http://dx.doi.org/10.1037/xan0000203

A Computational Implementation of a Hebbian Learning Network and its Application to Configural Forms of Acquired Equivalence

Jasper Robinson

School of Psychology

University of Nottingham

United Kingdom

Author Notes

Jasper Robinson, School of Psychology, University of Nottingham, UK.

David N. George, School of Life Sciences, University of Hull, UK.

Dietmar Heinke, School of Psychology, University of Birmingham, UK.

This work was supported a British Academy/Leverhulme Small Research Grant [SG163455].

We thank Sara Bru Garcia for her comments on a draft version of this report.

Submitted to: Journal of Experimental Psychology: Animal Learning and Cognition.

Submission Dates: August 30 2018 and January 15 2019

Manuscript Number: XAN-2018-1483.

Abstract

We describe and report the results of computer simulations of the three-layer Hebbian network informally described by Honey, Close, and Lin (2010): A general account of discrimination that has been shaped by data from configural acquired equivalence experiments that are beyond the scope of alternative models. Simulations implemented a conditional principle-components analysis (CPCA) Hebbian learning algorithm and were of four published experimental demonstrations of configural acquired equivalence. Experiments involved training rats on appetitive bi-conditional discriminations in which discrete cues, (w and x) signaled food delivery (+) or its absence (-) in four different contexts (A, B, C and D): Aw+ Bw- Cw+ Dw- Ax- Bx+ Cx- Dx+. Contexts A and C acquired equivalence. In three of the experiments acquired equivalence was evident from subsequent revaluation, from compound testing or from whole-/part-reversal training. The fourth experiment added concurrent bi-conditional discriminations with the same contexts but a pair of additional discrete cues (y and z). The ***congruent*** form of the discrimination, in which A and C provided the same information about y and z, was solved relatively readily. Parametric variation allowed the network to successfully simulate the results of each of the four experiments.

Honey, R., Close, J., & Lin, T. C. E. (2010). Acquired distinctiveness and equivalence: a synthesis. In C. Mitchell & M. Le Pelley. Oxford: Oxford University Press.

Keywords: Acquired equivalence, configural, discrimination, Hebbian, learning.

A Computational Implementation of a Hebbian Learning Network and its Application to Configural Forms of Acquired Equivalence

Behavior established to a stimulus may also be provoked, albeit at a reduced magnitude, by a second stimulus (see, e.g., Guttman & Kalish, 1956; Hanson, 1959). Knowledge of this stimulus generalization has been used to develop psychological models, whose scope includes descriptions of stimulus representation, discrimination learning and perceptual learning (e.g., Hall, 1991; Harris, 2006; McLaren & Mackintosh, 2002; Rescorla, 1976). Although these models differ in important ways, they share their conception of stimulus representation: within the limits of the organism's sensorium, stimuli will be elementally coded and generalization based on stimulus similarity. For example, high and low-pitched tones, may be conceived of as being represented by a finite number of representational elements, which are partially overlapping. Subjectively or physically similar stimuli are assumed to have a relatively great proportion of overlapping stimulus elements and this will afford relatively good stimulus generalization. For dissimilar stimuli, which are well discriminated (i.e., stimulus generalization is weaker), the proportion of overlapping elements is assumed to be less.

Honey and Hall (1989) reported findings that challenged this standard view of stimulus generalization. The design of one of their experiments is summarized in Table 1. Two groups of rats received discrimination training with three auditory stimuli, A, B and N, with Group N+'s discrimination having the form: A+ B- N+ and Group N-'s discrimination having the form: A- B+ N-. Food was delivered ("+") on termination of some trials but not on others ("-"). Notice that: the groups differed in whether or not N was food reinforced; the pattern of reinforcement and non-reinforcement across A, B and N was complemented in the two groups; for both groups, A and N were treated ***equivalently***. Following appetitive discrimination training, rats received aversive training in which N was used to signal a foot-shock; no food pellets were delivered. After shock training a free-operant instrumental response was established to assess generalized conditioned suppression to A and to B. Because A and B were counterbalanced across the rats within each of the groups, standard stimulus-generalization accounts predict that there will be no difference in generalized conditioned responding to A and B. Or put another way, on average, A and B may each be regarded as equally physically similar to N. However, both groups demonstrated an "acquired equivalence" effect: suppression from N generalized better to A than to B. Thus, the common appetitive training history of A and N appeared either to have enhanced generalization between them, or the difference in appetitive training between B and N had reduced stimulus generalization ("acquired distinctiveness").

(*) TABLE 1 ABOUT HERE PLEASE (*)

Honey and Hall (1989) also suggested a mechanism that can accommodate acquired equivalence within the framework of elemental accounts of learning. Holland (e.g., Holland, 1990; 2008); (see also, Ward-Robinson & Hall, 1998) demonstrated that the associatively activated representation of a stimulus can replace a conditioned stimulus in Pavlovian conditioning. In Honey and Hall's Group N+, the presentation of N during aversive conditioning would provoke the associatively activated representation of the food pellet, that could enter into association with the foot-shock. On test, A would elicit greater suppression than B because it would associatively activate a representation of the, now-aversive, food pellet. This mediated-conditioning process could happen in addition to any stimulus generalization based on the overlapping elements that were common to N and to A and to N and to B. Ward-Robinson and Hall (1999) replicated Honey and Hall's experiment and found direct evidence for the mediated conditioning account (see also, e.g., Hall, Mitchell, Graham, & Lavis, 2003).

In this report we summarize specialized, configural acquired equivalence experiments that are not amenable to the mediated conditioning analysis and are beyond the scope of two important accounts of configural learning (e.g. Brandon, Vogel, & Wagner, 2000; Pearce, 2002), which we examine in the General Discussion. These extant experiments all used appetitive, context-based procedures with rats (e.g., Coutureau et al., 2002; Honey & Ward-Robinson, 2001; Honey & Ward-Robinson, 2002; Honey & Watt, 1998; Honey & Watt, 1999; Iordanova, Killcross, & Honey, 2007; Ward-Robinson & Honey, 2000). An alternative, three-layer network analysis has been proposed (e.g., Honey & Ward-Robinson, 2002) to explain these special cases of acquired equivalence. The network is assumed to operate in much the same way as traditional three-layer Hebbian networks, with the modification that ***output*** units operate on the hidden layer in the same way as input units do. This modification produces a feedback signal that trains the network's hidden layer. As a result, hidden units come to be shared when their input units are correlated with activity in other input and output units. This allows the model to account for these special, configural forms of acquired equivalence and make some novel predications. Until now, the model has only been described informally but we describe a new, formal version of the model and its application to four examples of configural acquired equivalence. We were particularly interested in understanding whether the predictions derived from the informal description could be confirmed with our computational implementation. We were also interested in uncovering any novel predictions and important qualifications, such as the model’s failure to explain data except under a restricted range of simulation parameters.

Four demonstrations of configural acquired equivalence

1. Revaluation

Based on an experiment by Honey and Watt (1998), Ward-Robinson and Honey (2000) reported acquired equivalence using the experimental design summarized in Table 2 (first row). Their design is similar to Honey and Hall's (1989; see Table 1): Acquired equivalence is established in an initial appetitive stage of training before one of the stimuli is aversively revalued. And, again, equivalence is assessed by comparing generalization of the aversive response to the stimulus sharing the now-aversive stimulus' appetitive training history. The principle difference is in the use of its initial configural discrimination to establish acquired equivalence, rather than a simple discrimination. In Ward-Robinson and Honey's experiment, a single group of rats was trained in four Skinner boxes with differently patterned walls (A-D). Two brief, auditory stimuli (w and x) signalled food and its absence in the four contexts. The task was a pair of concurrent biconditional discriminations with the forms: Aw+ Bw- Ax- Bx+ and Cw+ Dw- Cx- Dx+), where "+" and "-", respectively represent ***Food*** and ***No-Food***. Notice that each context and auditory stimulus equally often signalled both outcomes and that it was the ***configuration*** of the specific context and auditory stimulus that indicated the outcome on any particular trial. Notice also that there were two pairs of equivalent contexts: A and C and B and D. Following mastery of the appetitive discrimination, all rats received trials on which contexts A and B were presented in the absence of the auditory cues w and x (Stage 2). A foot-shock was presented in context A but not in context B. Generalization of fear, measured by freezing behavior was assessed in a test of contexts C and D. Fear from context A better generalized to context C than to context D: the acquired equivalence effect.

(*) TABLE 2 ABOUT HERE PLEASE (*)

We noted above that the acquired equivalence effect reported by Honey and Hall (1989) is explicable as a form of mediated conditioning (e.g., Holland, 2008) in which a representation of the food reinforcer (or its absence) was elicited and entered into an association with the foot-shock during revaluation. This analysis is an inadequate explanation of configural demonstrations of acquired equivalence (Honey & Watt, 1998; Ward-Robinson & Honey, 2000) because it requires context A to elicit a representation of, say, food during stage-2 revaluation and context C to elicit the same representation of the food on test. It is unclear that context A would elicit the representation of ***Food*** any more than it would elicit the representation of ***No-Food***, the two outcomes were equally occurrent outcomes during appetitive training. But even if, say, the food representation alone was able to enter into association with the shock, on test, context C would be no more likely than context D to re-elicit the now-aversive representation of food. Thus, the greater responding to context C than to context D cannot be based upon mediated conditioning.

This finding has been described in terms of the operation of a three-layer Hebbian network, summarized in Figure 1 (e.g., Honey, Close, & Lin, 2010; Honey & Ward-Robinson, 2002). The analysis of this finding will be described below with reference to the computational implementation. According to this account, a layer of input units that codes for the context (A-D) and auditory cues (w and x), is activated immediately upon application of each particular stimulus. Each input unit is connected to each unit in a second, hidden, layer of units. The initial weighting of each connection is given a small, random, value; but, with training, co-occurrent input units will develop stronger connections with the same hidden units. During early stages of the training the input-to-hidden unit weightings will be random but weak. A pair of input units may initially generate activity in the same hidden units or they may begin training activating separate hidden units. A pair of equivalent input units, for example, A and C, could solve the discriminations involving w+ using separate hidden units but it is possible for them to share a hidden unit, the basis of the acquired equivalence phenomenon in this implementation. This process occurs because, for example, on an Aw+ trial at a sub-asymptotic level of training, the input units will generate some activity in three hidden units that code for trials that include those elements (i.e., "ACw+", "ACx-", and "BDw-"). The further changes in input-hidden layer weightings are determined simply by the units' contiguous activation (cf., Brandon et al. 2000; Hebb, 1949). However, the "correct" hidden unit, ***ACw+***, enjoys greater temporal overlap with the ***A*** and ***w*** input units because it receives the additional activation from the output unit, for ***Food*** (*+*). The pattern of connections and number of hidden units that are recruited will vary from simulation to simulation but we can think of an idealized solution using ***four*** hidden units, because contexts A and C and B and D will share theirs. This creates the acquired equivalence effect.

(*) FIGURE 1 ABOUT HERE PLEASE (*)

The transfer of fear responding from Stage-2 of Ward-Robinson and Honey's (2000; see also Honey & Watt, 1998) procedure is explained by the network by first assuming that the presentation of context A generates partial activation in hidden units ACw+ and ACx-, see Figure 1. This allows their association with the new foot-shock outcome (not shown in Figure 1). Testing with context C will provoke partial activation of same, fear-eliciting hidden units, which will not be elicited by context D.

2. Congruent/incongruent context combinations

Our starting point for understanding the effects reported by Honey and Ward-Robinson (2002; see Table 2, second row) is the state of the network at the conclusion of Stage-1 training on the appetitive bi-conditional discriminations (see lower panel of Figure 1). Honey and Ward-Robinson gave rats a pair of appetitively reinforced bi-conditional discriminations with visual (A and B) or thermal (C and D) contexts and auditory stimuli (w and x) before testing magazine entry during visual-thermal (e.g., AC or AD) combinations of the contexts in the absence of w and x (see also, Hodder, George, Killcross, & Honey, 2003)

Each of the network’s input units partially activates a pair of hidden units. Activation by a single input unit is only sub-threshold but the combined force of pairs of inputs is sufficient to trigger appropriate activity in the hidden layer. Notice that there is no special status given to any of the input units in their governance of performance in the bi-conditional discriminations: inputs simply activate their hidden unit within the scope of their connections' weightings. And because acquired equivalence reported by Ward-Robinson and Honey (2000) and by Honey and Watt (1998) indicates that input units, for example, A and C, operate on the same hidden unit (i.e., ACw+ and ACx-), Honey and Ward-Robinson (2002) reasoned that the "congruent" presentation of A and C, together but in the absence of w or of x, would summate in their activation of the ***ACw+*** and ***ACx-*** hidden units. Because the two hidden units provoke activity in, respectively, ***Food*** and ***No-Food*** output units, and because neither ***w*** nor ***x*** is present to activate either hidden unit more than the other, rats' patterns of activity were predicted to be relatively variable: some rats may strongly anticipate food; others may strongly anticipate no-food. This prediction was supported by the observation that the mean absolute deviation of the rates of appetitive behavior (magazine activity) was relatively great. In contrast, "incongruent" visual-thermal context combinations, produced less variable responding. Incongruent pairs of visual-thermal contexts were assumed to activate four ***different*** hidden units incompletely. Again, the anticipated ***Food*** and ***No-Food*** outcome units were expected to be evenly split; but the input layer's division of activity across the hidden layer would produce only sub-threshold activity, less likely than the congruent combinations to trigger activity in each hidden unit. Thus, the even split of ***Food*** and ***No-Food*** output activation, is muted in the incongruent context combination trials and variability is less extreme than for the congruent context combination.

3. Whole/part reversal

Figure 1 indicates that the learning of a pair of bi-conditional discriminations will produce acquired equivalence by the sharing of hidden units by context input units whose auditory cue-outcome combinations match. The tuning of the hidden layer to achieve this is a crucial part of the discrimination's solution. The other significant feature of the solution is the accuracy of each hidden unit's selection of its output unit. Honey and Ward-Robinson (2001; see Table 2, third row) gave groups of rats a reversal treatment, after their mastery of the pair of bi-conditional discriminations. For group Whole Reversal, the outcomes ***Food*** and ***No-Food*** for each of the eight trial types were reversed; for group Part Reversal, only one pair of the bi-conditional discriminations (i.e., four of the eight trial types) was reversed. From one point of view, recovery of performance in Part Reversal should be most quickly established because less new knowledge is required: fewer of the hidden unit → output unit connection weightings require modification. In fact, this discrimination was solved more slowly that group Whole Reversal's discrimination, a widely reported finding (e.g., Delamater & Joseph, 2000; Honey & Ward-Robinson, 2002; Nakagawa, 1986; Robinson & Owens, 2013; Zentall, Sherburne, Steirn, Randall, & Roper, 1992; Zentall, Steirn, Sherburne, & Urcuioli, 1991)

Honey and Ward-Robinson (2001) maintained that the Whole Reversal would retain the original hidden layer structure, while only connections to the outcomes would require re-learning. But the hidden layer for the Part Reversal would require restructuring because hidden units that had previously shared context inputs were no longer equivalent; for example, context A was equivalent to context C in the original discrimination but, after the reversal, context A was equivalent to context D.

4. Congruent/incongruent acquisition

Honey and Ward-Robinson (2001) reported a second study that employed a similar logic to their whole/part reversal study (see Table 2, fourth row). Two groups of rats, Congruent and Incongruent, received an expanded version of the bi-conditional discriminations used in other experiments (e.g., Honey & Watt, 1998; Ward-Robinson & Honey, 2000), which used a pair of visual stimuli (steady or pulsed lamp illumination), y and z, in addition to the auditory stimuli, w and x. Notice from the design for the Congruent group's treatment retains the equivalence relationships of the previous experiments: contexts A and C and contexts B and D give equivalent information about the discrete stimuli, w-z, and their outcomes. Notice also that equivalence relationships appear in the Congruent group's treatment with regard to the discrete stimuli: w and y share outcomes when presented in the same contexts as do x and z. Honey and Ward-Robinson found rats’ performance of the congruent discrimination to be superior to that of the incongruent discrimination (see also, e.g. Delamater & Joseph, 2000; Honey & Ward-Robinson, 2002; Nakagawa, 1986; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991).

The network anticipates that this Congruent discrimination will be solved by the tuning of only ***four*** hidden units, like that in Figure 1, receiving input from a pair of contexts ***and*** a pair of discrete cues. That is, these hidden units might be represented as: ACwy+, ACxz- BDwy-, BDxz+. During training, hidden unit ➔ output learning occurring on one trial (e.g., Aw+) would benefit ***three*** other trial types too (i.e., Ay+, Cw+ and Cy+), thereby accelerating learning.

The group Incongruent's discrimination is entirely comparable to group Congruent's in that it, too, includes sixteen trials types in the form of bi-conditional discriminations. But notice that no pair of contexts, nor either pair of discrete stimuli is equivalent: each pattern of food (+) and no-food (-) in the contexts (rows) and the discrete stimulus (columns) is unique. The network assumes that this discrimination will be solved more slowly because ***eight*** hidden units are required (i.e., either Awy+, Axz-, Bxz+, Bwz-, Cwz+, Cxy-, Dxy+, Dwz-, or ACw+, BDw-, BDx+, ACx-, ADy+, BCy-, BCz+, ADz-). Notice also that hidden unit → output learning on one trial (e.g., Aw+) can benefit only ***one*** other trial type (e.g., Ay+), not the ***three*** trial types that benefit in the congruent discrimination.

Model Description

The Hebbian network has been carefully described (e.g., Honey, Close, & Lin, 2010; Honey & Ward-Robinson, 2002) and supplies plausible accounts of both the phenomena that it was designed to account for and those that it predicted. However, its dynamic and interacting ingredients raise the possibility that extant verbal accounts could be prone to some unforeseen error. The current report describes one possible version a formal, computational account of the Hebbian model and describe its successes and failures in modelling acquired equivalence data. The simulations were run using MATLAB (MathWorks Inc., Natick, MA) programs written by one of the authors and are available for download from the 'HebbianNN' repository on GitHub (GitHub Inc., San Francisco, CA) at https://github.com/DavidNGeorge/HebbianNN.

The network consisted of three layers of units: an input layer in which individual units represented the discrete stimuli and contexts used in the experiments; a hidden layer of units; and an output layer in which units represented the outcomes of various conditioning trials (e.g., food or no-food). Activation of input and output units by stimuli and outcomes was binary; they were either on with a value of 1, or off with a value of 0. There were feed-forward connections between successive layers (input-to-hidden and hidden-to-output), as well as feed-back connections from the output layer to the hidden layer. Each conditioning trial consisted of four phases. First, the input and output units corresponding to the appropriate stimuli and outcome were clamped on (i.e., set to their maximal value 1). Second, activity from these units was propagated through their projections to the units in the hidden layer. Third, a form of winner-takes-all (WTA) competition was applied to the units within the hidden layer in order to increase the contrast between activity in different units. Because of this competition, in a well-trained network, a single unit would become fully active while all other units would have minimal activity. Fourth, weights between all units in adjacent layers were updated according to a conditional principle-components analysis (CPCA) Hebbian learning algorithm. Probe test trials were also conducted in order to generate predictions that could be compared to the behavior of animals in the experiments that were simulated. On these trials, only units in the input layer were clamped on. Following propagation of activity from the input layer to the hidden layer, and the application of WTA competition at the hidden layer, activity was propagated from the hidden layer via its projections to the output layer. WTA competition was applied to the output layer units in the same manner as for hidden layer units. This outcome may be seen as a partial version of the system of mutually exclusive "antinodes" described by Konorski (1967).

Activity in hidden and output units (when not clamped on) was directly proportional to the activity in units that projected to them multiplied by the strength of their connection. That is, these units had linear activation functions. Equation 1 shows how the activation level, y_j, of hidden unit j was determined by activation of input units and output units, where x_i is the activation of input unit i, z_k is the activation of output unit k, w_{ij} is the weight of the connection between input unit i and hidden unit j, and w_{kj} is the weight of the connection from output unit k back to hidden unit j.

$$y_j = \sum_i x_i w_{ij} + \sum_k z_k w_{kj} \qquad (1)$$

WTA competition was applied to hidden (and, on probe test trials, output) unit activation to enhance the selectivity of these units using Equation 2. The activity of a unit, $\boldsymbol{y_j}$, was converted to a proportion of the most active unit within the layer, y_{max}, and raised to the fourth power. For example, initial values of $\boldsymbol{y_j}$ and $\boldsymbol{y_{max}}$ of, .3 and .6 would become, respectively .0625 ($[.3/.6]^4$) and 1 ($[.6/.6]^4$). Because of this competition, the activity level of the most active hidden unit was always equal to 1.

$$y_j = \left(\frac{y_j}{y_{\max}} \right)^4 \tag{2}$$

Weight changes in all three layers of the network were governed by the conditional principle component analysis (CPCA) learning algorithm shown in Equation 3. Here, Δw_{ij} is the change in the weight of the connection between input unit i and hidden unit j. ε is a learning rate parameter. It had a fixed value during simulations for each set of connections and was restricted in the range: $0 < \varepsilon \leq 1$.

$$\Delta w_{ij} = \varepsilon \left[y_j \left(x_i - w_{ij} \right) \right] \tag{3}$$

The CPCA algorithm calculates the conditional probability that the sending unit, i — from either the input- or the output-layer — is active given that the receiving-unit, j — from either the hidden or the output layer — is active. Hence, when unit j is inactive no change will be made to the connection weight. When receiving-unit j is active, the connection weight will move in the direction of the sending unit activation. For example, in a network with initially low, random, connection weights between units, when unit i and j are both active the weight w_{ij} will increase. If sending-unit i is inactive and receiving unit j is active, then w_{ij} will decrease.

A limitation of the CPCA algorithm is that weights have a restricted dynamic range and so they do not lead to strong differentiation between input patterns. To maximize the network's differentiation between input patterns, weights between uncorrelated units in adjacent layers should be equal to approximately .5, with weights between positively correlated units being greater than .5 and weights between negatively correlated units less than .5. With the CPCA algorithm, however, the strength of weights between uncorrelated units is dependent upon the sparsity of activity within layers of units. Equation 3 can be re-written as Equation 4:

$$\begin{aligned}\Delta w_{ij} &= \varepsilon\left[y_j x_i - y_j w_{ij}\right] \\ &= \varepsilon\left[y_j x_i\left(1-w_{ij}\right)+y_j\left(1-x_i\right)\left(0-w_{ij}\right)\right]\end{aligned} \tag{4}$$

The first term in Equation 4 has the effect of increasing the weight strength towards the maximal value of 1, whereas the second term decreases the weight towards the minimal value of 0. To compensate for sparseness of activity within a network layer, we increased the maximum value of the weight strength by replacing the value 1 in the first term of Equation 4 with a parameter, *m*, in Equation 5. The value of *m* is determined for each set of weights according to Equation 6, where α represents the average sparsity of activity across all sending units. For the input and output layers, α is the average proportion of input and output units active on each trial, respectively. Because WTA competition was implemented at the hidden layer, α for projections from the hidden layer to the output layer was equal to 1 divided by the number of hidden units.

$$\Delta w_{ij} = \varepsilon\left[y_j x_i\left(m-w_{ij}\right)+y_j\left(1-x_i\right)\left(0-w_{ij}\right)\right] \tag{5}$$

$$m = \frac{.5}{\alpha} \tag{6}$$

Although the discriminations described here required as few as four hidden units for simulations, the network was preconfigured for eight hidden units. This does not create any important departure in the functioning of the network from that described earlier; it merely means that, for example, a ***pair*** of hidden units work as a single hidden unit in representing the unique combination of two input units. Connection weights at the beginning of each simulation were determined randomly from a uniform distribution in the range 0 – 1. Uniformly distributed random noise was added to the activation level of non-clamped units at the beginning of each trial. For all simulations reported here, the noise ranged between 0 and .05. The addition of noise reduces the occasional tendency of the network to recruit a very small number of hidden units to represent many input patterns. High levels of noise, however, restrict the ability of the network to learn the mapping between input and output patterns.

The learning-rate parameter, ε, was varied systematically to capture performance on the four forms of configural acquired equivalence phenomenon summarized in Table 2. Values ranged from .05 to .25 at different points of the network. We do not give an exhaustive description of the results of all values of ε but focus on values that either permitted or prevented the Honey network from successfully capturing results.

Unless otherwise specified, results of simulations for each network configuration were averaged over 1000 simulated networks with different randomly assigned initial connection weights, each over either 50 or 100 epochs of training. We did not consider here any special performance rules to translate simulated output activity into conditioned responses; the reader may take the generation of explicit responding to be monotonically related to the activity of the corresponding output unit.

Results

1. Revaluation

The network was applied to the Stage-1, appetitive discrimination summarized in Table 2 (first row). The weight strength for the connections from the input layer to the hidden layer, following only a single simulation run, are displayed in Table 3. The pattern of weight strengths match those of Figure 1 and allow successful solution of discrimination. Although only four hidden units are required for the discrimination, our simulations employed eight. Notice that relatively high weightings occupy connections between the inputs for stimulus ***w*** and ***x*** at ***complementary*** hidden units, for stimulus w: 1, 3, 4, and 7; and for stimulus x: 2, 5, 6, and 8. The largest four weightings for both context A and context C occupy hidden units 1, 3, 5, and 8. And this pattern of hidden unit weightings is almost ***complemented*** for contexts B and D (i.e., hidden units 2, 4, and 7). Thus, the contexts appear to be becoming either equivalent or distinct. Context A and stimulus w's hidden units of greatest weight connections are 1 and 3 and these may be taken to be one of the pair of notional hidden units described above (e.g., ACw+). The hidden units that context A and stimulus x are maximally weighted at are 5 and 8 and these may be taken to be the alternative hidden unit (e.g., ACx-). Notice also that hidden unit 6, “incorrectly” responds to context A. This was a temporary feature of the sub-asymptotic training given, which is also evident in unit 6’s relatively poor weighting discrimination, relative to its partner unit, unit 2. Inspection of the Table's weightings uncovers similar correspondences to informal descriptions of the Hebbian network (e.g., Honey, Close, & Lin, 2010; Honey & Ward-Robinson, 2002). Each input unit, whether it represents a context or a stimulus, forms a strong connection with approximately half of the hidden units. Because each stimulus is present on twice as many trials as is each context, the conditional probability that a particular stimulus is present when a given hidden unit is active is twice the conditional probability that any individual context is present (cf. Equations 3 – 6). It is for this reason that the connection weights between stimulus w and stimulus x and the appropriate hidden units are about twice as strong as the connection weights for each context. That some connection weights are greater than 1 reflects the influence of the weight renormalization described by Equations 5 and 6.

(*) TABLE 3 ABOUT HERE PLEASE (*)

Simulations were conducted in which the network was given a total of 50 epochs of Stage-1 training and then 2 epochs of the Stage-2, aversive revaluation: A → Shock, B → No-Shock. The network was subsequently tested with presentations of the four contexts in turn, and in the absence of either stimulus w or stimulus x. Activation of each of the four output units (***Food***, ***No-Food***, ***Shock***, ***No-Shock***) in the presence of each context are shown in Table 4. These figures are averaged over three sets of 1000 simulation runs in which the learning rate parameter, ε, for different sets of connections was manipulated. Irrespective of the values of ε, the network showed Stage-2 discrimination, which transferred to the equivalent contexts. The top row of Table 4 shows the results of simulations in which *ε* = .10 for all three sets of connections. ***Before Stage-2 revaluation***, each context resulted in activity (≥ .62) in both the ***Food*** and ***No-Food*** units, reflecting their involvement in the Stage-1 appetitive discrimination (i.e., they are substantially greater than zero); but, without the input from the Stimuli ***w*** and ***x***, they are undifferentiated with regard to those units. Activity in the ***Shock*** and ***No-Shock*** output units are also undifferentiated before Stage-2 training but, because aversive training had not then occurred, activity is negligible (≤ .05).

(*) TABLE 4 ABOUT HERE PLEASE (*)

The Stage-2, aversive revaluation produced a slight reduction in activity in the ***Food*** and ***No-Food*** units in response to the presentation of each context. More importantly, the Stage-2 training appropriately adjusted the connection weights of contexts A and B: Each generated more activity in the trained outcome unit (≥ .47), than in the alternative unit. An analysis of variance (ANOVA) confirmed this description of the summary data: Neither overall activity in response to contexts A and B, ***F***(1, 999) = 1.3; ***p*** > .239, nor, activity in the outcome units for Shock or ***No-shock***, ***F*** < 1, differed but the interaction between these main effects was reliable, ***F***(1, 999) = 4493.0; ***p*** < .001.

The network also exhibited acquired equivalence: The Stage-2 discrimination is mirrored in the acquired equivalence test to contexts C and D. Activity was higher in the ***Shock*** outcome unit than in the ***No-Shock*** unit for context C, whereas the reverse was true for context D. ANOVA yielded no main effects of outcome or context but a reliable interaction between those main effects, ***F***(1, 999) = 3559.4; ***p*** < .001.

Discrimination and acquired equivalence were also found with the two other sets of learning rate parameters. The central row of Table 4 displays the corresponding data from simulations run with ε = .10, ε = .20, and ε = .20 for the input-to-hidden layer, hidden-to-output layer, and output-to-hidden layer projections, respectively. The left columns show that, before Stage-2 revaluation, there was no differentiation in the activity in the

Food and ***No-Food*** output units in response to presentation of each of the four contexts.

Following context A → Shock, context B → No-Shock discrimination training, the response of the ***Food*** and ***No-Food*** output units to each context was decreased. Context A generated strong activity in the ***Shock*** output unit (.93) and no activity in the ***No-Shock*** outcome unit. Context B generated the opposite pattern of activity. ANOVA confirmed that there was no effect of output (***Shock*** vs. ***No-Shock***) or of context (A vs. B), *F*s < 1, but a reliable interaction between those variables, ***F***(1, 999) = 62290.4; ***p*** < .001. Importantly, contexts C and D, which has received no training with the ***Shock*** and ***No-Shock*** outputs generated equivalent patterns of activity to contexts A and B, respectively. Again, ANOVA yielded neither outcome nor context main effects, ***F*** < 1, but a reliable interaction between these factors, ***F***(1, 999) = 26401.1; ***p*** < .001.

The bottom row of Table 4 displays the corresponding data using simulations with even more exaggerated differences in learning-rate parameters: Respectively, input-to-hidden layer, ε = .05; hidden-to-output layer, ε = .25; and, output-to-hidden layer, ε = .25. The discrimination between contexts A and B and the ***Shock***/***No-Shock*** outcomes was reflected in a reliable Context x Outcome interaction, ***F***(1, 999) = 4493.0, ***p*** < .001. There was no overall main effect of context, ***F*** < 1, nor of outcome, ***F***(1, 999) = 1.3; ***p*** > .249. Acquired equivalence was successfully simulated. Results for contexts C and D indicated acquired equivalence, with context C and D mirroring their equivalent contexts' revalued outcomes. This was reflected in Context x Outcome interaction, ***F***(1, 999) = 3559.4; ***p*** < .001. There was no overall main effect of context or of outcome, ***F***s > 1.

The acquired equivalence effect was similar in the second pair of simulations and improved relative to the first. This is most simply determined by comparing the difference in activity in the ***Shock*** output unit between context C and context D. By these means the top, middle and bottom simulations of Table 4 yield acquired equivalence discriminations of, respectively: .39, .87, and .88. The finding that discrimination values doubled in the second pair of discriminations relative to the first is simply understood because they echo the corresponding results for context A and B. In particular, with the amount of training given, the top discrimination produced limited learning. The higher learning-rate parameters for the reciprocal connections between the hidden and output layers in the other two sets of simulations allowed improved context A → Shock learning, which improved scope for generalization to context C in the acquired equivalence test. Notice also that as the context A → Shock, context B → No-Shock discrimination improves across the three sets of simulations, activity in the ***Food*** and ***No-Food*** output units in response to each context decreases. This is the result of new weight changes between these output inputs and the hidden units during revaluation. The relative size of the discrimination between directly-conditioned contexts A and B is only slightly larger than that between contexts C and D. Empirically derived acquired equivalence results are less distinct than this. However, a *quantitative* comparison could be misleading because of, for example, the lack of specification about the translation of the simulated learning into behavior. Instead, it is the *qualitative relationships* between the contexts and their outcomes that are most meaningful, and they mirror empirically derived data. We examined two more adapted versions of this network. One was a repeat of the previous simulations but with the weightings from the output to the hidden layer units clamped off. Acquired equivalence was abolished once this source of feedback from the output layer to the hidden layer was removed. We can be confident, therefore, that this is the crucial feature of the networks ability to produce configural acquired equivalence. The second variant of the main simulation was of a non-configural acquired equivalence (cf., Honey and Hall, 1989), in which stimuli w and y were absent and contexts A and C, predicted food and contexts B and D predicted no-food during stage-1 training. The transfer of shock learning from stage 2 to contexts B and D during the test was very low – a consequence of our not including the common elements necessary to mediated primary stimulus generalization. However, the shock outcome activity was greater to context C than to context D – that is, the standard, non-configural acquired equivalence findings. This means both configural and non-configural acquired equivalence can be accommodated by this simulation.

Food Revaluation. The simulations reported here were based on an experimental design from Ward-Robinson and Honey (2000), which used a foot-shock revaluation. Acquired equivalence has also be reported with revaluation using ***the same food reinforcer*** as in the initial discrimination (e.g., Coutureau et al., 2002; Honey & Watt, 1999; Iordanova et al., 2007). Our examination of the aversive acquired equivalence simulations indicate that revaluation can interfere with Stage-1 learning and it seems possible that this would be yet more marked when the same Food/No-Food outcomes are re-used in revaluation. This would undermine our simulation of the Hebbian network in its departure from the empirical findings. We, therefore, simulated the same acquired equivalence experiment but used the same Food and No-Food outcomes for both the initial equivalence discrimination and the revaluation. The simulations were otherwise identical to those described here for shock revaluation and their results are in Table 5. The top row shows results for the acquired equivalence simulation with learning-rate parameters of ε = .10 at all three sets of connections. Appetitive revaluation of context A and B resulted in the appropriate discrimination. The Context x Outcome interaction was reliable, $\boldsymbol{F}(1, 999) = 32761.7$; $\boldsymbol{p} < .001$, but neither main effect was. The main effects of outcome and context were, respectively, $F(1, 999) = 3.2$, $p > .069$; and, $F(1, 999) = 2.8$, $p > .099$. Unlike the aversively revalued acquired equivalence simulations, this simulation resulted in ***a reverse acquired equivalence effect***, with context D generating the greater activity in the ***Food*** output unit. Unlike the previous simulations there was a small bias in activity toward the ***Food*** output unit over the ***No-Food*** output unit, $F(1, 999) = 5.7$; $p < .018$. Context C also generated more overall activity across output units that did context D, $F(1, 999) = 4.05$; $p < .005$. These two variables also reliably interacted, $F(1, 999) = 179.5$, $p < .001$.

(*) TABLE 5 ABOUT HERE PLEASE (*)

The center row of Table 5 shows the simulations results where the learning-rate parameters were as follows: input-to-hidden layer, $\varepsilon = .10$; hidden-to-output layer, $\varepsilon = .20$; and, output-to-hidden layer, $\varepsilon = .20$. Here activation of the ***Food*** and ***No-Food*** output units in the presence of contexts A and B interacted, $F(1, 999) = 126897.9$, p < .001, indicated the expected discrimination established by revaluation. Neither the context main effect nor the Outcome main effect was reliable, both ***F***s < 1. Here, a reliable acquired equivalence effect was evident from the output activations for context C and context D. These variables interacted reliably, $F(1, 999) = 26.6$; $p < .001$ but both constituent main effects were unreliable, both ***F***s < 1. Although a reliable acquired equivalence effect was obtained in this simulation its magnitude, for example the absolute difference in activations between context C and context D, was conspicuously smaller than in the aversive simulations.

The bottom row of Table 5 shows the results of the simulations above but with the learning-rate parameters of: input-to-hidden layer, $\varepsilon = .05$; hidden-to-output layer, $\varepsilon = .25$; and, output-to-hidden layer, $\varepsilon = .25$. The context and Outcome main effects, reflecting the Stage-2 revaluation, were unreliable, both ***F***s < 1 but the interaction between those variables was reliable, ***F***(1, 999) = 36558.8; ***p*** < .001. As in the previous appetitively-revalued simulation, but not the first, there was an acquired equivalence effect: This was evident in a reliable Context x Outcome interaction for contexts C and D, ***F***(1, 999) = 906.5; ***p*** < .001. Neither of the constituent main effects was reliable, both ***F***s < 1. Thus, unlike the aversively revalued acquired equivalence simulations, the appetitive revaluation simulations were parameter dependent, producing an acquired equivalence results, matching empirical reports (Coutureau et al., 2002; Honey & Watt, 1999; Iordanova et al., 2007) and also a ***reverse acquired equivalence*** effect.

Overtraining revaluation. Sensory preconditioning (Brogden, 1939), like acquired equivalence uses a three-stage procedure to demonstrate learning about the co-occurrence of relatively neutral stimuli. For example, Ward-Robinson and Honey (2000) gave rats presentations of an auditory and a thermal stimulus. Subsequently, the auditory stimulus served as the conditioned stimulus for a foot-shock. The thermal stimulus, despite never being paired with the shock, elicited freezing behavior, indicating some form of learning about the initial audio-thermal co-occurrence. Rescorla (1983) has shown that additional stage-2 revaluation ***reduces*** the sensory preconditioning effect. In light of this paradox, we thought it important to investigate parallels with acquired equivalence. The appetitively reinforced simulation with weights of ε = .10 at the input-to-hidden layer and εs = .20 at the other two layers was run for another two sets of 1000 simulations runs with four and with five epochs of revaluation. The results of these simulations appear, respectively, in the center and bottom rows of Table 6. The top row repeats the center row from Table 5 which uses only two epochs of training during revaluation. The two-epoch appetitive revaluation simulation produced a modestly sized but reliable acquired equivalence effect (statistical analysis is reported above). By contrast, increasing the number of epochs of revaluation training either abolished or reversed the acquired equivalence effect. With four epochs of revaluation the explicitly trained context A/context B discrimination produced no main effect of context, ***F***(1, 999) = 1.4; ***p*** > .229, and no main effect of outcome, ***F*** < 1. But the interaction between those variables was reliable, ***F***(1, 999) = 1.98 x 10^{8}; ***p*** < .001. In the transfer test with contexts C and D, there was no evidence of acquired equivalence with neither the context main effect nor the Outcome main effects reaching reliability, both ***F***s < 1 and with no interaction between those variables, ***F***(1, 999) = 1.4; ***p*** > .219. Thus far, the pattern of simulations run parallel to Rescorla's attenuation of sensory preconditioning seen with over-trained. However, the five-epoch revaluation ***reversed*** the acquired equivalence effect, albeit with a relatively small discrimination between contexts C and D. The explicitly trained context and outcome main effects were both unreliable, ***F*** < 1 but the interaction was reliable, ***F***(1, 999) = 5358.8; ***p*** < .001. Similarly, both the context and outcome main effects were unreliable, ***F***s, respectively, ***F***(1, 999) = 1.3; ***p*** > .239, ***F***(1, 999) = 1.6; ***p*** > .199, but their interaction was reliable, ***F***(1, 999) = 35.4; ***p*** < .001. Thus, this feature of our simulations generates a new experimental question: ***Would extensive revaluation attenuate, or even reverse, acquired equivalence?***

(*) TABLE 6 ABOUT HERE PLEASE (*)

Like the variations in inter-layer learning-rate parameter, there was a parameter dependency associated with the extent of revaluation training, when appetitive revaluation was given: Acquired equivalence was correctly simulated with minimal revaluation (i.e., two epochs of training). More extensive revaluation either eliminated discrimination (four epochs of training) or reversed discrimination (five epochs of training). This was not a feature of the aversively revalued acquired equivalence simulation. It is important to note that there is no sense in which our labels for the type of reinforcement, aversive or appetitive, appear in the simulations. They are merely labels for a pair of binary outcomes. Rather, the distinct patterns of the first *"aversive revaluations"* and current *"appetitive revaluations"* are more accurately thought of as revaluation with either the ***same*** ("appetitive") or ***different*** ("aversive") reinforcer as Stage-1's acquired equivalence discrimination. This feature of the simulations delivers a second, testable prediction from the Hebbian network: ***Would the use of two different outcomes in the two stages enhance the acquired equivalence effect?*** In some ways, this over-training effect is paradoxical: We might suppose that, at least up to the point of asymptote, extra revaluation training should only enhance the acquired equivalence effect. The explanation lies in the fact that the acquired equivalence effect relies upon the weight matrices established in Stage-1 (in addition to the Stage-2 revaluation). Using the same Food/No-Food outcomes during revaluation that served in the Stage-1 discrimination allows the Stage-1 weight changes to be modified during revaluation. Clearly, with minimal levels of revaluation, there is a sufficient balance of the necessary weight strengths from both stages to produce the acquired equivalence effect.

2. Congruent/incongruent context combinations

As in the previous revaluation simulations (cf. top panel of Table 2), the current simulation began with the eight-trial appetitive discrimination. The simulation, which results from an intermediate amount of training, is summarized in Table 3 and was commented on above. Rather than use context revaluation to detect acquired equivalence, testing involved the measurement of appetitive responding during ***different combinations*** of pairs of contexts, in the absence of the discrete stimuli, w and x (cf., Hodder et al. 2003; Honey & Ward-Robinson 2001; see second row of Table 2). The appetitive discrimination involved giving contexts A and C and contexts B and D equivalent roles in outcome signalling when in combination with stimulus w and x. This created two types of test context pairs: congruent (i.e., A and C, and B and D) and incongruent (i.e., A and D, and B and C). The simulations successfully detected differences in the behavior of the network in the two trial types, which matched the empirically derived data. Simulated data are summarized in Table 7.

(*) TABLE 7 ABOUT HERE PLEASE (*)

Again, three sets of simulation were performed with learning-rate parameters between the input-to-hidden unit, hidden-to-output unit, and output-to-hidden unit of, respectively: εs = .10, .10, .10; εs = .10, .20, .20; and, εs = .05, .25, .25. Irrespective of the learning rate parameters used at each of the networks' layers, the mean weightings between the input and hidden units were similar and they were similar for both ***Food*** and ***No-Food*** outcomes. Honey and Ward-Robinson (2001) similarly found undifferentiated appetitive responding and commented that this is to be expected when both outcomes are equally well announced by the constituent contexts. However, Honey and Ward-Robinson did find evidence of differences in ***variability*** in responding to congruent and incongruent context combinations and this was seen in our simulations also. For the two simulations with εs of .10 between at the input-to-hidden layer, variance in the activation levels between the input and hidden layer was greater for congruent context pairs than for incongruent context pairs. And this was true for the Food and the

No-Food hidden units, which were similar. For the simulation with markedly different learning-rate parameters at the input-to-hidden layer, there was no differentiation in the input-to-hidden layer weightings.

Honey and Ward-Robinson (2001) captured differences in variability in rats' responses using mean absolute differences, rather than variance. By this measure too, the acquired equivalence effect was seen. For the simulation with equivalent learning-rate parameters at all three network layers ($\varepsilon = .10$), the mean absolute difference in activation of the ***Food*** and ***No-Food*** output units was greater, .64, for the congruent context pairs than for the incongruent context pairs, .58. The mean difference was .055, 95% CI [.039, .071], $t(999) = 6.6$; $p < .001$. For the simulations whose learning rate parameters were, εs = .10, .20, and .20 between, respectively, the input-to-hidden layer, the hidden-to-output layer, and the output-to-hidden layer, equivalence was also demonstrated. Here the mean absolute differences for congruent and incongruent context compounds, were respectively .66 and .60, with a mean difference of .059 95% CI [.043, .076]. This difference was reliable, $t(999) = 7.0$; $p < .001$. However, acquired equivalence was not present for simulations with the network whose corresponding learning-rate parameters were: εs = .05, .25, and .25. Here the mean absolute deviation were .64 and .65, respectively for the congruent and incongruent context compounds, $t < 1$.

3. Whole/part reversal

The design of this experiment is summarized in the third row of Table 2. It began with the appetitive discrimination, used in the previous three simulations. The results of the simulations are summarized in Figure 2. Each of the three sets of simulations, differing in their inter-layer learning-rate parameters, includes three lines that indicate the mastery of the Stage-1 discrimination and the two forms of reversal learning: whole and part. With all learning-rate parameters and all parts of the discrimination, error rates tended to decrease, that is, the network learned. The acquisition of the stage-1 discrimination was very similar, a reflection of the large sample of simulations used. Of more importance is the relative rate of acquisition of whole and part reversal learning.

For each set of simulations with different learning-rate parameters, 1000 networks were simulated with different, random starting weights. Each network was trained for 50 epochs on the Stage-1 discrimination. The network was then cloned and the two identical copies were separately trained for another 50 epochs on either a whole- or partial-reversal of the original discrimination. This method provides a within-networks comparison of the rate of acquisition of whole- and partial-reversal learning.

(*) FIGURE 2 ABOUT HERE PLEASE (*)

For the simulations with learning-rate parameters of ε = .10 for all network layers, the whole reversal was learned ***more slowly*** than the part reversal. This is the ***opposite*** finding to that reported by Honey and Ward-Robinson (2001). When averaged over all 50 training epochs, the average error rates for each simulation were: Stage-1: .152 (standard deviation of the mean: .045); whole reversal: .209 (.015); and part reversal: .132 (.054). The average part-whole difference was .077, 95% CI [.076, .078]. A paired-sample ***t*** test confirmed the apparent advantage of the part, over the whole treatment, ***t***(999) = 161.8; ***p*** < .001. 99% of the part-reversal simulations had lower overall error rates than their whole-reversal twin.

Inspection of Figure 2 reveals the same advantage for part- over whole-reversal learning for the simulations with learning rates of εs = .10 at the input layer and .20 at the hidden-output layer. Averaged over all 50 epochs of training, the average error rates for each simulation were: Stage-1: .125 (standard deviation of the mean: .038); whole reversal: .192 (.050); and part reversal: .134 (.050). The average part-whole difference was .067, 95% CI [.064, .070]. A paired-sample ***t*** test confirmed the apparent advantage of the part, over the whole treatment, ***t***(999) = 48.5; ***p*** < .001. 81% of part-reversed networks had error-rates that were higher than their twin network.

However, with the learning-rate parameters of .05 at the input-to-hidden layer and .25 at the hidden-output layer, the simulation matched Honey and Ward-Robinson (2002) empirical result (see also Delamater & Joseph, 2000; Nakagawa, 1986; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991). When averaged over all of the training epochs, the mean error rates for each of the three simulations was: Stage-1: .157 (standard deviation of the mean: .043); whole reversal: .119 (.041); and part reversal: .187 (.035). The average part-whole difference was .068, 95% CI [.065, .071]. This superiority of whole- over part-reversal was reliable, ***t***(999) = 41.8; ***p*** < .001 and 91% of the whole reversal simulations had lower error rates then their part-reversal twin.

Thus, the simulation of the Hebbian network, matches the empirical findings (e.g., Delamater & Joseph, 2000; Honey & Ward-Robinson, 2001; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991), albeit only when specific learning-rate parameters are employed between the three layers of the network. Our selection of these three trios of learning-rate parameters was somewhat arbitrary. They were the first three that we simulated with these four acquired equivalence phenomena and, because they produced a mixture of successful and unsuccessful simulations, we elected not to examine any further learning-rate parameters. It is important to consider the significance of our finding with alternative learning-rate parameters in delivering faster discrimination learning in the part- than the whole-reversal. In particular, should this be taken as a challenge to the Hebbian network in general or our particular simulation of it? One view is that the Hebbian network can allow part reversal to be superior to whole reversal because it depends on the rates at which: 1. hidden-units are re-mapped onto sets of input and output patterns; and 2. hidden-to-output layer weightings are adjusted. The part reversal's hidden-unit re-mapping will be more extensive than the whole reversals but the part reversal's hidden-to-output layer weight change requirements should be less intensive. This is because the whole reversal requires no changes in hidden unit mapping (merely weight changes to the alternative ***Food***/***No-Food*** outcome units) and, because, the part reversal has half the number of hidden-to-output unit weightings to adjust. Thus, no pattern of results disconfirms either the Hebbian network, or its current simulation. But the finding that, under any circumstances, whole-reversal acquisition is superior to part-reversal acquisition challenges many alternative accounts of configural learning (e.g., Brandon et al., 2000; Pearce, 2002).

4. Congruent/incongruent acquisition

The design of this discrimination is summarized in the bottom row of Table 2 and was demonstrated empirically by Honey and Ward-Robinson (2001); (see also, Delamater & Joseph, 2000; Hodder et al., 2003; Nakagawa, 2005; Robinson & Owens, 2013). Unlike the previous four acquired equivalence designs considered here, the current demonstration occurs in a ***single***, sixteen-trial-type, discrimination, having two forms: congruent and incongruent. Despite both forms of the discrimination having the same number of trial types, there was a difference in their rates of learning. In each of the three sets of learning-rate parameters that we examined, the congruent variant of the discrimination was mastered more quickly than its incongruent variant.

For each set of simulations with different learning-rate parameters, 1000 networks were simulated with different, random starting weights. After initialization, each network was cloned and two identical copies were separately trained for 100 epochs on either the congruent or incongruent version of the discrimination. This method provides a within-networks comparison of the rate of acquisition of the two discrimination tasks.

The simulations with learning-rate parameters of $\varepsilon = .10$ at each layer of their networks, are summarized in the leftmost panel of Figure 3. The error-rate for both types of discrimination declined with training, but the improvement was more marked for the congruent form of the discrimination than its incongruent form. Over all 100 epochs of training the average root-mean-square error rates were .143 (standard deviation of the mean: .072) and .191 (.073) for congruent and incongruent discriminations. These values had a mean difference of .049; 95% CI [.043, .055]. A paired ***t*** test confirmed the reliability of this difference, ***t***(999) = 17.1; ***p*** < .001. 73%.

(*) FIGURE 3 ABOUT HERE PLEASE (*)

The simulations with εs = .10 (input-to-hidden layers) and .20 (hidden-output layers), are summarized in the center panel of Figure 3 and are similar to both other simulations. Averaged over the 100 epochs of training, the congruent discrimination's error rate was .121 (standard deviation of the mean: .060) and the distinct discrimination's error rate was .157 (.066). The mean difference was .036; 95% CI [.031, .041], ***t***(999) = 13.5; ***p*** .001. 69% of the congruent networks solved their discrimination faster than their distinct twin. The rightmost panel of Figure 3 summarizes the simulations that used εs of .05 at the input-to-hidden layer and .25 between the hidden and output layers. The error rates over all 50 epochs of training were .131 (standard deviation of the mean .051) and .167 (.167), which had a mean difference of .036; 95% CI [.032, .040]. This difference was reliable, ***t***(999) = 17.6; .001 and 73% of the congruent networks had lower error rates than their twin.

Thus, like the shock-revalued discrimination described above, the congruent/incongruent form of acquired equivalence produced results matching the empirical findings (Delamater & Joseph, 2000; Hodder et al., 2003; Honey & Ward-Robinson, 2001; Nakagawa, 2005; Robinson & Owens, 2013) at all of the learning-rate parameters that we examined.

General Discussion

Our current work provides a successful, formal implementation of a Hebbian network model (e.g., Honey, Close, & Lin, 2010; Honey & Ward-Robinson, 2001): We found it to appropriately accommodate four findings from configural, acquired equivalence experiments (e.g., Coutureau et al., 2002; Honey & Ward-Robinson, 2001; Honey & Ward-Robinson, 2002; Honey & Watt, 1998; Honey & Watt, 1999; Iordanova et al., 2007; Ward-Robinson & Honey, 2000). Although originally aimed at explaining acquired equivalence and distinctiveness (e.g., Honey & Hall, 1989) it is more accurately regarded as a ***general*** model of discrimination learning, which has been uniquely informed by the analysis of acquired equivalence.

We found some circumstances where the model’s success was dependent on the particular parameters used. For example, simulation of advantage of whole- over part-reversal learning (Delamater & Joseph, 2000; Honey & Ward-Robinson, 2001; Nakagawa, 1986; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991) was unsuccessful unless the learning rate, ε, was relatively large. This parameter-dependency does not serve to challenge the model's interpretation of extant empirical findings – because we cannot know the organism's learning-rate parameters However, we noted two, new testable predictions that could serve to support or refute our implementation of the Hebbian network. One prediction is that acquired equivalence will be weaker when it is assessed with the ***same outcome*** in the both initial bi-conditional discrimination and the revaluation stage (Coutureau et al., 2002; Honey & Watt, 1999; Iordanova et al., 2007) than when ***two different outcomes*** are used (Honey & Watt, 1998; Ward-Robinson & Honey, 2000) (cf., Table 2, first row). This extant experimentation cannot address this prediction because effect-size statistics from the two classes of experiment confound the use of two outcomes with their other properties. For example, if effect size statistics were larger for two- than for one-outcome acquired equivalence demonstrations would this be because the model is correct in that regard, or because the foot-shock outcome used in the two-outcome experiments is a more potent reinforcer? There can, currently be no answer to that ambiguity. However, this question could be answered by systematically varying the role of one- versus two outcomes. For example, rats could receive either food or sucrose outcomes in the bi-conditional discrimination training followed factorially to create four treatment groups, revaluation with either food or sucrose outcomes. The two groups with one outcome only are predicted by our implementation of the Hebbian network to show a weaker acquired equivalence effect than the two groups whose outcomes changes during revaluation. The second prediction derived from this simulation of the Hebbian network was that overtraining the revaluation stage in the design summarized in Table 2, first row, with the same outcome as in the initial discrimination, should attenuate acquired equivalence. This prediction of the network simulation could be evaluated using a modified version of the procedure reported by (Coutureau et al., 2002; Honey & Watt, 1999; Iordanova et al., 2007) with systematically increased sessions of appetitive revaluation in different treatment groups.

We also noted that the whole- versus part-reversal acquired equivalence procedure summarized in the third row of Table 2 (see, e.g. Delamater & Joseph, 2000; Honey & Ward-Robinson, 2001; Nakagawa, 1986; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991) was sensitive to the specific values of learning-rate parameter, ε, at each layer of the network: The commonly reported superiority of whole- over part-reversals was evident only when ε for the input-to-hidden layer was .05 and the remaining layers' εs were .25. In the other two simulations, whose εs were different, the part-reversal was solved more rapidly than the whole reversal. Nakagawa (1986) reported a similar mixture of experimental results from Y-maze experiments, which was the result of variation in training. Rats were first trained on two successive discriminations in which choices between black versus white stimulus cards and between vertically versus horizontally striped stimulus cards were appetitively reinforced. Evidence of acquired equivalence came from Nakagawa's finding that reversal of the vertical/horizontal discrimination, which all rats received, was accomplished more rapidly if their black/white discrimination was also reversed. A control group received no reversal of their black/white discrimination. However, whole-reversal performance was superior to part-reversal performance only when rats' original discrimination was trained for an additional twelve days after reaching criterion. Rats trained only to criterion on the initial forms of the discriminations performed better on the reversed vertical/horizontal discrimination when they did not also receive the reversed black/white discrimination. That is, when the initial discrimination was not overtrained, the part-reversal was superior to the whole-reversal. It is possible that initial discrimination training in the other experiments that show whole-reversal learning to be superior to part-reversal learning (Delamater & Joseph, 2000; Honey & Ward-Robinson, 2001; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991) was overtrained – but there are no means to assess this. It is also possible that the variations in εs at different network layers capture the effects of overtrained discrimination training. On grounds of parsimony, our εs were fixed but several empirically based theories (e.g., Mackintosh, 1975; Pearce & Hall, 1980) assume that εs will be modified during training. It might be fruitful to apply such considerations to, for example, attention-like phenomena (e.g., Duffaud, Killcross, & George, 2007; George & Pearce, 1999).

In addition to the challenge to mediated conditioning accounts of acquired equivalence that we outlined at the beginning of this report, the configural acquired equivalence phenomena that we consider here, challenge the generality of two general classes of account of configural learning. Brandon et al., 2000) proposed that stimulus combinations produced novel patterns of stimulus coding. For example, the presentation of context A and stimulus w in one of the configural acquired equivalence experiments described here may *add* unique stimulus elements (to those that are not present when A and w are presented alone) and to *subtract* other elements (those that are present when A and w are presented alone). These positive and negative changes in stimulus coding allow distinctively different sets of elements to gain and lose associative strength with their trial's outcome. In providing this solution to configural learning problems, Brandon et al.'s model has successfully accommodated data from discrimination learning and compound Pavlovian conditioning experiments (but see, e.g., George 2018; Haselgrove, Robinson, Nelson, & Pearce, 2008). Pearce (e.g., 2002) proposed a different conception of configural learning in which stimulus representations gain and lose associative strength *only* when they are explicitly paired with a trial outcome; the associative strength of the individual elements that comprise those representations do not change (as they do with Brandon et al.'s model). For example, Pearce's model does not allow changes in associative strength to stimulus w, which is never paired with an outcome in the absence of another stimulus. However, representations for Aw and for A, which are paired respectively with food and shock, will undergo changes in associative strength. They are represented by two separate, albeit similar, representational units. The stimulus elements that comprise representations are, however, important for the outcome of discrimination learning because they govern the *generalization* of the associative strength among these representations. Using these assumptions, Pearce's model effectively captures a great deal of discrimination learning data and has made and confirmed novel predictions. However, as Allman, Ward-Robinson, & Honey (2004) remark, neither of these general classes of account is able to accommodate the configural acquired equivalence phenomena that we summarize here. For example, according to Brandon et al., the revaluation form of configural acquired equivalence will result in distinct representational coding for Aw+, Ax-, which will support learning about the two different outcomes during the appetitive discrimination. During context A's pairing with foot shock, the absence of w and x will result in both the removal of some elements and the addition of others. A's conditioning will result in those remaining elements gaining associative strength which will be able to generalize to context C, to the extent that it shares some of the same elements that context A generated. However, there is nothing in Brandon et al's model to predict that the overlap in A and C's elements is any different to the overlap between A and D's. Similarly, though for different reasons, Pearce's model, is unable to explain configural, revaluation acquired equivalence. Appetitive discrimination will be solved when the eight necessary configural units acquire sufficient associative strength to offset the generalization among then based on their similarity. The subsequent pairing of context A with shock will result in a ninth configural representation for A entering into association with the shock's representation. Testing fear responding to contexts C and D is wrongly predicted by Pearce (2002) to be equivalent because both have equivalent similarity to context A's configural unit and the generalization that this supports.

The bi-conditional discriminations described are explicable as forms of "occasion setting" (see, e.g., Bonardi, Robinson, & Jennings, 2017; Bouton & Nelson, 1998; Holland, 1983); see also, (Rescorla, 1990)). For example, stimulus w might be expected to have ***two*** associations, one with each of the two outcomes, food and no-food. Neither association may necessarily be effective without the accompanying presentation of one or more of the contexts. Context A or context C would act as an occasion setter for the w → food association, whereas contexts B and D would act as occasion setters for a w → no-food association. Here, the occasion setter does not operate on the food representation, ***directly***, rather it operates on the entire w → food association, facilitating its operation. We might think of this class of account as a complement for configural accounts of such discriminations whose explanation requires modification to the stimulus representations (e.g., a single, configural representation of the stimulus configuration of context A and stimulus w), but with no unusual assumptions about the associative structures involved. We note that if such an account were true for the discriminations described here and if binary associations could be subject to mediated conditioning (e.g., Holland, 1990; 2008), then the mediated conditioning account of (simple) acquired equivalence demonstrated by Ward-Robinson & Hall (1998) and by Hall et al. (2003) could be applied to the configural acquired equivalence effect (e.g., Coutureau et al., 2002; Honey and Watt, 1998; Honey and Watt, 1999; Iordanova et al., 2007; Ward-Robinson and Honey, 2000). That is, there would be no necessity to invoke the Hebbian system described by Honey, et al. (2010) and Honey & Ward-Robinson (2002) and the current instantiation would be a forlorn enterprise. However, if we were to accept the occasion-setting/mediated conditioning account of the revaluation forms of configural acquired equivalence, it must also be applicable in the other procedures. But it is unclear how the occasion setting account would apply to finding that congruent/incongruent context combinations generated different patterns of variability (Hodder et al., 2003; Honey & Ward-Robinson, 2002) or acquisition (e.g., Delamater and Joseph, 2000; Hodder et al., 2003; Honey and Ward-Robinson, 2001; Nakagawa, 2005; Robinson and Owens, 2013); or why the speed of reacquisition of whole versus part reversal should differ (Delamater & Joseph, 2000; Honey & Ward-Robinson, 2001; Nakagawa, 1986; Robinson & Owens, 2013; Zentall et al., 1992; Zentall et al., 1991). Our conclusion is, therefore, that the occasion setting analysis of the acquired equivalence demonstrations described here is unlikely to be true.

We presented the use of ***configural*** acquired equivalence tasks as a means of demonstrating that mediated learning (e.g., Hall et al., 2003; Ward-Robinson & Hall, 1999) was not the sole mechanism of acquired equivalence. However, some ***non-configural*** forms of acquired equivalence are also inexplicable in terms of mediated learning (e.g., Delamater, 1998; Nakagawa, 1986; Vaughan, 1988) and it will be instructive to examine our simulation of the Hebbian network on, for example, Delamater's experiments to further test our simulation of the Hebbian network.

References

Allman, M. J., Ward-Robinson, J., & Honey, R. C. (2004). Associative change in the representations acquired during conditional discriminations: Further analysis of the nature of conditional learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(2), 118.

Bonardi, C., Robinson, J., & Jennings, D. (2017). Can existing associative principles explain occasion setting? Some old ideas and some new data. *Behavioural processes*, *137*, 5-18.

Bouton, M. E., & Nelson, J. B. (1998). The role of context in classical conditioning: Some implications for cognitive behavior therapy. *Learning and behavior therapy*, 59-84.

Brandon, S. E., Vogel, E. H., & Wagner, A. R. (2000). A componential view of configural cues in generalization and discrimination in Pavlovian conditioning. *Behavioural Brain Research*, *110*, 67-72.

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, *25*, 323-332.

Coutureau, E., Killcross, A. S., Good, M., Marshall, V. J., Ward-Robinson, J., & Honey, R. C. (2002). Acquired equivalence and distinctiveness of cues: II. Neural manipulations and their implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*, 388-396.

Delamater, A. R. (1998). Associative mediational processes in the acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 467-482.

Delamater, A. R., & Joseph, P. (2000). Common coding in symbolic matching tasks with humans: Training with a common consequence or antecedent. *The Quarterly Journal of Experimental Psychology Section B*, *53*(3b), 255-273.

Duffaud, A. M., Killcross, A. S., & George, D. N. (2007). Optional-shift behaviour in rats: A novel procedure for assessing attentional processes in discrimination learning. *Quarterly Journal of Experimental Psychology*, *60*, 534-542.

George, D. N. (2018). Stimulus similarity affects patterning discrimination learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *44*(2), 128.

George, D. N., & Pearce, J. M. (1999). Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*, 363-373.

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*, 79-88.

Hall, G. (1991). *Perceptual and associative learning*. Oxford: Clarendon Press.

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: evidence for associative mediation. *Journal of Experimental Psychology: General*, *132*(2), 266.

Hanson, H. M. (1959). Effects of Discrimination-Training on Stimulus-Generalization. *Journal of Experimental Psychology*, *58*, 321-334.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review*, *113*, 584-605.

Haselgrove, M., Robinson, J., Nelson, A., & Pearce, J. M. (2008). Analysis of an ambiguous-feature discrimination. *The Quarterly Journal of Experimental Psychology*, *61*(11), 1710-1725.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: John Wiley & Sons, Inc.

Hodder, K. I., George, D. N., Killcross, A. S., & Honey, R. C. (2003). Representational blending in human conditional learning: Implications for associative theory. *The Quarterly Journal of Experimental Psychology Section B*, *56*(2b), 223-238.

Holland, P. C. (1983). Occasion setting in Pavlovian feature positive discriminations. In *4* (pp. 183-206). Cambridge, MA: Ballinger.

Holland, P. C. (1990). Event representation in Pavlovian conditioning: Image and action. *Cognition*, *37*, 105-131.

Holland, P. C. (2008). Cognitive versus stimulus-response theories of learning. *Learning & Behavior*, *36*(3), 227-241.

Honey, R., Close, J., & Lin, T. C. E. (2010). Acquired distinctiveness and equivalence: a synthesis. In C. Mitchell & M. Le Pelley. Oxford: Oxford University Press.

Honey, R. C., & Hall, G. (1989). The acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*, 338-346.

Honey, R. C., & Ward-Robinson, J. (2001). Transfer between contextual conditional discriminations: An examination of how stimulus conjunctions are represented. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*, 196-205.

Honey, R. C., & Ward-Robinson, J. (2002). Acquired equivalence and distinctivenes of cues: I. Exploring a neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(4), 378.

Honey, R. C., & Watt, A. (1998). Acquired relational equivalence: Implications for the nature of associative structures. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 325-334.

Honey, R. C., & Watt, A. (1999). Acquired relational equivalence between contexts and features. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*(3), 324.

Iordanova, M. D., Killcross, A. S., & Honey, R. C. (2007). Role of the medial prefrontal cortex in acquired distinctiveness and equivalence of cues. *Behavioral neuroscience*, *121*(6), 1431-1436.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*, 177-200.

Nakagawa, E. (1986). Overtraining, extinction and shift learning in a concurrent discrimination in rats. *The Quarterly Journal of Experimental Psychology Section B*, *38*(3b), 313-326.

Nakagawa, E. (2005). Emergent, Untrained Stimulus Relations In Many-To-One Matching-To-Sample Discriminations In Rats. *Journal Of The Experimental Analysis Of Behavior*, *83*(2), 185-195.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*, 73-95.

Rescorla, R. A. (1976). Stimulus generalization: Some predictions from a model of Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*, 88-96.

Rescorla, R. A. (1983). Effect of Separate Presentation of the Elements on within- Compound Learning in Autoshaping. *Animal Learning & Behavior*, *11*, 439-446.

Rescorla, R. A. (1990). Evidence for an Association between the Discriminative Stimulus and the Response Outcome Association in Instrumental Learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*, 326-334.

Robinson, J., & Owens, E. (2013). Diminished acquired equivalence yet good discrimination performance in older participants. *Frontiers in psychology*, *4*(October), 1-8.

Vaughan, W. (1988). Formation of equivalence sets in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(1), 36.

Ward-Robinson, J., & Hall, G. (1998). Backward sensory preconditioning when reinforcement is delayed. *Quarterly Journal of Experimental Psychology Section B-Comparative and Physiological Psychology*, *51*, 349-362.

Table 1.

Design of an acquired equivalence experiment by Honey and Hall (1989). Two groups of rats first received discrimination training in which three auditory stimuli (A, B and N) signalled either food reinforcement (+) or no outcome (-). Subsequently stimulus N signalled delivery of a foot-shock. Differential generalization of fear responding was assessed to the remaining pair of stimuli, A and B. In both groups, free-operant responding was less during stimulus A than during stimulus B.

Group	Training			Result
	Appetitive Training	Aversive Training	Testing	
Group N+	A+	---	A?	Conditioned suppression
	B-	---	B?	---
	N+	N → shock	---	---
Group N-	A-	---	A?	Conditioned suppression
	B+	---	B?	---
	N-	N → shock	---	---

Table 2.

The designs of four types of configural acquired equivalence experiment, which are not amenable to a mediated-conditioning account. Letters A-D signify context stimuli that were differentiated on visual or thermal features. Letters w-z represent discrete auditory or visual stimuli, used in appetitive discrimination training. "+" and "-" represent the delivery of food reinforcement on termination of stimuli w-z. The Revaluation experiment includes a foot-shock discrimination with shock delivery indicated, respectively by, "→ shock" and "-". *1. Revaluation*: For example, Ward-Robinson and Honey (2000) gave rats initial appetitive training on a pair of biconditional discriminations involving the contexts A-D and the auditory stimuli, w and x. After this, Context A and B were both presented successively, and equally often, but in the absence of w and y. Context A was revalued by its pairing with a foot-shock during its presentations. Generalization of freezing, the conditioned response to the foot-shock, from A to context C was greater than from A to context D. In some variants food outcomes were used during Stage 1 and Stage 2. *2. Congruent/incongruent context combination*. Honey and Ward-Robinson (2002) used a similar Stage-1 procedure to that of Ward-Robinson and Honey (2000). The aversive training was omitted and, instead, the context stimuli were tested as compounds (i.e., one visual, A or B, with one thermal, C or D). Variability in appetitive responding (magazine activity) was greater in the combinations of contexts that had indicated the same w/x-reinforcement contingencies (*congruent*) than in combinations of contexts that had indicated the different w/y-contingencies (*incongruent*). *3. Whole/part reversal*. Honey and Ward-Robinson (2001) used a similar Stage-1 procedure to that of Ward-Robinson and Honey (2000). After sufficient training required to master the pair of biconditional discriminations, the food-reinforcement contingencies were reversed. For some rats (*Whole Reversal Group*), all eight trial types reversed; for other rats (*Part Reversal Group*), only four of the trial types were reversed and the remaining four trial types continued to signal their original contingencies. Despite their having more new information to learn, the Whole Reversal Group mastered their reversed discrimination more quickly that the Part Reversal Group. *4. Congruent/incongruent acquisition*. Honey and Ward-Robinson (2001) used a similar Stage-1 procedure to that of Ward-Robinson and Honey (2000), but additional stimuli, y and z, were included, giving sixteen trial types (four biconditional discriminations). Rats were divided into two treatment groups, whose biconditional discriminations were arranged differently. For the *Congruent Group* two pairs of contexts (A and C, and B and D) were either reinforced or non-reinforced with same discrete stimulus, w-z; but for the *Incongruent Group*, no two contexts had the same reinforcement relationship with w-z. Despite both groups being matched in having four biconditional discriminations to solve, the *Congruent Group's* acquisition was superior to the *Incongruent Group's.*

Experiment	Group	Stage 1 (appetitive)	Stage 2	Result
1. Revaluation	Within Subject	Aw+ Ax-	A → shock	---
		Bw- Bx+	B-	---
		Cw+ Cx-	C?	Higher freezing
		Dw- Dx+	D?	Lower freezing
2. Congruent/incongruent context combinations	Within Subject	Aw+ Ax- Bw- Bx+ Cw+ Cx- Dw- Dx+	AC? v AD? BD? v BC?	Variability in appetitive responding is greater in congruent context combinations (AC and BD) than in incongruent combinations (AD and BC).
3. Whole/part reversal	Whole Reversal	Aw+ Ax- Bw- Bx+ Cw+ Cx- Dw- Dx+	Aw- Ax+ Bw+ Bx- Cw- Cx+ Dw+ Dx-	Faster reversal learning
	Part Reversal		Aw+ Ax- Bw- Bx+ Cw- Cx+ Dw+ Dx-	Slower reversal learning
4. Congruent/incongruent acquisition	Congruent	Aw+ Ax- Ay+ Az- Bw- Bx+ By- Bz+ Cw+ Cx- Cy+ Cz- Dw- Dx+ Dy- Dz+		Faster acquisition
	Incongruent	Aw+ Ax- Ay+ Az- Bw- Bx+ By- Bz+ Cw+ Cx- Cy- Cz+ Dw- Dx+ Dy+ Dz-		Slower acquisition

Table 3.

Weight strengths (*W*s) between the six input units and the eight hidden units of the Hebbian network. Each hidden unit's largest pair of context input *W*s and the larger of the two discrete stimulus input *W*s are in bold, indicating the context-stimulus combinations most likely to activate each unit. Values come from a simulation based on Stage-1 of the discrimination used by Honey and Ward-Robinson (2000), which is summarized in Table 2, row 1.

Input Unit	Hidden Unit							
	1	2	3	4	5	6	7	8
	ACw+	BDx-	ACw+	BDw+	ACx-	BDx-	BDw+	ACx-
Context A	**.765**	.052	**.526**	.067	**.758**	**.414**	.052	**.601**
Context B	.069	**.691**	.073	**.682**	.072	**.493**	**.671**	.070
Context C	**.595**	.057	**.826**	.064	**.599**	.225	.053	**.761**
Context D	.007	**.699**	.074	**.687**	.070	.368	**.724**	.067
Stimulus w	**1.473**	.021	**1.469**	**1.476**	.026	.280	**1.478**	.025
Stimulus x	.027	**1.479**	.031	.024	**1.473**	**1.219**	.0218	**1.475**

Table 4.

Mean activation levels of the output units of 1,000 Hebbian network on test trials with each of the four context stimuli (A – D), before and after the aversive revaluation stage of an acquired equivalence experiment (cf. Honey and Ward-Robinson, 2000; see Table 2, row 1). The networks were trained with *Food* and *No-Food* outcomes during Stage 1 training, and *Shock* and *No-Shock* outcomes in Stage 2 training. Three pairs of simulations were run with a trio of different learning-rate parameters (ε) between the Input-to-Hidden, Hidden-to-Output, and Output-to-Hidden layers, which are specified in the leftmost column. The center quartet of columns shows the state of the network's activations before Stage-2 revaluation training occurred; the rightmost quartet of columns shows the effect of Stage-2 revaluation, during the test, on the activations. One of the primary comparisons of each set of simulations is shown in bold.

Input → Hidden	Hidden → Output	Output → Hidden	Test Stimulus	Output Unit							
				Before Revaluation				During Test			
				Food	No-Food	Shock	No-Shock	Food	No-Food	Shock	No-Shock
.10	.10	.10	Context A	.65	.62	.04	.04	.62	.59	.48	.02
			Context B	.64	.63	.04	.05	.61	.60	.02	.47
			Context C	.64	.62	.04	.04	.61	.59	**.41**	.02
			Context D	.63	.64	.04	.05	.60	.61	**.02**	.42
.10	.20	.20	Context A	.63	.63	.04	.04	.36	.36	.93	.00
			Context B	.62	.63	.04	.04	.35	.37	.00	.93
			Context C	.62	.63	.04	.04	.40	.41	**.87**	.00
			Context D	.62	.64	.04	.04	.38	.42	**.00**	.86
.05	.25	.25	Context A	.65	.64	.05	.04	.29	.28	.96	.00
			Context B	.65	.64	.04	.05	.29	.28	.00	.95
			Context C	.65	.64	.04	.04	.35	.37	**.88**	.00
			Context D	.64	.65	.04	.05	.37	.36	**.00**	.87

Table 5.

Mean activation levels of the output units of 1,000 Hebbian network on test trials with each of the four context stimuli (A – D), before and after the appetitive revaluation stage of an acquired equivalence experiment (cf. Coutureau et al., 2002; Honey and Watt, 1999; Iordanova et al.,2007). The networks were trained with *Food* and *No-Food* outcomes during Stage 1 and Stage 2 training. Three pairs of simulations were run with a trio of different learning-rate parameters between the Input-to-Hidden, Hidden-to-Output, and Output-to-Hidden layers, which are specified in the leftmost column. The center pair of columns shows the state of the networks' mean activation levels before Stage-2 revaluation training occurred; the rightmost pair of columns shows the effect of Stage-2 revaluation, during the test. One of the primary comparisons of each set of simulations is shown in bold. The outcome of the acquired equivalence test for the three simulations is indicated by the signs (✔ and ✗) in the rightmost column.

Input → Hidden	Hidden → Output	Output → Hidden	Test Stimulus	Output Unit				Acquired Equivalence?
				Before Revaluation		During Test		
				Food	No-Food	Food	No-Food	
.10	.10	.10	Context A	.68	.65	1.00	.09	✗
			Context B	.68	.64	.11	1.00	
			Context C	.68	.65	**.55**	.74	
			Context D	.68	.64	**.76**	.50	
.10	.20	.20	Context A	.67	.64	1.00	.04	✔
			Context B	.63	.67	.04	1.00	
			Context C	.66	.64	**.69**	.60	
			Context D	.64	.66	**.60**	.69	
.05	.25	.25	Context A	.65	.64	1.00	.08	✔
			Context B	.64	.67	.08	.99	
			Context C	.66	.65	**.84**	.40	
			Context D	.65	.66	**.40**	.85	

Table 6.

Mean activation levels of the output units of 1,000 Hebbian network on test trials with each of the four context stimuli (A – D), before and after the appetitive revaluation stage of an acquired equivalence experiment (cf. Coutureau et al., 2002; Honey and Watt, 1999; Iordanova et al.,2007). The networks were trained with *Food* and *No-Food* outcomes during Stage 1 and Stage 2 training. A trio of simulations was run with two, four or five epochs of Stage-2 revaluation training. In all three simulations, the learning-rate parameter was respectively: input-to-hidden layer, **ε** = .10; hidden-to-output layer, **ε** = .20; and, output-to-hidden layer, **ε** = .20. One of the primary comparisons of each set of simulations is shown in bold. The outcome of the acquired equivalence test for the three simulations is indicated by the signs (✔ ≈ and ✗) in the rightmost column.

Epochs of Appetitive Revaluation	Test Stimulus	Output Unit				Acquired Equivalence?
		Before Revaluation		During Test		
		Food	No-Food	Food	No-Food	
2	Context A	.67	.64	1.00	.04	✔
	Context B	.63	.67	.04	1.00	
	Context C	.66	.64	**.69**	.60	
	Context D	.64	.66	**.60**	.69	
4	Context A	.64	.65	1.00	.00	≈
	Context B	.65	.65	.00	1.00	
	Context C	.65	.65	**.61**	.65	
	Context D	.64	.66	**.64**	.63	
5	Context A	.67	.61	1.00	.00	✗
	Context B	.66	.64	.00	1.00	
	Context C	.66	.64	**.58**	.67	
	Context D	.67	.65	**.69**	.57	

Table 7.

Mean activation levels of the output units of 1,000 Hebbian network on test trials with each of four permutations of pairs of context stimuli (cf., Hodder et al., 2003; Honey & Ward-Robinson, 2002; see Table 2, row 2). Two pairs of contexts were equivalent (i.e., Contexts AC and Context BD) and two were distinct (i.e., Contexts AD and Contexts BC). The networks were trained with *Food* and *No-Food* outcomes during Stage 1 training, and the mean and variance of their activation levels is shown in the center and right pairs of columns, respectively. Three pairs of simulations were run with a trio of different learning-rate parameters between the Input-to-Hidden, Hidden-to-Output, and Output-to-Hidden layers, which are specified in the leftmost column. One of the primary comparisons of each set of simulations is shown in bold (equivalent) and italic (distinct). The outcome of the acquired equivalence test for the three simulations is indicated by the signs (✔ and ≈) in the rightmost column.

Input → Hidden	Hidden → Output	Output → Hidden	Test Stimulus	Mean		Variance		Equivalent Context Combinations Inflate Variance?
				Food	No-Food	Food	No-Food	
.10	.10	.10	Contexts AC	.67	.70	**.15**	.15	✔
			Contexts BD	.66	.70	**.15**	.15	
			Contexts AD	.69	.73	*.13*	.12	
			Contexts BC	.70	.72	*.12*	.12	
.10	.20	.20	Contexts AC	.69	.66	**.15**	.15	✔
			Contexts BD	.68	.66	**.15**	.15	
			Contexts AD	.72	.68	*.13*	.13	
			Contexts BC	.71	.69	*.13*	.13	
.05	.25	.25	Contexts AC	.69	.67	**.14**	.14	≈
			Contexts BD	.68	.68	**.14**	.14	
			Contexts AD	.68	.67	*.14*	.14	
			Contexts BC	.70	.66	*.14*	.15	

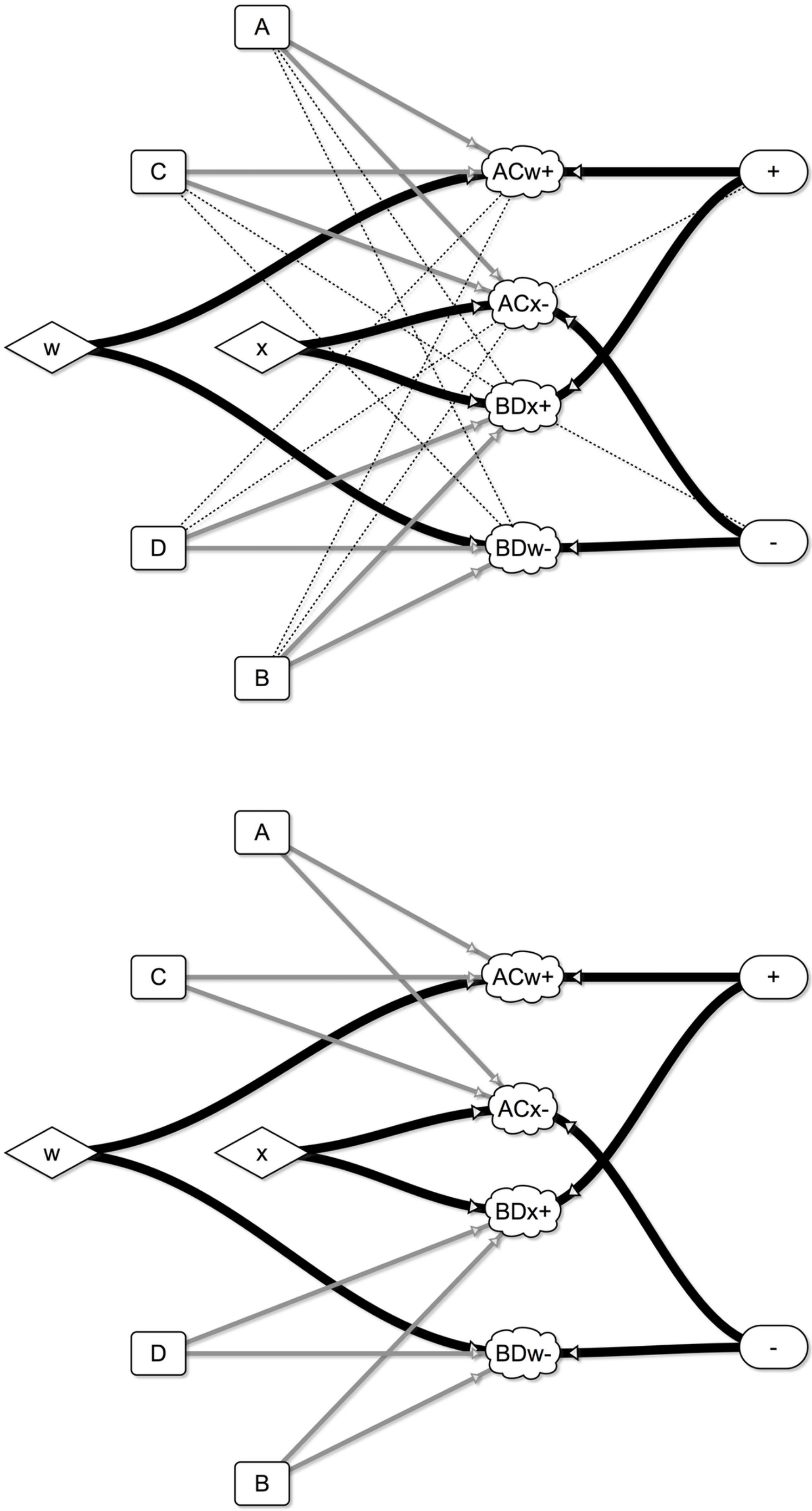

Figure 1. Top panel: Depiction of simulations of the Hebbian network after training on a pair of biconditional discriminations: Aw+ Bw- Cw+ Dw-, Ax- Bx+ Cx- Dx+. Letters A-D represent visually or thermally distinguished context stimuli and letters w and x represent discrete auditory stimuli. "+" and "-", respectively represent the delivery or non-reinforcement of a trial type with food. The cloud-shaped units represent "configural" or "hidden" units, which code for particular combinations of stimuli on particular trial types. Stimuli A-D, w and y, and "+" and "-" act as input units to these hidden units. Strong weightings are: black, broad, curved, unbroken, and arrowed. Moderate weightings are: gray, intermediately wide, straight, unbroken, and arrowed. Weak weightings are: black, fine, straight, broken, and not arrowed. Bottom panel: This depiction is identical to that displayed in the top panel except that weak weighting are removed.

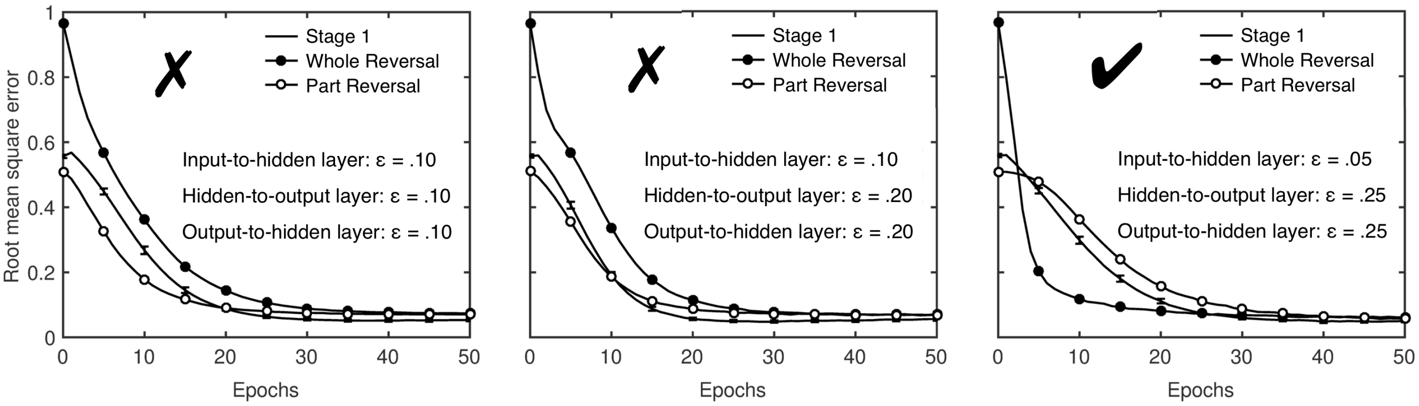

Figure 2. Simulations of Stage-1 acquisition of the appetitive discrimination (see third row of Table 2) and subsequent reversal learning. In the whole reversal all eight trial types signalled the alternative Food or No-Food outcomes; in the partial reversal only four of the eight trial types' outcomes were reversed. The dependent variable is the error rate, which declines as the network learns. Each of the three panels show the results of simulations with different learning-rate parameters (ε) between the three network layers. The figure includes error bars of the variances of each mean, which are partially or entirely occluded by each mean's data point. The success or failure of each simulation is indicated by, respectively, ✔ or ✗.

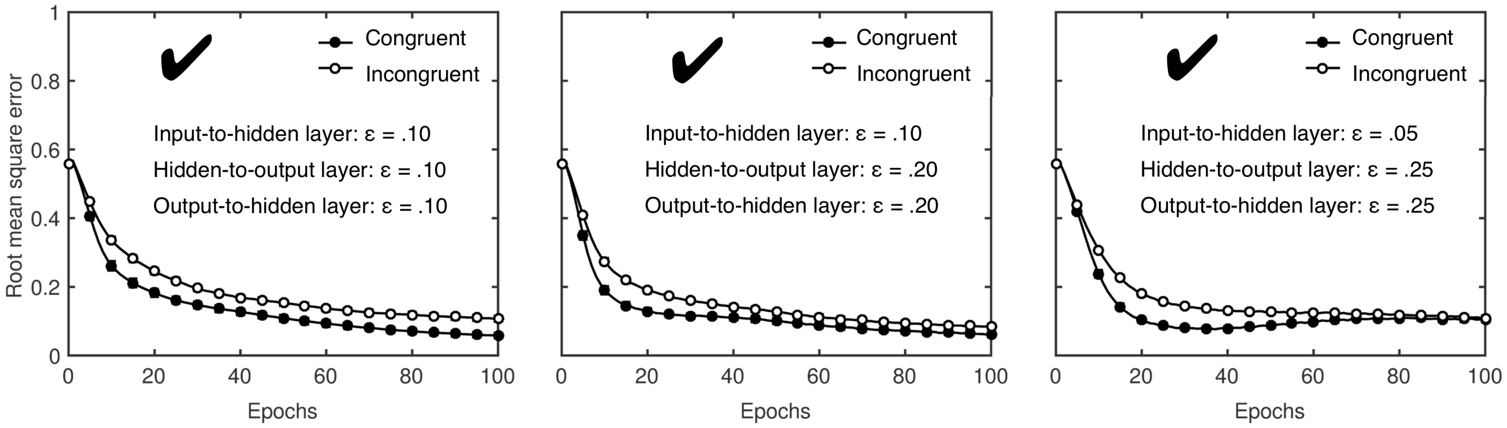

Figure 3. Simulations of acquisition of the congruent versus incongruent acquired equivalence discrimination (see bottom row of Table 2). The three figures show the acquisition by networks on the congruent discrimination (Aw+, Ax,- Ay+, Az-, Bw-, Bx+, By-, Bz+, Cw+, Cx-, Cy+, Cz-, Dw-, Dx+, Dy-, Dz+), in which the Contexts (A-D) and the discrete stimuli (w-z) serve some equivalent roles and the incongruent discrimination (Aw+, Ax-, Ay+, Az-, Bw-, Bx+, By-, Bz+, Cw+, Cx-, Cy-, Cz+, Dw-, Dx+, Dy+, Dz-), in which neither contexts nor discrete stimuli share roles in the discrimination. The dependent variable is the error rate, which declines as the network learns. Each of the three panels shows the results of simulations with different learning-rate parameters (ε) between the three network layers. The figure includes error bars of the variances of each mean, which are partially or entirely occluded by each mean's data point. The success of the three simulations is indicated by ✔.