymbol{V} = \langle V_0, \ldots, V_n \rangle$ (one for each cue) to be an *equilibrium of the R–W model for a probability distribution* provided that, for every cue, the expected value (if we assume random presentation of the cases) of the change in the associative strength of that cue with the outcome is zero, that is, if and only if $\forall i \ (E(\Delta V_i) = 0)$. The equilibrium values will be dependent on the frequency with which various combinations of cues appear in the actual experiment, and, for the purposes of this paper, I will regard an experimental design as governed by a joint probability distribution on the cues and the outcome, regarded as binary variables. I want to emphasize, however, that a learner using the Rescorla–Wagner model will, in many cases, *not* converge on the equilibrium for that experimental design, but rather will only converge on a neighborhood of the equilibrium (where the size of the neighborhood is dependent on, e.g., the learning rates).

3. Equilibria of the Rescorla–Wagner model

In this section, I provide the general form for the long-run equilibria of the R–W model in terms of the

[3] By 'logically independent,' I simply mean that the description of one cue does not include another one. For example, we cannot have $C_1 =$ "green light flash", and $C_2 =$ "green light flash and red light flash", where the green flash in the two cues is the same.

[4] The R–W model actually only requires that each cue have a marked case, which need not correspond to the presence of the cue. However, the methods used in this paper are not changed by assuming that the marked state is the presence of the cue, and in fact this assumption holds true of most experiments in this area.

[5] The problem with the Widrow–Hoff method is that it starts with the asymptotes, and works backwards to find a procedure that will produce those asymptotes. Hence, it is (to a certain extent) just good fortune that the R–W model is equivalent to the Widrow–Hoff rule for certain parameters.

probability distribution, as well as an example demonstrating its use. In the following section, I prove that this method correctly yields the equilibria of the model. For the purposes of this presentation, I assume that $\beta_1 = \beta_2 = \beta$, and that $\lambda = 1$. These are standard theoretical assumptions (though they are not necessarily used when simulations are performed), and they simplify the derivations. I will show in a later section how to relax these assumptions. I do not, in contrast to Cheng (1997), assume that the α_i's are equal.

Given a probability distribution over the cues and outcome, a vector of associative strengths is an equilibrium of the R–W model for that distribution if and only if it is a solution to the following set of $n+1$ equations:

$$P(O|C_i) - \sum_{j=0}^{n} P(C_j|C_i)V_j = 0. \tag{2}$$

In matrix form, the system of $n+1$ equilibrium equations is

$$\begin{pmatrix} P(C_0|C_0) & P(C_1|C_0) & \dots & P(C_n|C_0) \\ P(C_0|C_1) & P(C_1|C_1) & \dots & P(C_n|C_1) \\ \dots & \dots & \dots & \dots \\ P(C_0|C_n) & P(C_1|C_n) & \dots & P(C_n|C_n) \end{pmatrix} \begin{pmatrix} V_0 \\ V_1 \\ \dots \\ V_n \end{pmatrix} = \begin{pmatrix} P(O|C_0) \\ P(O|C_1) \\ \dots \\ P(O|C_n) \end{pmatrix}. \tag{3}$$

The equilibrium values of the V_i's, can therefore be determined by using standard Gaussian elimination on the following augmented matrix (noting that $P(C_j|C_j) = 1$):

$$\left(\begin{array}{cccc:c} 1 & P(C_1|C_0) & \dots & P(C_n|C_0) & P(O|C_0) \\ P(C_0|C_1) & 1 & \dots & P(C_n|C_1) & P(O|C_1) \\ \dots & \dots & \dots & \dots & \dots \\ P(C_0|C_n) & P(C_1|C_n) & \dots & 1 & P(O|C_n) \end{array}\right).$$

We can then determine the value of each of the V_i's by reducing the part on the left of the dotted line to an upper triangular matrix. Matrix programming languages provide routines for easily solving the above system of equations. A Java applet implementing the fully general algorithm (which determines equilibria even outside of the above parameter restrictions) is available at: http://www.phil.cmu.edu/guests/ddanks/RWCalculator.html.

To illustrate the practical implementation of the above method, consider an experiment from Spellman (1996). Her cover story is that there are two liquids (red and blue) that are potentially fertilizers, and the experimental participant is given the rates at which flowers bloom for the four possible conditions (no liquid, red liquid, blue liquid, and both liquids). Let us define the cues to be:

C_0: the constant background (e.g., the pots in which the flowers are grown),

C_1: the red liquid,

C_2: the blue liquid.

In the experiment, the summary statistics for blooming are (p. 173; first experiment in A):

- red and blue liquid present: 5/5 plants flowered,
- red liquid only: 10/15 plants flowered,
- blue liquid only: 5/15 plants flowered,
- no liquid present: 0/5 plants flowered.

Using these contingencies, we can compute the probabilities needed for the augmented matrix that represents the system of 3 equations. That matrix is

$$\begin{pmatrix} P(C_0|C_0)=1 & P(C_1|C_0)=0.5 & P(C_2|C_0)=0.5 & P(E|C_0)=0.5 \\ P(C_0|C_1)=1 & P(C_1|C_1)=1 & P(C_2|C_1)=0.25 & P(E|C_1)=0.75 \\ P(C_0|C_2)=1 & P(C_1|C_2)=0.25 & P(C_2|C_2)=1 & P(E|C_2)=0.5 \end{pmatrix}.$$

Straightforward Gaussian elimination on this matrix yields:

$$V_0 = 0,$$

$$V_1 = 2/3,$$

$$V_2 = 1/3,$$

which are the associative strengths the R–W model should approach in the long run for her experimental design.

4. Derivation of the equilibrium algorithm

In this section, I show that the above method correctly determines the equilibria. Given some particular pattern of cues and outcome, and the current associative strengths, Eq. (1) defines the change in V_i for any of the cues. Define δ_i to be 1 if cue C_i appears, and 0 if it is absent. Similarly, define δ_O to be 1 if the outcome occurs, and 0 if it is absent. Then, for any particular stage, we can rewrite Eq. (1) as

$$\Delta V_i = \delta_i \alpha_i \beta \left(\delta_O - \sum_{\delta_j=1} V_j \right). \tag{4}$$

Define a cue pattern as the combination of cues (or their absences) that appear in a trial. I will use '$\sim$' before a cue to indicate the absence of the cue. Thus, if we have four possible cues, then one possible cue pattern is: $C_0 \sim C_1 C_2 C_3$ (i.e., all of the cues except C_1 occur). I define *ExpDesign* to be the set of all possible cue patterns allowed in the experimental design. So, for example, if we have two possible cues $\{C_1, C_2\}$ that can appear separately, and the constant background cue C_0, then $ExpDesign = \{C_0C_1C_2, \quad C_0 \sim C_1C_2, \quad C_0C_1 \sim C_2,$

$C_0 \sim C_1 \sim C_2\}$. Also, define $ExpDesign(X)$ as the set of all possible cue patterns allowed in the experimental design in which the cue(s) X appears (unnegated). So, to use the previous example of two cues, $ExpDesign(C_1) = \{C_0C_1C_2, C_0C_1 \sim C_2\}$. Finally, for some cue pattern p, we will define $Shown(p)$ to be the set of indices of the cues that appear in the pattern. For example, $Shown(C_0 \sim C_1) = \{0\}$, and $Shown(C_0C_1) = \{0, 1\}$.

Recall that a particular vector of associative strengths is an equilibrium if and only if the expected change for all of the strengths is zero. Returning to Eq. (4), ΔV_i is zero for all patterns $p \notin ExpDesign(C_i)$. Hence, we can restrict our attention to $p \in ExpDesign(C_i)$. We need to compute the expected change for an arbitrary pattern $p \in ExpDesign(C_i)$. δ_O will be 1 with probability $P(O|p)$, and 0 with probability $P(\sim O|p)$. Therefore, for a particular pattern p, the expected value of Eq. (4) is

$$E(\Delta V_i) = P(O|p)\delta_i\alpha_i\beta\left(1 - \sum_{\delta_j=1} V_j\right) + P(\sim O|p)\delta_i\alpha_i\beta\left(-\sum_{\delta_j=1} V_j\right). \quad (5)$$

Taking advantage of the fact that $P(O|p) + P(\sim O|p) = 1$, we can rewrite (5) as

$$E(\Delta V_i) = \delta_i\alpha_i\beta\left(P(O|p) - \sum_{\delta_j=1} V_j\right). \quad (6)$$

Eq. (6) is the expected change for a particular pattern. Equilibria were defined as those values for which the *overall* expected change is zero. Since the patterns are all mutually exclusive, the total expected change is just the sum of these changes, weighted by the probability with which each pattern occurs (i.e., $P(p)$). In other words, for every cue C_i, we have that

$$E(\Delta V_i) = \sum_{p \in ExpDesign(C_i)} P(p)\alpha_i\beta \times \left(P(O|p) - \sum_{j \in Shown(p)} V_j\right). \quad (7)$$

To find the equilibria of the R–W model for a particular probability distribution P, we set each of these $E(\Delta V_i) = 0$, and solve for the V_i's. For any particular i, α_i and β appear in every term of the sum, and so we can divide through by them (and since the $E(\Delta V_i) = 0$, this division has no effect on the right-hand side). We can then separate the constant terms from the V_j terms, and we have

$$E(\Delta V_i) = 0 = \sum_{p \in ExpDesign(C_i)} P(p)P(O|p) - \sum_{p \in ExpDesign(C_i)} P(p) \sum_{j \in Shown(p)} V_j. \quad (8)$$

Consider just the first sum in (8). First, note that, since $P(p)P(O|p) = P(O\&p)$, each term in the sum is the joint probability distribution of O and the particular pattern p, where the only condition on the p's is that they must all contain cue C_i. But since the sum is over *all* possible patterns that contain C_i, this sum is just equal to $P(O\&C_i)$: the joint probability of the cue C_i and the outcome.

Consider the second sum in Eq. (8). Note that V_j only occurs when $j \in Shown(p)$ for a pattern p. Hence, we can rewrite this term as $\sum_{j=0}^{n} \sum_{p \in ExpDesign(C_iC_j)} P(p)V_j$.[6] For each V_j, the coefficient is just the sum of the probabilities of the occurrences of patterns with C_i and C_j. But this is just equal to $P(C_i\&C_j)$: the joint probability of C_i and C_j. Hence, we can rewrite Eq. (8) as

$$E(\Delta V_i) = 0 = P(O\&C_i) - \sum_{j=0}^{n} P(C_j\&C_i)V_j. \quad (9)$$

Since $P(A\&B) = P(A|B)P(B)$, we can divide both sides by $P(C_i)$, which results in the equations described in (2).

5. Generalizations of the Rescorla–Wagner model

Throughout these derivations, I have assumed that $\beta_1 = \beta_2 = \beta$, and that $\lambda = 1$. I will now relax those assumptions. First, let us assume that $\lambda \neq 1$. Then Eq. (7) becomes

$$E(\Delta V_i) = \sum_{p \in ExpDesign(C_i)} P(p)\alpha_i\beta \times \left(\lambda P(O|p) - \sum_{j \in Shown(p)} V_j\right). \quad (10)$$

Since λ only appears in the term without any V_j's, the only difference in the final equations is that the $P(O\&C_i)$ term is multiplied by λ. Hence, we can rewrite Eq. (2) as

$$\lambda P(O|C_i) - \sum_{j=0}^{n} P(C_j|C_i)V_j = 0. \quad (11)$$

The matrix changes similarly.

The situation in which $\beta_1 \neq \beta_2$ is important since it is often thought (on theoretical grounds) that $\beta_1 > \beta_2$. This inequality corresponds to the claim that cases in which the outcome occurs are more salient to the organism than cases in which the outcome is absent (though see Lober and Shanks (2000) for arguments that the absence of the effect can be more salient than its presence). To incorporate this inequality, we need to reconsider Eq. (10), in which we collapsed situations in which the outcome occurs and those in which it is absent. In order to allow for $\beta_1 \neq \beta_2$, we must separate those two sets of

[6] Recall that $ExpDesign(C_iC_j)$ is the set of all patterns in which *both* C_i and C_j occur.

events. The outcome occurs with probability $P(O|p)$, and it is absent with probability $P(\sim O|p)$. Hence, we have

$$E(\Delta V_i) = 0 = \sum_{p \in ExpDesign(C_i)} P(p)\alpha_i \Bigg[P(O|p)\beta_1 \times \left(\lambda - \sum_{j \in Shown(p)} V_j\right) + P(\sim O|p)\beta_2 \left(-\sum_{j \in Shown(p)} V_j\right)\Bigg]. \quad (12)$$

As before, we can cancel the α_i factors. Noting that $P(X|p)P(p) = P(X \& p)$ and rearranging the terms some, we have

$$\begin{aligned} E(\Delta V_i) &= 0 \\ &= \sum_{p \in ExpDesign(C_i)} \lambda P(O \& p)\beta_1 \\ &\quad - \sum_{p \in ExpDesign(C_i)} (P(O \& p)\beta_1 \\ &\quad + P(\sim O \& p)\beta_2) \sum_{j \in Shown(p)} V_j. \end{aligned} \quad (13)$$

Since V_j terms only appear when the pattern contains C_j, we can further reduce Eq. (13) to

$$\begin{aligned} E(\Delta V_i) = 0 = &\sum_{p \in ExpDesign(C_i)} \lambda P(O \& p)\beta_1 \\ &- \sum_{j=0}^{n} \sum_{p \in ExpDesign(C_i C_j)} (P(O \& p)\beta_1 \\ &+ P(\sim O \& p)\beta_2) V_j. \end{aligned} \quad (14)$$

Compare Eq. (14) to Eq. (7). If $\beta_1 = \beta_2 = \beta$, and $\lambda = 1$, then (14) reduces to (7), as we would expect.[7] Also, (14) shows how to modify the matrix to take account of the differing β's, since the constant and V_j terms are already separated. Hence, we have a method for determining, for arbitrary parameters, the equilibria of the R–W model.

6. The matrix method of equilibrium determination

The above derivation of the equilibria of the R–W model suggests a different characterization using matrices.[8] Suppose we have an enumeration (in arbitrary order) of the possible cue patterns in our experimental design, and further suppose that there are m different cue patterns. We define the matrix $\mathbf{M}$ such that $M_{ji} = \delta_i(p_j)$, where δ_i is as defined earlier. That is, the element in the ith column and jth row is 1 (0) if C_i is present (absent) in cue pattern j. Define the $m \times 1$ matrix $\mathbf{O}$ such that $O_j = P(O|p_j)$. Finally, define diagonal matrices for the saliences: $\mathbf{A}$ such that $A_{ii} = \alpha_i$, and for the probabilities of the patterns: $\mathbf{P}$ such that $P_{jj} = P(p_j)$. Given these matrices, a vector $\mathbf{V}$ of associative strengths is an equilibrium of the R–W model if and only if

$$\mathbf{M}^T\mathbf{PMV} = \mathbf{M}^T\mathbf{PO}. \quad (15)$$

To see how this method works, consider the earlier example from Spellman (1996). In that case, we had two distinct cues which could each appear independently of the other. So, we can arbitrarily enumerate the possible cue patterns as $p_0 = C_0C_1C_2$; $p_1 = C_0C_1 \sim C_2$; $p_2 = C_0 \sim C_1C_2$; $p_3 = C_0 \sim C_1 \sim C_2$. The left-hand side of Eq. (15) then becomes

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} P(p_0) & 0 & 0 & 0 \\ 0 & P(p_1) & 0 & 0 \\ 0 & 0 & P(p_2) & 0 \\ 0 & 0 & 0 & P(p_3) \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} V_0 \\ V_1 \\ V_2 \end{pmatrix}.$$

Similarly, the right-hand side of (15) becomes

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} P(p_0) & 0 & 0 & 0 \\ 0 & P(p_1) & 0 & 0 \\ 0 & 0 & P(p_2) & 0 \\ 0 & 0 & 0 & P(p_3) \end{pmatrix} \times \begin{pmatrix} P(O|p_0) \\ P(O|p_1) \\ P(O|p_2) \\ P(O|p_3) \end{pmatrix}.$$

Substituting in the probabilities from the actual experiment, we find that $\mathbf{V}$ is an equilibrium if and only if it satisfies

$$\begin{pmatrix} 1.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.125 \\ 0.5 & 0.125 & 0.5 \end{pmatrix} \begin{pmatrix} V_0 \\ V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.375 \\ 0.25 \end{pmatrix}.$$

The $\mathbf{V}$ that satisfies this equality is the same as the one earlier found using the method of $n+1$ equations in $n+1$ unknowns.

7. Derivation and generalization of the matrix method

Eq. (6) gives the expected change for a particular variable, given some particular pattern. Translating that

[7] This happens because every term now has a β, so we can cancel them out. Since $\lambda = 1$, we no longer have to include it. And since $P(O \& p) + P(\sim O \& p) = P(p)$, we can use the same argument as before to show that the coefficient for each V_j is $P(C_iC_j)$.

[8] I am indebted to an anonymous reviewer for *Journal of Mathematical Psychology* for pointing out and describing this alternative formulation.

equation into the matrix terminology, we have

$$(\text{for pattern } p_j)E(\Delta V_i) = A_{ii}\beta\left(O_j - \sum_{M_{jk}=1} V_k\right)M_{ji}. \tag{16}$$

Note that we cannot move the final M_{ji} term into the sum in the brackets, since that term ensures that the expected change is zero for all variables that do not appear in the pattern p_j. Given this formula for the expected change for a particular cue-pattern pair, the weighted sum of changes for a particular cue (for all patterns) can be represented in matrix terms as

$$E(\Delta V_i) = \beta A_{ii}\sum_{p=0}^{m} P_{pp}(O_p - [MV]_p)M_{pi}, \tag{17}$$

where $[\mathbf{X}]_i$ is the ith row of matrix $\mathbf{X}$. Therefore, the overall change in the vector $\mathbf{V}$ is given by

$$E(\Delta\mathbf{V}) = \beta\mathbf{A}\mathbf{M}^T\mathbf{P}(\mathbf{O} - \mathbf{M}\mathbf{V}). \tag{18}$$

If we then set the overall change in $\mathbf{V}$ equal to $\mathbf{0}$ (to find the equilibrium), then we can immediately see that $\mathbf{V}$ is an equilibrium if and only if

$$\beta\mathbf{A}\mathbf{M}^T\mathbf{P}\mathbf{M}\mathbf{V} = \beta\mathbf{A}\mathbf{M}^T\mathbf{P}\mathbf{O}. \tag{19}$$

If we then multiply (from the left) both sides of Eq. (19) by $1/\beta$ and $\mathbf{A}^{-1}$ (given by $A_{ii}^{-1} = 1/\alpha_i$), then we have Eq. (15) as a characterization of the equilibria of the R–W model.

Throughout the above derivation, I again assumed that $\beta_1 = \beta_2 = \beta$ and that $\lambda = 1$. As in the first characterization of the R–W equilibria, we need to relax these assumptions. Consider first the case in which $\lambda \neq 1$. In this case, we simply multiply $\mathbf{O}$ by λ, so (15) becomes

$$\mathbf{M}^T\mathbf{P}\mathbf{M}\mathbf{V} = \lambda\mathbf{M}^T\mathbf{P}\mathbf{O}. \tag{20}$$

Now let us assume that $\beta_1 \neq \beta_2$. To determine the equilibria in this case, we need to define two new matrices: the $m \times m$ diagonal matrix $\mathbf{N}$ such that $N_{jj}^* = P(\sim O|p_j)$, and the diagonal matrix $\mathbf{O}^*$ with $O_{ii}^* = O_i$ (that is, $\mathbf{O}^*$ is just $\mathbf{O}$ converted to a diagonal matrix). For a particular pattern p_j, we then have

$$E(\Delta V_i) = A_{ii}\left[O_j\beta_1\left(\lambda - \sum_{M_{jk}=1} V_k\right) + \beta_2 N_{jj}\left(-\sum_{M_{jk}=1} V_k\right)\right]M_{ji}. \tag{21}$$

Taking the weighted average of (21) over all possible patterns, we then have

$$E(\Delta V_i) = A_{ii}\sum_{p=0}^{m} P_{pp}[\beta_1 O_p(\lambda - [\mathbf{MV}]_p) - \beta_2 N_{pp}[\mathbf{MV}]_p]M_{pi}. \tag{22}$$

Therefore, the expected change of the whole strength vector is given by

$$E(\Delta\mathbf{V}) = \mathbf{A}\mathbf{M}^T\mathbf{P}[\beta_1\lambda\mathbf{O} - (\beta_1\mathbf{O}^* + \beta_2\mathbf{N})\mathbf{M}\mathbf{V}]. \tag{23}$$

So the expected change will equal $\mathbf{0}$ just when

$$\beta_1\lambda\mathbf{M}^T\mathbf{P}\mathbf{O} = \mathbf{M}^T\mathbf{P}(\beta_1\mathbf{O}^* + \beta_2\mathbf{N})\mathbf{M}\mathbf{V}. \tag{24}$$

Hence, we can also use a matrix representation for the equilibria of the R–W model for arbitrary parameters.

8. Infinitely many equilibria

One problem with the above methods is that they do not necessarily determine a unique equilibrium. Formally, there are infinitely many equilibria if and only if the determinant of the strength matrix is zero. That is, there are infinitely many equilibria if and only if $\det(\mathbf{M}^T\mathbf{P}\mathbf{M}) = 0$. This is a relatively simple computation, though it does require us to know the probabilities with which each pattern appears. The R–W model is fully deterministic, though, and so even if there are infinitely many equilibria, only one will actually occur. Therefore, we can ask: if there are an infinite number of equilibria, can we determine in advance which one will occur? It turns out that the answer to this question is "yes", for at least some situations. Specifically, we can determine the actual equilibrium if the infinity of possible equilibria are due to the presence of a set S of coextensive cues, where by "coextensive" I mean that the cues in the set S always and only appear together. For example, if $ExpDesign = \{X, XAB\}$, then A and B are coextensive, since they only appear together. There are infinitely many equilibria whenever we have coextensive cues, since the determinant of $\mathbf{M}^T\mathbf{P}\mathbf{M}$ will always be zero. Nevertheless, if the infinity of equilibria is due solely to the coextensive cues, then we can calculate which equilibrium will actually occur.

First, we need to define a "new" cue C_S, which is just the composite cue composed of all of the cues in S. This "cue" always exists, since the actual cues only ever appear together. So, in the example in the previous paragraph, we would now have $ExpDesign = \{X, XC_S\}$. By assumption, there is a unique solution for the model when the coextensive cues are replaced with V_S. And since cue strength is additive in the R–W model, we know that

$$V_S = \sum_{s\in S} V_s. \tag{25}$$

In other words, the strength of the composite cue can be decomposed into the sum of the strengths of the individual cues. The natural next question is how much strength will each cue in the composite receive? To answer this question, consider the ΔV_i equation for some $i \in S$. For any particular pattern p shown at time t,

we can express this equation as $\Delta V_i^t = \alpha_i \times Error_t$, where $Error_t = \lambda - \sum_{j \in Shown(p)} V_j^t$; or 0, if C_i was not shown. Since all of the cues in S are always shown together, $Error_t$ will be the same for all of them, for all times t.

Now, consider the sum over time of the changes in V_i for some $i \in S$. At time T, we have

$$V_i^T = \sum_{t=0}^{T} \Delta V_i^t = \alpha_i \sum_{t=0}^{T} Error_t. \tag{26}$$

In other words, the strength at any time T is just α_i times the sum of the errors for every time before T. However, as noted above, $Error_t$ is the same for all of the cues in S, and so the sum of the errors is also the same. Hence, we have

$$(\forall i, j \in S) \quad \left(\frac{V_i}{\alpha_i} = \frac{V_j}{\alpha_j}\right). \tag{27}$$

Combining (25) and (27), we find that

$$\begin{aligned} &\forall i \in S \quad \left(V_S = \sum_{j \in S} \frac{\alpha_j V_i}{\alpha_i}\right) \\ &\Rightarrow \forall i \in S \quad \left(V_i = \frac{\alpha_i}{\sum_{j \in S} \alpha_j} V_S\right). \end{aligned} \tag{28}$$

So, if the infinity of equilibria is due to a set S of coextensive cues, then the equilibrium that will actually occur will be one in which the strengths of the "composite cue" and the remaining cues are determined using one of the methods described above, and the cues in S divide up the strength of their "composite cue" in proportion to their saliences.

We might think that the conditions for this case are quite restrictive. However, the standard blocking experiment meets these conditions. In a standard blocking design, we have $ExpDesign = \{XA, XAB\}$ (where X is the background), and typically the effect always occurs. These experiments are designed to test whether cue B acquires any associative strength. The interesting feature of these experiments for this section is that we have a set $S = \{X, A\}$ of coextensive cues which results in infinitely many equilibria for the model as a whole, and so the equilibrium that actually occurs will depend on the relative saliences of X and A. Hence, two blocking experiments that use identical statistics, but that have different cover stories, should result in different estimates of the strength of X and A.

9. Equilibria of an augmented Rescorla–Wagner model

There are known empirical results that the R–W model cannot explain. One of the most significant of these is retrospective updating. According to the R–W model, V_i only changes when C_i occurs. Hence, it cannot explain data in which strengths change, even though the cue is not presented. Van Hamme and Wasserman (1994) give a clear account of a case in which strengths change when patterns without the cue are shown. To correct this shortcoming, several authors (e.g., Tassoni, 1995; Van Hamme & Wasserman, 1994; Wasserman et al., 1996) have argued that the R–W model should be altered to allow updating when the absence of a cue is "informative" in some sense. The intuitive idea is that sometimes (but not always) we learn something from the absence of a cue.

Both Tassoni (1995) and Van Hamme and Wasserman (1994) offer specific modifications of the R–W model to account for the problematic data. Their theories are not exactly the same, however. First, Tassoni allows that the "counterpart to λ" in the equations when the outcome does not occur can be different than zero. In other words, he allows that the absence of the *outcome* (as opposed a particular cue's absence) might be particularly salient. Second, Tassoni allows for the possibility that not every cue absence is salient. Since Van Hamme and Wasserman's theory is just the special case of Tassoni's theory in which the absence of the outcome is not more salient (than in the standard R–W model) and every cue absence *is* salient, I will focus on Tassoni's model. Throughout this discussion, I will derive the equilibria using the probability equations, rather than the matrices, since it simplifies the presentation. All of the results in this section can also be represented using matrices.

Let us consider Tassoni's first modification of the standard R–W model, since its effect on the long-run equilibria is simply to introduce a correction factor, as in the case of $\lambda \neq 1$. To incorporate the first modification, we need to allow the constant term to be some value $\mu < 0$ when the cue appears and the outcome is absent. In other words, we need to rewrite the standard R–W ΔV_i equation as

$$\Delta V_i = \begin{cases} 0 & \text{if the cue is absent,} \\ \alpha_i \beta \left(\lambda - \sum_{C_j \text{ is shown}} V_j\right) & \text{if the cue appears and the outcome is present,} \\ \alpha_i \beta \left(\mu - \sum_{C_j \text{ is shown}} V_j\right) & \text{if the cue appears and the outcome is absent,} \end{cases} \tag{29}$$

where $-1<\mu<0$. This implies that we can rewrite the $E(\Delta V_i)$ equations as

$$E(\Delta V_i) = \sum_{p \in ExpDesign(C_i)} P(p)\alpha_i\beta \left[(\lambda P(O|p) + \mu P(\sim O|p)) - \sum_{j \in Shown(p)} V_j \right] = 0. \quad (30)$$

In effect, we need only correct for the fact that we previously assumed that the contribution to the constant factor of the "no outcome" cases was zero. We can rewrite (30) using the same equivalencies as in the standard R–W model to get

$$E(\Delta V_i) = \lambda P(O|C_i) + \mu P(\sim O|C_i) - \sum_{j=0}^{n} P(C_i|C_j)V_j = 0. \quad (31)$$

We simply add a (negative) correction factor to the constant term, and solve the equations as before.

Now consider Tassoni's second modification: namely, that some cue absences are salient. For Tassoni, the salience of the absence of a cue is a function purely of the cue itself (and not, for example, of the previously observed patterns). Therefore, we use γ_i to denote the salience of the absence of cue C_i. The modified R–W equations (with $\beta_1 = \beta_2 = \beta$, $\lambda = 1$, and $\mu = 0$) are thus

$$\Delta V_i = \begin{cases} -\gamma_i\beta\left(1 - \sum_{C_j \text{ is shown}} V_j\right) & \text{if the cue is absent and the outcome is present,} \\ -\gamma_i\beta\left(0 - \sum_{C_j \text{ is shown}} V_j\right) & \text{if the cue is absent and the outcome is absent,} \\ \alpha_i\beta\left(1 - \sum_{C_j \text{ is shown}} V_j\right) & \text{if the cue appears and the outcome is present,} \\ \alpha_i\beta\left(0 - \sum_{C_j \text{ is shown}} V_j\right) & \text{if the cue appears and the outcome is absent.} \end{cases} \quad (32)$$

In this model, the absence of a cue can be salient primarily for two reasons: either the subject has come to expect that cue based on previous experience with the data, or because the experimenter explicitly points out to the subject that the cue was absent. That is, a cue's absence is not always salient. Therefore, define $q_{i,p}$ to be the probability that the absence of C_i is salient, given that we have pattern p. We thus have the following expected change equation:

$$E(\Delta V_i) = \sum_{p \in ExpDesign(C_i)} P(p)\alpha_i \times \left(P(O|p) - \sum_{j \in Shown(p)} V_j\right) - \sum_{p \in ExpDesign(\sim C_i)} \times P(p)\gamma_i q_{i,p} \left(P(O|p) - \sum_{j \in Shown(p)} V_j\right) = 0. \quad (33)$$

Noting that the first sum is the same as the standard R–W model (with the α_i factor included), we can reduce (33) to

$$E(\Delta V_i) = \alpha_i P(O \,\&\, C_i) - \alpha_i \sum_{j=0}^{n} P(C_i \,\&\, C_j)V_j - \sum_{p \in ExpDesign(\sim C_i)} P(p)\gamma_i q_{i,p} \left(P(O|p) - \sum_{j \in Shown(p)} V_j\right) = 0. \quad (34)$$

We can only simplify (34) further if we constrain $q_{i,p}$ in some way. One plausible constraint is that the probability that a cue absence is salient is independent of the pattern shown. In that case, we can reduce (34) to

$$E(\Delta V_i) = (\alpha_i - q_i\gamma_i)P(O \,\&\, C_i) - \sum_{j=0}^{n} (\alpha_i P(C_i \,\&\, C_j) - q_i\gamma_i P(C_j \,\&\, \sim C_i))V_j = 0. \quad (35)$$

Eq. (35) is in a form that we can easily compute and solve.

Before leaving this discussion of Tassoni's model, I should note that one drawback of the augmented models is that they do not necessarily make the same predictions as the standard R–W model on experiments known to support the standard model. For example, Baker et al.'s (1993) experimental condition PR.5/.4, which they claim supports the standard R–W model, cannot be automatically assumed to support Tassoni's augmented R–W model.[9]

[9] Specifically, the augmented R–W models can make differing predictions for any experiments for which there is not a perfect equilibrium (defined in the next section).

10. Perfect equilibria

The conditional ΔP theory of human causal judgment predicts that the causal strength rating for a particular factor will be (proportional to) the conditional contrast for that factor (see, e.g., Spellman, 1996). The general form of the conditional contrast for a particular cue is given by $\Delta P_{C.\{X\}} = P(O|C \,\&\, X) - P(O|\sim C \,\&\, X)$, where X ranges over the possible states of the other cues. So, for example, if we have two potential cues, C_1 and C_2, then there are two conditional contrasts for C_1: $\Delta P_{C1.\{C2\}} = P(O|C_1 \,\&\, C_2) - P(O|\sim C_1 \,\&\, C_2)$ and $\Delta P_{C1.\{\sim C2\}} = P(O|C_1 \,\&\, \sim C_2) - P(O|\sim C_1 \,\&\, \sim C_2)$. Some conditional contrasts may not have definite values for a particular cue, and even if all of the contrasts have definite values, they may not agree. Nevertheless, in order to compare the R–W model with the conditional ΔP theory, we should determine the conditions under which the equilibria of the R–W model consist of conditional contrasts (given that they are defined and equal regardless of X). Throughout the remainder of this section, I will assume that $\beta_1 = \beta_2 = \beta$ (though I place no constraints on α_i or λ).

We can best express the desired condition by defining a particular type of equilibrium: namely, a perfect equilibrium. I earlier pointed out that, at an equilibrium, the expected change of a cue strength is only necessarily zero for the weighted sum of all of the patterns. Even at equilibrium, the expected change for any particular pattern might be quite different from zero. For some equilibria, however, the expected change is zero regardless of the pattern. Hence, we have the following definition:

A strength vector $\mathbf{V} = \langle V_0, \ldots, V_n \rangle$ is a *perfect equilibrium* if and only if $\forall p \in ExpDesign(\lambda P(O|p) = \sum_{j \in Shown(p)} V_j)$. In matrix terminology, $\mathbf{V}$ is a *perfect equilibrium* if and only if $\lambda \mathbf{O} = \mathbf{MV}$.

Not all experiments have a perfect equilibrium. Consider the case in which the outcome occurs if and only if two cues have the same value (either present or absent), and in which we see all four cases. If there were a perfect equilibrium, then the strength of the background cue would have to be 1 (because the outcome occurs when both of the variable cues do not occur), and the strength of each variable cue would have to be -1 (since the outcome is absent when exactly one variable cue occurs). But then $V_0 + V_1 + V_2 = -1 \neq P(O|C_0 C_1 C_2)$.

It is important to note that a vector's status as a perfect equilibrium does not depend on the α_i parameters in the model. Given the concept of a perfect equilibrium, we have the following theorem (proof provided in the appendix):

Theorem. *For a particular experimental design, an equilibrium* $\mathbf{V}$ *of the R–W model consists of* λ*-scaled conditional contrasts (to the extent possible) if and only if* $\mathbf{V}$ *is a perfect equilibrium.*

Hence, if we can determine the class of experimental designs that have a perfect equilibrium, then we will have found the class of designs for which the R–W model computes (to the extent possible) λ-scaled conditional contrasts, regardless of the α_i parameters. Given that λ is typically assumed to equal 1 for human causal learning experiments, the class of perfect equilibrium designs is thus the same as the class of conditional contrast designs for the R–W model. Here (without proofs) are three classes of experiments that always have a perfect equilibrium:

1. Experiments with a nested cue design (see Cheng, 1997, for more details about what constitutes a nested cue design).
2. Experiments in which every possible combination of cues appears and, for each cue, all of that cue's conditional contrasts are the same.
3. Experiments in which there are three or fewer cues (not counting the background cue), and, for each cue, all of the conditional contrasts for that cue are the same.

The fact that the R–W model computes (to the extent possible) conditional contrasts in the first two classes of experiments was previously demonstrated in Cheng (1997, Appendices A and B) for equal α_i parameters and $\lambda = 1$. Almost certainly, these three classes of experiments do not form an exhaustive list of experiments with a perfect equilibrium. Determining such an exhaustive list remains an open research question.

11. Conclusion

The methods described in this paper allow researchers to determine the long-run behavior of the R–W model under a wide range of parameter values, and even a range of potential modifications to the model's structure, but always assuming that cases are presented to the subject in random order. Nevertheless, there are some significant open questions about the long-run behavior of the R–W model that remain. Some particularly interesting questions are as follows:

1. I gave a procedure to determine, for restricted conditions, which of infinitely many equilibria will actually occur when the cases occur in a random order. Is there a general procedure that makes no assumptions about the experiment for determining which of infinitely many equilibria will occur (under these restricted presentation conditions)?
2. Throughout this paper, I have assumed that the data are passed to the method in a random order—that is,

the patterns are independently distributed. In reality, almost no experiments are done this way. Typically, the cases are "chunked," in that subjects will see several of one pattern, then several of another pattern, and so on. Moreover, when there are infinitely many equilibria, chunking will play some role in determining the actual equilibrium. It is also known that, when cases are chunked, the initial weights can play a significant role in determining the long-run behavior. Can we give a characterization of the effect of chunking on the discussions of both the standard and augmented R–W models?

3. This chunking of cases helped prompt the augmented R–W models, since retrospective updating was first noticed in experiments to test backwards blocking, which requires chunking.[10] As a result, it has been claimed that the augmented R–W models give the "right" answer on a wider range of data than the standard R–W model. Is this claim true?

Acknowledgments

Thanks to Clark Glymour for an offhand comment that sparked this paper, as well as numerous discussions about both the content and presentation of the paper. Thanks to Patricia Cheng and an anonymous reviewer for *Journal of Mathematical Psychology* for their helpful critiques and suggestions. Thanks to Mara Harrell for comments on earlier versions of this paper. Portions of this paper previously appeared in a chapter in my Ph.D. Dissertation at the University of California, San Diego.

Appendix

Theorem. *For a particular experimental design, an equilibrium* **V** *of the R–W model consists of conditional contrasts* (*to the extent possible*) *if and only if* **V** *is a perfect equilibrium.*

Proof. We begin by defining a new concept. Define a *cue pattern graph* by the following construction: Write down all of the patterns that are shown in the experiment. Let r index the pattern pairs, $\langle p_i^r, p_{\sim i}^r \rangle$, that differ by only the presence or absence of a single cue C_i. For each r, and each i, connect p_i^r and $p_{\sim i}^r$ with a line whose associated value, denoted $A_{i,r}$, is the difference in outcome probabilities for p_i^r and $p_{\sim i}^r$. In other words, if two patterns, p_1 and p_2, differ only in that p_1 contains C_i and p_2 does not, then connect them with a line whose associated value is $P(O|p_1) - P(O|p_2)$. Consider each cue C_i. If there are r and s such that $A_{i,r} \neq A_{i,s}$, then we will say that the cue pattern graph is *invalid*, and the construction stops. Otherwise, we assign $W_i = A_{i,r}$. For those j such that $A_{j,r}$ is never defined (i.e., there are no two patterns that differ only in cue C_j), determine the range of values that are consistent with

$$\forall p \in ExpDesign(C_j) \left(\sum_{k \in Shown(p)} W_k = P(O|p) \right).$$

If there is a j such that there are no possible values for W_j, then the cue pattern graph is invalid and the construction stops.

If the cue pattern graph is connected,[11] then we are done. Otherwise, consider all pairs of maximally connected subsets.[12] Find the two patterns, p_i and p_j, in the different subsets whose symmetric difference[13] is least (or arbitrarily pick a pair, if there is more than one). Connect p_i and p_j by an undirected edge if and only if $\sum_{i \in Shown(p_1)} W_i - \sum_{j \in Shown(p_2)} W_j = P(O|p_1) - P(O|p_2)$, where we may have to consider a range of values for some of the W_j's. If the edge exists, associate with it the above difference.

An equilibrium defined by the W_i's of a valid, connected cue pattern graph will be composed of conditional contrasts (to the extent possible). Furthermore, note that the R–W model assumes that there is always a constant background cue. Hence, we need only the following lemma to finish the proof.

Lemma A.1. *Assume there is a constant background cue. There is a perfect equilibrium if and only if there is a valid, connected cue pattern graph, and that perfect equilibrium is given by the W_i's defined in the construction.*

Proof. ($\Rightarrow$) Assume that there is a particular perfect equilibrium **V**. To construct a valid, connected cue pattern graph, consider first all of the pairs of patterns that differ by only one cue. For all of these pairs of patterns that, for example, differ only in cue C_k, $p_k^r, p_{\sim k}^r$, we assign $A_{k,r} = P(O|p_k^r) - P(O|p_{\sim k}^r)$. Since **V** is perfect, we also know that, for all r, $A_{k,r} = \sum_{j \in Shown(p_k^r)} V_j - \sum_{j \in Shown(p_{\sim k}^r)} V_j = V_k$. Hence, for all cues C_j such that at least one A_j is defined, each A_j will be the same, and

[10] Backwards blocking involves showing two cues together with an outcome, and then showing just one of the cues with the outcome. Intuitively, the second block of data tells the subject that the cue shown only in the first block was not doing anything.

[11] A graph is connected if and only if, for every pair of variables X and Y, there is a path between X and Y.

[12] M is a maximally connected subset of S if and only if (i) M is connected, and (ii) adding any variable from $S \backslash M$ (the variables in S but not in M) to M would result in an unconnected set.

[13] The symmetric difference of two sets, A and B, is defined as the union of the elements of A that are not in B, and the elements of B that are not in A. Since we can think of patterns as sets of cues, this notion is well defined here. For example, if $p_1 = C_0 C_1 C_2$, and $p_2 = C_0 \sim C_1 C_2$, then the symmetric difference of the two patterns is $\{C_1, \sim C_1\}$.

will be equal to V_j. If this produces a connected graph, then we are done.

So, we can assume that we have multiple maximally connected subsets. Consider connecting any two of them. There will be an edge between the two patterns (whose symmetric difference is least) if and only if $\sum_{i \in Shown(p_1)} V_i - \sum_{j \in Shown(p_2)} V_j = P(O|p_1) - P(O|p_2)$. But since **V** is a perfect equilibrium, this is always true. Hence, we can always connect the subsets to produce a fully connected graph.

($\Leftarrow$) Assume there is a valid, connected cue pattern graph. Assume for *reductio* that the W_j's picked out by this cue pattern graph do not form a perfect equilibrium. In other words, assume that there is a pattern p such that $\sum_{i \in Shown(p)} W_i \neq P(O|p)$. Lemma A.2 will be used to show that this is impossible.

Lemma A.2. *If we have a connected, valid cue pattern graph, then the difference* $\sum_{i \in Shown(q)} W_i - P(O|q)$ *has the same value for all patterns* q.

Proof. Consider any arbitrary pattern q, and consider an adjacent pattern r along one of the edges that ends at q. There are three possible cases:

Case 1: C_i appears in q, but not in r. Then, since we have defined $W_i = P(O|q) - P(O|r)$, we have

$$\begin{aligned}\sum_{j \in Shown(q)} W_j - P(O|q) &= P(O|q) - P(O|r) \\ &\quad + \sum_{j \in Shown(q) - \{i\}} W_j - P(O|q) \\ &= \sum_{j \in Shown(r)} W_j - P(O|r).\end{aligned}$$

Case 2: C_i appears in r, but not in q. As in case 1, we can use the fact that $W_i = P(O|q) - P(O|r)$ to show that the differences are the same.

Case 3: q and r differ by more than one cue. Since the cue pattern graph is valid and there is an edge between q and r, it must be the case that there are W_i's such that $\sum_{i \in Shown(q)} W_i - \sum_{j \in Shown(r)} W_j = P(O|q) - P(O|r)$, and so immediately we have that the differences are equal.

Since we made no assumptions about the patterns q or r, and since there is a path (possibly of length greater than one) connecting any pair of patterns, the differences must all be equal. □

Using this lemma and the above assumption, we can conclude that, for all patterns p and q, $\sum_{i \in Shown(q)} W_i - P(O|q) = \sum_{i \in Shown(p)} W_i - P(O|p) = a \neq 0$. Now, since the cue pattern graph is valid, for every W_i, either an edge (or edges) determines it, or else there is a range of values that it could take on. The constant background cue appears in every pattern, and so there cannot be an edge that determines it.[14] However, by construction of the cue pattern graph, the range of values that W_0 can take on must satisfy $\forall p \in ExpDesign(C_0)\left(\sum_{k \in Shown(p)} W_k = P(O|p)\right)$. Therefore, for those patterns, the difference must be zero. Since we showed that the differences must be non-zero (on our *reductio* assumption), we have a contradiction. Hence, the solution picked out by the cue pattern graph must be a perfect equilibrium. □

References

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 414–432.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, *18*, 537–545.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Gluck, M. A., & Bower, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General*, *119*, 105–109.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*, 195–212.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla–Wagner learning rule? Comment on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1398–1410.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin*, *117*, 363–386.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H., & Prokasy, W. F. (Eds.), *Classical conditioning II*: *Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Shanks, D. R. (1995). Is human learning rational? *The Quarterly Journal of Experimental Psychology*, *48A*, 257–279.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In Bower, G. (Ed.), *The psychology of learning and motivation*, Vol. 21 (pp. 229–261). San Diego, CA: Academic Press.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin & Review*, *3*, 314–321.

Spellman, B. A. (1996). Conditionalizing causality. In Shanks, D. R., Holyoak, K. J., & Medin, D. L. (Eds.), *Causal learning: the psychology of learning and motivation*, Vol. 34 (pp. 167–206). San Diego, CA: Academic Press.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In Rumelhart, D. E., McClelland, J. L., &

[14] We cannot do without this assumption here, since otherwise there are probability distributions for which the differences are all equal and non-zero. Consider the following probability distribution: $P(O|AB) = 1$; $P(O|A \sim B) = 0.5$; $P(O| \sim AB) = 0.3$. Then, $V_A = 1 - 0.3 = 0.7$, and $V_B = 1 - 0.5 = 0.5$. But then $P(O|A \sim B) - V_A = P(O| \sim AB) - V_B = P(O|AB) - V_A - V_B \neq 0$.

the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1: Foundations (pp. 444–459). Cambridge, MA: The MIT Press.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–170.

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 193–204.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*, 127–151.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 174–188.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In Shanks, D. R., Holyoak, K. J., & Medin, D. L. (Eds.), *Causal learning: The psychology of learning and motivation*, Vol. 34 (pp. 207–264). San Diego, CA: Academic Press.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON convention record* (pp. 96–104). New York: IRE. (Reprinted in Anderson, J. A., & Rosenfeld, E. (Eds.) (1988). *Neurocomputing: Foundations of research* (pp. 123–134). Cambridge, MA: MIT Press.)

Yamaguchi, M. (1999). New methods for solving the Rescorla–Wagner model. *Behavior Research Methods, Instruments, & Computers*, *31*, 684–688.