*LETTER* Communicated by Anne R. Schutte

# A Dynamic Neural Gradient Model of Two-Item and Intermediate Transposition

**Gavin Jenkins**
*gjenkins@sfu.ca*
**Paul Tupper**
*pft@sfu.ca*
*Mathematics Department, Simon Fraser University, Burnaby, BC V5A 1S6, Canada*



**Transposition is a tendency for organisms to generalize relationships between stimuli in situations where training does not objectively reward relationships over absolute, static associations. Transposition has most commonly been explained as either conceptual understanding of relationships (Köhler, 1938) as nonconceptual effects of neural memory gradients (as in Spence's stimulus discrimination theory, 1937). Most behavioral evidence can be explained by the gradient account, but a key finding unexplained by gradients is intermediate transposition, where a central (of three) stimulus, "relationally correct response," is generalized from training to test. Here, we introduce a dynamic neural field (DNF) model that captures intermediate transposition effects while using neural mechanisms closely resembling those of Spence's original proposal. The DNF model operates on dynamic rather than linear neural relationships, but it still functions by way of gradient interactions, and it does not invoke relational conceptual understanding in order to explain transposition behaviors. In addition to intermediate transposition, the DNF model also replicates the predictions of stimulus discrimination theory with respect to basic two-stimulus transposition. Effects of wider test item spacing were additionally captured. Overall, the DNF model captures a wider range of effects in transposition than stimulus discrimination theory, uses more fully specified neural mechanics, and integrates transposition into a wider modeling effort across cognitive tasks and phenomena. At the same time, the model features a similar low-level focus and emphasis on gradient interactions as Spence's, serving as a conceptual continuation and updating of Spence's work in the field of transposition.**

## 1 Introduction

Animals of many species, as well as preverbal humans, can learn and generalize differences between stimuli as if understanding relationships like "bigger than" or "brighter than." This ability is called transposition. The

*Neural Computation* **30, 1961–1982** (2018) 
doi:10.1162/neco_a_01093

most canonical transposition task is to reinforce a subject when presented with only one of a pair of training stimuli. For example, response to the brighter of two training circles may be reinforced but not the darker. After training to criterion, the subject is tested with another pair of stimuli shifted along the differing feature dimension, such as the bright circle reinforced earlier along with an even brighter circle. One circle in this example is the exact match of the reinforced training stimulus (the absolute response, which includes any stimulus at test absolutely closest to the reinforced one), but the other matches the reinforced relationship of being brightest (the relational response). Many species prefer the relational response under certain circumstances in favor of the absolute response, including pigeons (Lazareva, Wasserman, & Young, 2005; Marsh, 1967), rats (Gulliksen, 1932), monkeys (Klüver, 1933), crows (Coburn, 1914), chickens, chimpanzees, and young children (Kohler, 1939).

Relational responding in transposition tasks can be tenuous, however, often only somewhat measuring above-chance performance or being easily disrupted by changing task parameters. Cognitive psychologists have therefore questioned for a century the degree to which this ability represents an abstract conceptual understanding of relationships between stimuli or whether and when transposition can be explained as the product of simpler associative principles and neural circuits, without conceptual understanding.

A seminal nonconceptual explanation of transposition was Spence's (1937, 1945, 1950) stimulus discrimination theory. Spence suggested that if excitatory memory traces formed by reinforced stimuli sum together with inhibitory memory traces formed by nonreinforced stimuli, and if animals preferred test pairs proportionally to this summed activation, then reinforcement-driven learning gradients alone can explain relational responding in the basic transposition task. Spence's model is most easily understood graphically, as in Figure 1. As shown in the figure, the sum of excitatory and inhibitory training gradients creates a maximal level of excitation at a feature value other than the reinforced value, a theoretical phenomenon known as peak shift.

Due to peak shift, stimulus discrimination theory predicts that a test pair of stimuli varying along a single dimension like brightness (such as S7 and S8 in Figure 1) and shifted farther in the reinforced direction from the training stimuli (S6 and S7) will yield the relational answer (S8), because the sum of excitation and inhibition is higher for that response than for the absolute response (S7). Spence's theory predicts that, furthermore, more dramatically shifted pairs tested (S8 and S9 and so on) will still show relational responding until the shifted peak is passed (such as at S10 and S11), at which point subjects will switch to absolute responses as test pairs go down the summed activation slope, then eventually taper toward chance responding as the influence of the gradients wanes entirely (S12 and S13).

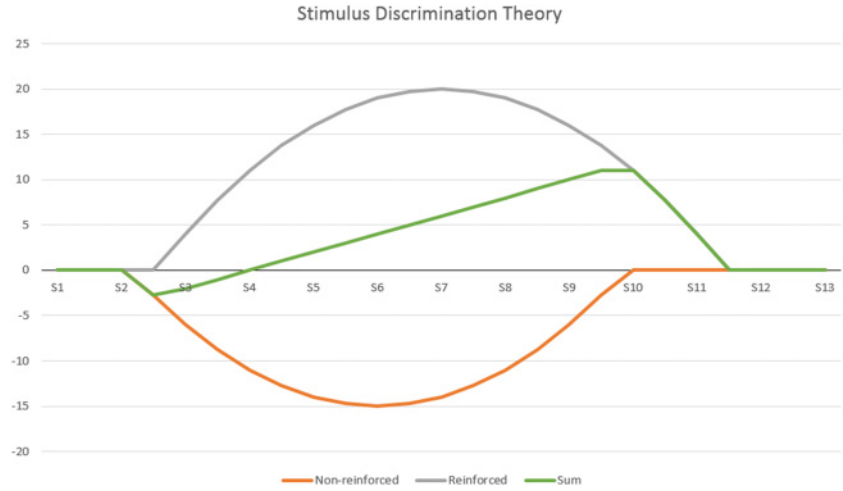


Figure 1: In this model of Spence's stimulus discrimination theory, stimulus S7 was reinforced during training, while stimulus S6 was presented but not reinforced. The proposed resulting excitatory (gray) and inhibitory (orange) learning gradients overlap one another, and when their influence is added together, the sum (green) shows a peak shifted to the right at S10. Thus, when shifted pairs like S7 versus S8 are tested, S8 will be preferred, the relational choice. Very far shifts like S10 versus S11 will revert to an absolute choice, since the sum gradient slopes back downward. Even more extreme shifts can approach chance performance as all gradients cease to have influence (e.g., S12 versus S13).

Spence's work prompted a large amount of experimental study throughout the twentieth century. The core prediction of stimulus discrimination theory—peak shift—has been repeatedly observed behaviorally in line with Spence's model (Spence, 1937, himself; Hanson, 1959; Thomas, 1962; Nicholson & Gray, 1971, in kids with line angles; Baron, 1973, in adults with tones).

Other findings are consistent with stimulus discrimination theory but were not actively predicted by Spence. Training stimuli for longer leads to sharper, stronger memory traces and results in less relational responding (Lazareva, 2012; Gerry, 1971), which is not described in Spence's (1937) theory, although Spence admitted a lack of knowledge of real, neurological gradient shapes and mechanisms. The concept of changing shape, spacing, and sharpness of gradients is included in stimulus discrimination theory, though, and such changes could be used to model the effects of different training times. Also, spacing test stimuli farther apart relative to training stimuli usually leads to less relational responding (Hanson, 1959; Thomas,

1962), which is not directly predicted by Spence's original theory. Spacing's importance is consistent with stimulus discrimination, however, since spacing changes the relative difference across gradients between test items: wider spacing means (usually) steeper gradient differences so should show sharper, more definitive relational or absolute responding.

Other transposition findings contradict stimulus discrimination theory. Subjects are also able to show relational-like responding to the intermediate of three test stimuli after being reinforced on the intermediate of three training stimuli (Brown, Overall, & Gentry, 1959; Gonzalez, Gentry, & Bitterman, 1954, in animals; Reese, 1961, 1962; Rudel, 1957, in children). These findings directly conflict with Spence's (1942) prediction and finding that chimpanzees chose the absolute test item when trained to the intermediate of three stimuli. Stimulus discrimination theory predicts an absolute response because when inhibition is present on both sides of the trained stimulus, there should be no peak shift in the summed activity, and the originally reinforced stimulus should always be chosen if available. The majority of experimental evidence suggests relational intermediate responding, however, and this has remained a significant challenge to stimulus discrimination theory.

In this letter, we introduce a novel computational model of transposition that instantiates many of the same basic neural-level principles as stimulus discrimination theory but in the context of modern, multilayer, dynamic neural networks. Dynamic neural field (DNF) models are a class of recurrent, multilayer neural networks organized by spatial or feature dimensions and performing cognitive processing mainly via dynamics between neighboring neural units over simulated time. Our DNF model of transposition does not include conceptual or symbolic representations of relations, instead relying on the results of neural activation gradients, similar to Spence's theory. With a single parameterization, the DNF model captures all of the original, basic transposition response patterns predicted by Spence (1937) and captures effects not predicted by Spence, including some test item spacing effects and a mechanistic explanation of relational intermediate transposition with three stimuli, contrary to Spence's predictions but still consistent with his primary mechanisms and the bulk of experimental findings.

## 2 A Dynamic Neural Field Model of Transposition

**2.1 Overview.** Dynamic neural field models are a type of recurrent, convolutional neural networks. Layers ("fields") are organized to mimic biological brain functions like attention, memory, or contrastive decision making. Fields are generally organized along perceptual, metric feature dimensions, also following the form of perceptual maps in the brain that are often organized by dimensions like size, color, pitch, or orientation.

Fields communicate with other fields via gaussian kernels across their shared metric dimensions, and fields also connect to themselves recurrently. A field's connections to itself or other fields use both excitatory and inhibitory connections. By varying the relative width and strength of inhibition and excitation, DNF models can establish and maintain precise and often self-stabilizing patterns of activity called peaks. Peaks can maintain themselves only through recurrent connections (good for simulating memory), or they can be tuned to require input (good for a task like attention), depending on the task and situation.

DNF models are also dynamic systems, with simulation unfolding as a function of time, rather than as a function of observations, trials, or discrete pieces of information. DNF models can therefore explain some behaviors or memory representations that traditional gradient descent neural networks cannot explain. When memories are known to drift over time in the absence of new input or when the same stimuli lead to different effects based only on their presentation timing, these are examples of situations where DNF models' dynamic nature gives them an explanatory advantage.

DNF models were designed based on early observations of the mechanics of cortical[1] tissues (Griffith, 1965; Wilson & Cowan, 1972; Amari 1977). In Amari's classic conception of a DNF model, neighboring units in neural populations (ones with similar receptive fields) send strong excitatory activation to each other narrowly, and they send weaker inhibition more broadly, creating stable peaks of activation maintained by the excitation but contained by the inhibition. The exact strength and breadth of these dynamics control whether peaks self-sustain without any input, as might be appropriate to memory fields, or if peaks are fleeting, purely input-driven peaks that are appropriate to attention fields, or something in between. Although early DNF models described fields composed of multiple layers of cooperating excitatory and inhibitory units, our model conceives of fields as single layers connected to themselves in different ways.

DNF models have successfully captured many behavioral effects, from executive control (Buss & Spencer, 2008) to motor planning (Erlhagen & Schöner, 2002) to spatial cognition (Spencer, Simmering, Schutte, & Schöner, 2007) to word learning (Samuelson, Spencer, & Jenkins, 2013) to object recognition (Faubel & Schöner, 2008). DNF models had not yet been applied to transposition tasks, however, despite the fact that DNF peaks and their interactions closely resemble the gradients in Spence's (1937) model and owe a historical debt to the tradition of stimulus discrimination theory. The current model addresses this gap in dynamic modeling and

[1] Our model is more abstract than Amari's (1977) and does not commit to representing a specific cortical brain region, since some species used in transposition tasks are not all traditionally conceived as having "cortical" tissue and/or may feature varying neural architectures. We use generic labels and roles of memory, perceptual, and decision fields instead.

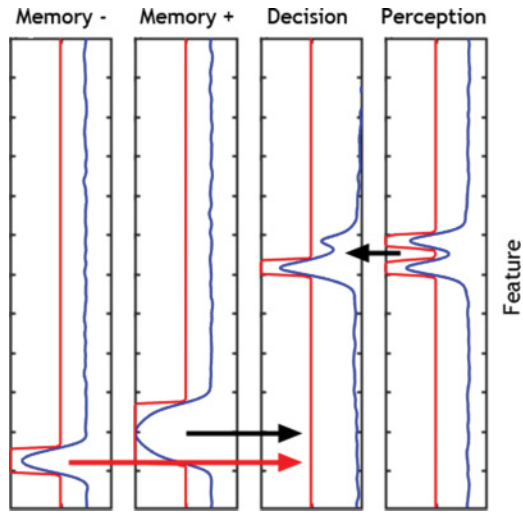


Figure 2: The DNF model of transposition during a test trial. Blue lines represent neural activity as defined in the appendix, function $a(x, t + 1)$. Red lines represent activation after passing through a sigmoid function. Memory from training is represented in the two memory fields to the left of the figure, the $+$field holding the value of any reinforced stimuli and the $-$field holding the value of nonreinforced or punished stimuli. Current test stimuli are represented as input-driven peaks in the perception field to the right of the figure. Perception and reinforced training memories excite the decision field at corresponding positions along the size dimension, while nonreinforced training memories inhibit the decision field (black and red arrows). The decision field has high global inhibition, making it competitive and forcing one peak to usually dominate, indicating an answer. In this example, the model chose an absolute response to a moderately far test pair.

advances a more comprehensive explanation of transposition behavior. Our DNF model is not the only modern model of transposition (see Lazareva, Young, & Wasserman, 2014), but it uniquely captures original and new effects in transposition using a purely neural-level and directly simulated set of mechanisms.

**2.2 Architecture.** Figure 2 depicts the detailed architecture of our DNF model of transposition (for a mathematical description of the model, refer to the appendix). The model consists of four neural fields, each with between 100 and 300 units.[2] The dimension along which the units are organized is

[2] Number of units is not a variable grounded in any empirical data for this task. We varied field size purely for speed of calculation versus maintaining sufficient resolution

left ambiguously labeled as "feature," since transposition tasks use stimuli varying in a variety of features such as size or brightness and since DNF models would treat these identically.[3]

Two "memory" fields represent memory from a prior training phase. The training process itself is only abstractly implied in the model, since nothing of theoretical interest was predicted in this portion of a transposition task. During training, an item pair would have been presented multiple times, with one member of the pair being reinforced and building a positive memory representation, the other not being reinforced and establishing a negative memory representation. Our model begins simulating the task immediately after this training phase, with memory representations already existing as peaks of activation at the appropriate positions in the "memory−" and "memory+" fields, as if receiving input from a (also implied, not explicitly modeled) long-term potentiated memory trace from training. In a more complex model, memory might be divided into these separate potentiated, nondynamic long-term memory fields and dynamic working memory fields, but due to the simplicity of this task, we used only one set of fields that, although dynamic, have fixed inputs driving them, as if working memory was being fed by simple, static memory traces.

A perceptual field receives input from currently observed stimuli. Since the model simulates the test phase specifically, this field always represents the feature values of test items during simulation. In Figure 2, a pair of test items is shown in the perception field moderately shifted from the training stimuli's feature values seen in the training memory fields. Peaks in this field are input driven by currently observed stimuli only.

Finally, the decision field receives input only from the other three fields: excitatory input from the memory+ and perceptual fields and inhibitory input from the memory− field. Thus, it combines current stimuli with prior knowledge to arrive at an educated decision. The decision field features high global inhibition, where any above-threshold activation flatly inhibits all activity in the field. This encourages winner-takes-all activity, aiding in decisive answering. Peaks in this field are input driven by values projected from memory fields and from perceptual fields, so the decision field only makes decisions about currently perceived test items in the context of recent training.

The memory fields project very widely to the decision field—the resulting sum of both training memory inputs results in a wave pattern that can

for the model to record clearly individual peaks of activity. The number is not connected to any theoretical claim.

[3] The one exception would be perceptually circular dimensions (like color hue) that would wrap around and treat units on the other end of the field as neighbors in a DNF model, as if the field were a circle. For this study, we modeled exclusively linear dimensions, which are the most common in transposition tasks, which tend to use features like brightness or size.

be (subtly) seen in Figure 2 in the heights of the many subthreshold peaks, rising higher near the bottom of the figure and falling in strength in the middle and nearer the top of the figure, similar to in Spence's (1937) model. The maximum point of this combined wave pattern is shifted higher than either of the training stimuli, as in Spence's model and as shown in Figure 1. This biases the input from the perceptual layer to favor one test item or the other. For this trial, the test items are shifted far enough that the combined wave is declining in amplitude, so the model has chosen the bottom-most test item, an absolute response. If the items had been shifted less, nearer the rising portion of the combined wave, the model would be more likely to show a relational response.

The decision field features subthreshold bumps of activity. During parameter fitting, these were at every possible response position, driven by the input format of the task given to a subject. During test, bumps were added only at valid answer positions. For example, if a rat is presented with discrete levers or search locations or a pigeon is given only discrete peckable buttons at test, this perceptual information can be represented as the subthreshold preshape in the decision field at those valid response values. Bumps were useful because without fixed anchor points, peaks tend to drift around a feature dimension, as if the response type was a freely sliding scale. This makes it very difficult to quantify when the model has chosen a valid answer, since the drifting occurs within a single trial. It also would make it difficult for the brain to send a motor command to hit a specific lever (rather than a paw reaching uselessly between the levers, for example). Subthreshold bumps anchor or snap peaks into valid locations and hold them there, allowing fewer errors and less ambiguity. This is always important at the actual, valid answer locations. It was also helpful at all of the other possible stimulus locations during parameter fitting, for more easily quantifying the errors the model made (off by one or two stimulus shifts, for example), which was used during model fitting. Having all bumps also made comparing two-item and three-item conditions easier, because the amount of unanchored peak drifting would have been greater among a wider three-peak zone versus a two-peak zone, which would have added confounds when interpreting variance between conditions.

The portion of the neural system that would translate from feature value of the strongest peak to a motor response has been abstracted out of our model, as no theoretically interesting dynamics are predicted during that process.

All fields featured spatially correlated noise (unrelated to field input). We wanted to verify that the model could potentially show relational responding sometimes and also occasionally show absolute responding sometimes for any given training and test stimuli. Both response types are mixed in most behavioral data from transposition tasks, so this is a realistic behavior we sought to emulate. A continuous shift in absolute versus relational

responding also reveals more detailed information about rate of shift in answer type, about which we have made some theoretical predictions discussed in section 3.1. A deterministic model would jump from 100% of one response type to 0% as the amount of test item shift was varied, without providing rate-of-change information.

In transposition tasks, experimental stimuli are usually spaced logarithmically along a feature dimension, but since the DNF model begins with perceptual information included in a single high-level field, we assume that raw logarithmic stimuli have already been transduced and transformed to a linear scale outside the scope of the model. Exact details of how and where this might occur differ by species, but our model is a generalized representation across primates, birds, and rats, so the single perceptual field is conceived as an end result of the perceptual process, a general summary of several possible processing streams cross-species.

All fields in the model form stable but input-reliant peaks of activation. The peaks do not sustain themselves in the absence of input. The memory fields are conceived as receiving input from potentiated long-term memory, the perceptual field is receiving input from sensory organs, and the decision field is receiving input from the other fields, so no independently self-sustaining peak dynamics are needed.

## 3 Simulation

**3.1 Conditions.** We simulated three experimental conditions with a single set of free parameter values across conditions using the DNF model. In the standard condition, we used two-item training pairs and equally (to training stimuli) spaced shifts of two-item test pairs. In the wide condition, we used two-item training pairs and more widely spaced two-item test pairs (three times the spread as trained). In the intermediate condition, we used three-item training triplets (center item reinforced) with equally spaced shifts of three-item test triplets.

Our standard condition tested the set of original predictions of Spence's (1937) stimulus discrimination theory: relational responding for near pairs without the need for conceptual understanding, reversal of this effect with far pairs, and approach to chance responding with extreme pairs. Our wide and intermediate conditions tested additional effects: lower relational responding from wider test pairs and intermediate transposition.

**3.2 Simulation Methods.** For every condition, the set of test items exactly matching the training stimuli (or in the wide condition, the test items centered on the training stimuli) was tested along with a large range of differently shifted test sets in the reinforced direction along the feature dimension relative to the reinforced training stimulus. Most transposition research and theories focus on shifts in the direction of the reinforced stimulus,

because a peak shift is unidirectional and thus the opposite direction is not expected to be theoretically interesting (both absolute and relational accounts agree on responses). We also investigated shifts in the unreinforced direction (not shown) but, as expected, found no interesting effects: responding was near 100% in the two-item tests and was a mirror of the reinforced direction in the three-item intermediate condition.

Once a sufficient level of activity was reached in the decision field for a given set of test items in a given trial of a given condition, the response was logged, the model was reset, and we simulated another identical test trial. We simulated 1000 such identical trials per experimental condition per amount of test shift. One thousand trials was a number that empirically allowed a smooth, continuous change in the proportion of absolute versus relational responding for each amount of test item shift (due to the random noise added to the model's fields), allowing test conditions to be more meaningfully compared. With no noise, responding would suddenly change between 0% and 100% relational responding for different test item shifts, and with fewer than 1000 simulation runs, model result curves were jagged and unnecessarily difficult to interpret.

**3.3 Criteria for Success.** We did not fit any one specific data set since our model is designed to generally capture typical transposition behavior across a wide variety of species. Instead of quantitative fit to a data set, then, parameterizations of the model (specific parameters are described in section 3.4) were evaluated according to three criteria. Two criteria were pass/fail tests, and the third was a continuous measure of balanced decisions.

First, the model was required to choose only valid response values for each test trial. For example, if test items were presented at possible stimulus positions 6 and 7, if the model ever chose a position 8, it was considered too imprecise, and that parameterization was rejected. The existence of subthreshold bumps at every possible test item position during parameter fitting made it possible for the model to clearly make and snap into such errors rather than an ambiguous peak that would drift position over the course of a trial with no constraints on discrete answer types. We expected to have to allow an acceptable threshold for errors, but many parameterizations achieved perfect results on this criterion, allowing us to use a threshold of 0 of this type of error.

Second, the model was also required to show the three-part qualitative pattern expected in the standard condition: high relational responding at low test item shifts (i.e., greater than zero and greater than all more extreme shifts), increasingly absolute responding (>50%) at midlevel shifts, and then at least some amount of recovery back in the direction of an even split in response types at extreme shifts. Any parameterizations that did not show these three distinct phases of responding were rejected.

Finally, parameterizations that showed an overall more even mixture of relational and absolute responding in the standard condition (closer to a

50-50 split on average across all test item shifts) were favored over parameterizations that showed more uneven amounts of the two response types. Behavioral tests for standard transposition consistently show both types of responses with different test item shifts, so an even split meant an appropriately calibrated model and guaranteed informative comparisons to the other two, more experimental conditions.

**3.4 Parameters Explored.** We first explored eight free parameters to determine a general region of acceptable (per section 3.3) model responses. The free parameters are summarized in Table 1. Two of the parameters—test stimulus width and decision field local excitation—were repeatedly found during initial exploration to have only very narrow ranges leading to acceptable results. The reasonable response region was therefore defined and fully explored in final fitting without these as free parameters. Results were also repeatedly acceptable only given a single ratio between excitatory and inhibitory memory projection widths, so the two parameters were yoked as one parameter at that fixed ratio for final fitting. Notably, the reinforced memory representation, being wider than the nonreinforced memory representation, showed the best results, which matches Spence's (1937) model. Only the five remaining parameters were fit to behavioral patterns for final model performance evaluation, as described in section 3.3.

**3.5 Adjustment of Final Parameter Fit.** Most behavioral transposition tasks involve training subjects to a certain level of performance accuracy before testing. This is an equalizing and stabilizing factor between different tasks and conditions within tasks, since research subjects dynamically receive more trials of harder tasks in order to reach a given performance criterion. After we found our best-fitting parameters via the criteria in section 3.3, we simulated this training to minimum accuracy in our model, since the model is applied to three different task conditions. Our simulations did not include an actual training phase, but peak width is a proxy for amount of training time. This is because stimuli that are trained more have been found to result in memory representations with narrower peaks (Lazareva, 2012). Thus, we applied condition-specific multipliers to the width of all training peaks, varying this multiplier until performance without shift was at or above two-thirds correct for each condition. We did not consider this to be a free parameter, because only one possible multiplier is valid by this rule per condition, and the rule was determined prior to any simulation results. For our final set of parameters, only the intermediate condition required narrowing (by a factor of 1.6) to meet the criterion; the other conditions were already above the criterion with no widening (a multiplier of 1). In behavioral terms, this would be analogous to both of the two-item conditions having reached sufficient accuracy after a minimum number of training trials, with the intermediate condition requiring extra training trials due to higher difficulty.

Table 1: Eight Initial Free Parameters and Five Fine-Tuning Free Parameters and Reasons for Searching Each.

| Parameter (free only) | Symbol | Final Fit? | Purpose for Varying | Initial Values Explored | Final Fit Values Tested | Best Value |
|---|---|---|---|---|---|---|
| (+) memory strength | $k_{excite,(+)memory}$ | Yes | Determining relative | 6, 7, 8, 9, 10 | 7, 9 | 9 |
| Test item strength | $k_{from\_each\_stim,perception}$ | Yes | relationship between (+) memory, (−) memory, and test item strength (df = 2) | 7, 8, 9, 10, 11, 12, 13 | 9, 11, 13 | 11 |
| (+) memory width | $\sigma_{excite,(+)memory}$ | Yes | Determining relative | 6, 7, 8, 9, 10 | 7, 9 | 9 |
| Test item width | $\sigma_{from\_each\_stim,perception}$ | No | relationship between (+) memory, (−) memory, and test item width (df = 2) | 2, 3, 4, 5, 6 | [3] | [3] |
| Decision self-excitation | $k_{excite,decision}$ | No | Controls importance of decision field local interactions | 7, 8, 9, 10, 11, 12 | [12] | [12] |
| Decision global inhibition | $k_{global,decision}$ | Yes | Guarantees giving one answer despite amount of input | 0.2, 0.4, 0.6 | 0.4, 0.6 | 0.4 |
| (+) memory to decision projection width | $\sigma_{from(+)memory,decision}$ | Yes | Allows for holding items crisply and distinctly in memory (like in the intermediate condition) but having those items heavily blend and overlap for decisions. | 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 | 70, 80, 90, 100 | 70 |
| (−) memory to decision projection width | $\sigma_{from(-)memory,decision}$ | No | Same as above | 20, 30, 40, 50, 60, 70, 80, 90, 100 | [Yoked to 20 lower than above] | [50] |

Note: Brackets indicate the parameters that were no longer free parameters in final fitting due to having only one acceptable value or yoked relationship found in prefitting.

## 4 Results and Mechanisms

We consider the model's success at each test condition below, which are also graphically displayed in Figure 3. The DNF model is emergent, dynamic, and nondeterministic, so exact mechanisms behind these results are not uniformly clear, but we also discuss the apparent mechanisms involved for each.

**4.1 Near-Pair Relational Responding.** All conditions show relational responding for near-shifted pairs. For the standard and wide conditions, this is due to gradient interactions similar to those in stimulus discrimination theory: the excitatory memory representations are somewhat wider than inhibitory gradients, as well as their projections to the decision field, so the peak decision excitation when the two inputs are summed together shifts to the right (in Figure 3) and until that peak is passed, the relational choice is closer to that shifted peak of excitation and is chosen more often.

For the intermediate condition, peak shift instead lowers relational responding, but a separate mechanism overwhelms this, described below. Figure 4 shows a broken-down view of activity in the model in the intermediate condition. Red is the input to the decision field from the perception field (the test items), blue is the combined input to the decision layer from training memory fields, and green is an example of the decision field's decision on one trial.

The trial shown in Figure 4 is a near-shifted triplet of test items. The gradient effect favors the model choosing the right-most test item (an absolute choice). The greater excitation of the middle test item (red line) than its neighbors is a stronger effect, however. This arises due to dynamic within-field interactions present in the DNF model but not in Spence's (1937) model. Units excite their neighbors in a dynamic neural field, so three very similar test items overlap in excitation, and the center stimulus receives more of this overlapping excitation than do the two extreme stimuli. This favors a center item, relational response despite gradient influences. This mechanism was emergent; we did not plann it prior to simulation. The mechanism does not affect the two-stimulus test conditions, because with two stimuli, both share an equal amount of excitation with one another, so neither gains an advantage.

**4.2 Far-Pair Absolute Response.** With further shifts, the ratio of relational to absolute responses changes. In the standard and wide conditions, responses dip into mostly absolute with far enough test stimuli shifts. This is because the test stimuli have passed the combined gradient from memory, and the absolute response is now closer to peak activation, despite the peak activation having shifted. Thus, the model begins making absolute choices.

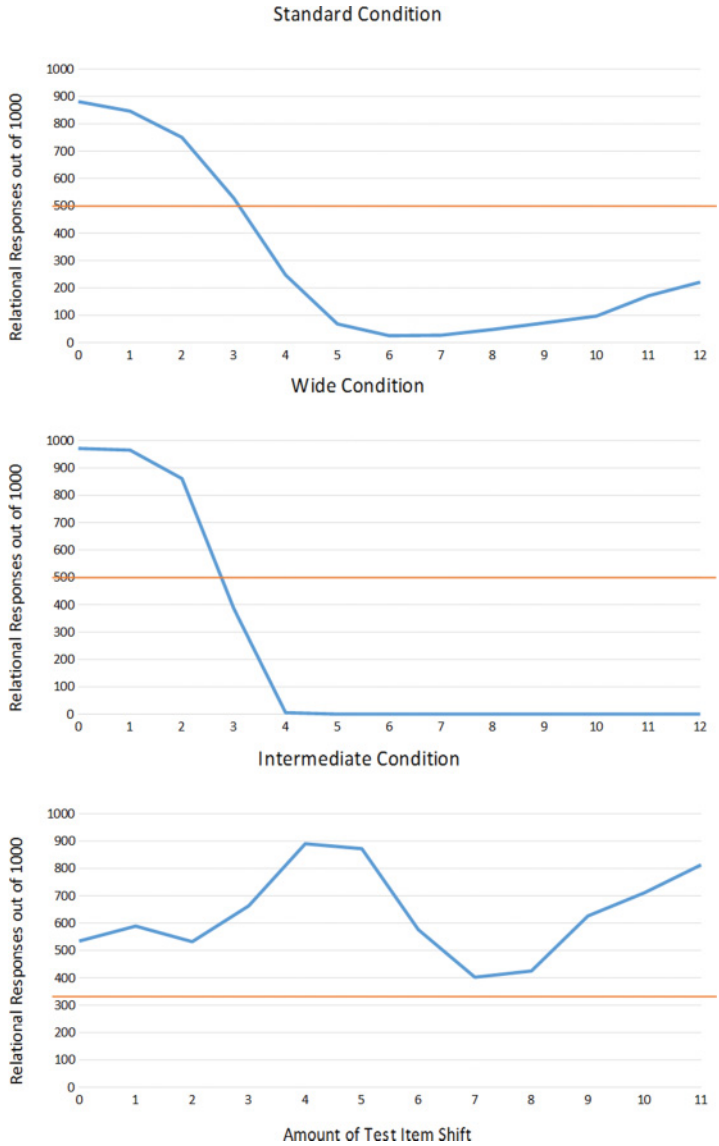


Figure 3: Proportion relational responding with different test item shifts. Zero shift refers to performance on the training stimuli. Each additional step of shift is evenly spaced toward the direction of the reinforced training stimulus (or in the intermediate case, shift direction is arbitrary). Orange lines represent chance performance: 50% for the two stimulus pair conditions, and 33% for the three stimulus condition. Blue lines represent number of relational responses in 1000 simulated trials.

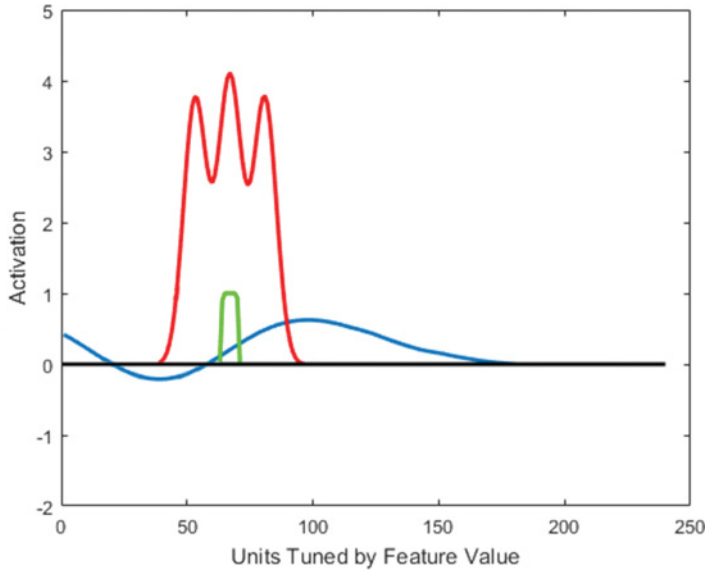


Figure 4: Input and output in the decision field of the DNF model in the intermediate condition. Blue indicates the sum of the input gradients to the decision field from the training memory fields (memory− and memory+ fields in Figure 2), with the center-trained stimulus having been at the minimum point of the blue line, near feature value 30. The three red (activation) peaks are the input from the perceptual field during a near-shifted test trial with test stimuli appearing at the feature positions of the maxima of the three red peaks. Shared dynamic excitation in the perceptual field causes the center item to project a stronger input to the decision field, despite the actual center stimulus having been identical to the flanking stimuli. The decision field's competitive choice for this test trial is shown in green. On this trial, the model made the center relational choice. Note that the characteristic local excitation, lateral inhibition Mexican hat shape of field dynamics is not visible in this figure, since all lines shown here have already been gated by sigmoid functions.

In the intermediate condition, the gradient still plays a role, but unless the central test stimulus lies directly on the peak of the combined inputs from the two different training inputs, gradients will always bias the model toward choosing one of the two flanking stimuli. However, a stronger effect than that of the gradients is the excitation shared among the three peaks in the perceptual field. The central field gets shared excitation from both flanking peaks, while the flanking peaks get only meaningful excitation from the central peak. Thus, the central peak (the relational choice) is quite commonly chosen in this condition, despite the gradients fighting against this choice in most trials. The influence of the gradients still causes all of the

minor perturbations in the results seen in Figure 3, though. At shifts 1 and 7, the gradients are at their steepest and strongest in encouraging flanking stimulus choices (absolute responding). At shift 4, the middle test item happens to line up with the peak of the gradients, so there is a rare boost to relational responding.

**4.3 Extreme-Pair Relational Recovery.** For the standard and wide conditions, responding begins to recover with very far shifts. The wide condition's recovery is difficult to see in Figure 3, but at the farthest tested pair, relational responding had risen from 0 to 5 out of 1000. In the standard condition, recovery is much more pronounced. Recovery is due to the test stimuli progressing beyond the (strongest) influence of memory gradients altogether, causing responding to eventually approach random chance.

In the intermediate condition, relational responding goes gradually up with very far shifts overall. This is similarly due to the gradients becoming shallower and less influential with far shifts (see Figure 4, shallow blue lines on the extreme right side). The shared-excitation mechanisms thus take greater precedence, and relational responding goes slightly higher.

**4.4 Wide Condition versus Standard Condition.** We predicted that the wide condition would lead to more absolute responding sooner due to wide test pairs passing the shifted peak excitation sooner than the standard condition test pairs. We also predicted it would lead to sharper, more distinct relational or absolute responding due to bigger gradient differences between wider test items. We found both effects. The wide condition also shows stronger relational responding with near shifts due to the steeper effective gradient difference between more widely spaced test pairs. It shows slower recovery as well, again because even with shallower sections of the gradient, more widely spaced test pairs effectively have a steeper gradient between them than in the standard condition, so gradients matter with further shifts.

## 5 General Discussion

In 1937, Kenneth Spence did not have access to powerful computational modeling resources, clear empirical knowledge of gaussian neural gradient mechanics, or experience with multilayer neural networks. Yet his stimulus discrimination theory nevertheless proved to be a powerful cautionary reminder of how seemingly complex conceptual understanding may instead be explained by simple cognitive mechanisms. Mere addition of overlapping reinforcement patterns was initially able to account for the same behavior as a sophisticated awareness of analogies.

Our DNF model is a modern technical and spiritual evolution of and upgrade to Spence's (1937) stimulus discrimination theory. The DNF model

replicates the achievements of Spence's model and surpasses them in capturing intermediate transposition, traditionally a thorn in the side of Spence's conclusions. The DNF model captures these new effects with neural mechanisms equally as simple as those featured in stimulus discrimination theory. In doing so, we continue a decades-long history of explaining seemingly complex cognitive effects via local neural dynamics (Spencer, Thomas, & McClelland, 2009).

Future work with the DNF model will focus in part on testing additional transposition effects not covered here. In particular, multiple-pair interactions have been tested and theories developed to explain them (e.g., Lazareva et al., 2014). Contrast manipulations between stimuli and their backgrounds have raised challenges to Spence's theory as well (Lawrence & DeRivera, 1954) that may be addressed with a DNF model.

The DNF transposition model also makes some novel behavioral predictions that can be tested with animals or human research participants. The shared excitation mechanisms should no longer function if four test items are used instead of three in intermediate transposition. The central two items (along a feature dimension) should receive more excitation than the extreme two, but if one were reinforced during training and the other not, the shared excitation should no longer give any beneficial boost or bias to correct relational responding between those two central stimuli.

Similarly, higher spread of test items in the intermediate case should reduce relational responding, since the shared excitation effect should reduce when peaks are farther from one another and less able to share excitation. The same result should obtain from overtraining well beyond a typical criterion, as peaks become overly narrow and stop sharing excitation with one another.

Our DNF model was not developed directly from Spence's stimulus discrimination theory, but the DNF mechanisms and implications establish our model as a close theoretical sibling to Spence's. Both employ neural-level gradients to describe a seemingly high-level conceptual behavior by way of small population mathematical unit interactions. Both subsequently suggest caution in overly quick interpretation of complexity in abstract-looking behaviors. Overall, we consider our DNF transposition model not just as a revalidation of Spence's 1937 specific work but as a modest next step in a broader quest to explain many of the most sophisticated behaviors with simple, neural mechanisms.

## Appendix: Mathematical Description of the Dynamic Neural Field Model

The developing activation (roughly equivalent to neural membrane potential) of the fields as a function of unit position $x$ and time $t$ was simulated in discrete time steps according to

$$a_f(x, t+1) = a_f(x, t) + \frac{-a_f(x, t) + h_f + i_f(x, t)}{\tau_f} + \epsilon_f \tag{A.1}$$

where $f$ is which of four fields in the model, $h_f$ is the resting level of a field, $\tau_f$ modifies rate of relaxation of a field, $\varepsilon_f$ is a random noise factor, and $i_f(x, t)$ is all inputs to a field, expanded as:

$$i_f(x, t) = k_{excite,f} \int g_{excite,f}(x - x')\lambda_{excite,f}(a_f(x', t))dx' \tag{A.2a}$$

$$- k_{global,f} \int \lambda_{global,f}(a_f(x', t))dx' \tag{A.2b}$$

$$- k_{inhib,f} \int g_{inhib,f}(x - x')\lambda_{inhib,f}(a_f(x', t))dx' \tag{A.2c}$$

$$+ \sum_{n=1}^{inputs} k_{n,f} \int g_{n,f}(x - x')\lambda_{n,f}(a_f(x', t))dx' \tag{A.2d}$$

The first two lines represent local, gaussian self-excitation and self-inhibition, the third line a flat global inhibition over the field from its own activation, and the fourth line input from any number of other fields, aligned along a common receptive dimension (such as size) shared with this field. Inputs can be other fields in the model or other inputs such as external visual stimuli (mathematically treated as if a field with a single peak at the stimulus' feature position). $g_{y,z}(x - x')$ and $\lambda_{y,z}(a_z(x', t))$ are generically expanded below:

$$g_{y,z}(x - x') = e^{-\frac{(x-x')^2}{2\sigma_{y,z}{}^2}} \tag{A.3}$$

$$\lambda_{y,z}(a_z(x, t)) = \frac{1}{1 + e^{(-\beta_{y,z}a_z(x,t))}} \tag{A.4}$$

$y$ and $z$ represent the various subscripts used in the previous equations for $g$ and lambda (sigmoid) functions. The sigma (width) term in each gaussian function, the $k$ (strength) coefficients in all input expressions, and the beta (sigmoid steepness) in each lambda function can vary parametrically for each field or projection between fields individually. However, only a very small number of these potential parameters need to be explored in fitting the model to new behaviors or tasks (i.e., are "free" parameters). Table 1 lists all free parameters that were varied and explored to fit transposition behavior separately from dynamic neural field fits to previous cognitive tasks mentioned in the main text. The final values of all parameters, both free and static, are listed in Table 2.

Table 2: All Parameters, Free and Static, in the Model.

| Parameter (both free and constrained) | Best/Final Value | Free Parameter |
|---|---|---|
| $\tau$ (for all fields) | 5 | No |
| Beta (all from perception field) | 2 | No |
| Beta (all from any other field) | 4 | No |
| $k_{excite,(+)memory}$ | 9 | Yes |
| $k_{excite,(-)memory}$ | 10 | No |
| $k_{excite,perception}$ | 5 | No |
| $k_{excite,decision}$ | 12 | Partially |
| $k_{inhib,(+)memory}$ | 3 | No |
| $k_{inhib,(-)memory}$ | 3 | No |
| $k_{inhib,perception}$ | 1 | No |
| $k_{inhib,decision}$ | 1 | No |
| $k_{global,(+)memory}$ | 0 | No |
| $k_{global,(-)memory}$ | 0 | No |
| $k_{global,perception}$ | 0.1 | No |
| $k_{global,decision}$ | 0.4 | Yes |
| $k_{from(+)memory,decision}$ | 20 | No |
| $k_{from(-)memory,decision}$ | 20 | No |
| $k_{from\_perception,decision}$ | 6 | No |
| $k_{from\_decision,perception}$ | 1 | No |
| $k_{from\_each\_stim,perception}$ | 11 | Yes |
| $k_{from\_each\_stim,(+)memory}$ | 9 | No |
| $k_{from\_each\_stim,(-)memory}$ | 9 | No |
| $h_{(+)memory}$ | −5 | No |
| $h_{(-)memory}$ | −5 | No |
| $h_{decision}$ | −7 | No |
| $h_{perception}$ | −5 | No |
| $\varepsilon_{(amplitude,for\ all\ fields)}$ | 2 | No |
| $\sigma_{excite,(+)memory}$ | 9 | Yes |
| $\sigma_{excite,(-)memory}$ | 3 | No |
| $\sigma_{excite,perception}$ | 8 | No |
| $\sigma_{excite,decision}$ | 4 | No |
| $\sigma_{inhib,(+)memory}$ | 6 | No |
| $\sigma_{inhib,(-)memory}$ | 6 | No |
| $\sigma_{inhib,perception}$ | 8 | No |
| $\sigma_{inhib,decision}$ | 10 | No |
| $\sigma_{from(+)memory,decision}$ | 70 | Yes |
| $\sigma_{from(-)memory,decision}$ | 50 | Partially |
| $\sigma_{from\_perception,decision}$ | 4 | No |
| $\sigma_{from\_decision,perception}$ | 4 | No |
| $\sigma_{from\_each\_stim,perception}$ | 3 | Partially |
| $\sigma_{from\_each\_stim,(+)memory}$ | 4 | No |
| $\sigma_{from\_perception,(-)memory}$ | 4 | No |

Table 2: Continued.

Notes: The exact values of the static parameters were all chosen prior to model fitting and are mostly of arbitrary importance. This is because the free parameters control most of the unique degrees of freedom relevant to the task. For example, the width of any gaussian in the model matters only with respect to the widths of other gaussians. Controlling $\sigma_{\text{excite,(+)memory}}$ controls the relative relationship between the memory fields, while the sigmas for each projection between fields control those fields' relationships. In other cases, static variables are easily estimated a priori; for example, the resting level ($h$) of the decision field must be slightly lower than the other fields, since its multiple input channels will tend to add up to larger total activity than elsewhere. Some static variables were also chosen based on typical values across many other past dynamic neural field models.



## Acknowledgments

G. J. was supported by an NSERC Discovery Accelerator Supplement. P. T. was supported by an NSERC Discovery Grant and a Tier 2 Canada Research Chair. We offer special thanks to Ed Wasserman, John Spencer, and Larissa Samuelson for early consultation on this project.

## References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77–87.

Baron, A. (1973). Postdiscrimination gradients on a tone continuum. *Journal of Experimental Psychology*, *101*, 337–342.

Brown, W. L., Overall, J. E., & Gentry, G. V. (1959). "Absolute" vs. "relational" discrimination of intermediate size in the rhesus monkey. *American Journal of Psychology*, *72*, 593–596.

Buss, A. T., & Spencer, J. P. (2008). *The emergence of rule-use: A dynamic neural field model of the DCCS*. Paper presented at the 30th Annual Conference of the Cognitive Science Society, Washington, DC.

Coburn, C. A. (1914). The behavior of the crow, Corvus americanus. *Journal of Animal Behavior*, *4*, 185–201.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*, 545–572.

Faubel, C., & Schöner, G. (2008). Learning to recognize objects on the fly: A neurally based dynamic field approach. *Neural Networks*, *21*, 562–576.

Gerry, J. E. (1971). Peak shift on the tonal-frequency continuum: The effects of extinction and punishment. *Psychonomic Science*, *23*, 33–34.

Gonzales, R. C., Gentry, G. V., & Bitterman, M. E. (1954). Relational discrimination of intermediate size in the chimpanzee. *Journal of Comparative and Physiological Psychology*, *47*, 385–388.

Griffith, J. S. (1965). A field theory of neural nets. II. *Bulletin of Mathematical Biophysics*, *27*, 187–195.

Gulliksen, H. (1932). Studies of transfer of response. *Journal of Genetic Psychology*, *40*, 37–51.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, *58*, 321–334.

Klüver, H. (1933). *Behavior mechanisms in monkeys*. Chicago: University of Chicago Press.

Kohler, W. (1939). Simple structural functions in the chimpanzee and in the chicken. In E. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 217–227). New York: Harcourt. (Original work published 1918)

Lawrence, D. H., & DeRivera, J. (1954). Evidence for relational discrimination. *Journal of Comparative and Physiological Psychology*, *47*, 465–471.

Lazareva, O. F. (2012). Relational learning in a context of transposition. *Journal of the Experimental Analysis of Behavior*, *97*(2), 231–248.

Lazareva, O. F., Wasserman, E. A., & Young, M. E. (2005). Transposition in pigeons: Reassessing Spence (1937) with multiple discrimination learning. *Learning and Behavior*, *33*(1), 22–46.

Lazareva, O. F., Young, M. E., & Wasserman, E. A. (2014). A three-component model of relational responding in the transposition paradigm. *Journal of Experimental Psychology and Animal Behavior Processes*, *40*(1), 63–80.

Marsh, G. (1967). Relational learning in the pigeon. *Journal of Comparative and Physiological Psychology*, *64*, 518–521.

Nicholson, J. N., & Gray, J. A. (1971). Behavioral contrast and peak shift in children. *British Journal of Psychology*, *62*, 367–373.

Reese, H. W. (1961). Transposition in the intermediate-size problem by preschool children. *Child Development*, *32*, 311–314.

Reese, H. W. (1962). The distance effect in transposition in the intermediate size problem. *Journal of Comparative and Physiological Psychology*, *55*, 528–531.

Rudel, R. G. (1957). Transposition of response by children trained in intermediate-size problems. *Journal of Comparative and Physiological Psychology*, *50*(3). 292–295.

Samuelson, L. K., Spencer, J. P., & Jenkins, G. W. (2013). A dynamic neural field model of word learning. In L. Gogate & G. Hollich (Eds.), *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence.* Hershey, PA: Information Science Reference/IGI Global.

Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, *44*, 430–444.

Spence, K. W. (1942). The basis of solution by chimpanzees of the intermediate size problem. *Journal of Experimental Psychology*, *31*, 257–271.

Spence, K. W. (1945). An experimental test of the continuity and non-continuity theories of discrimination learning. *Journal of Experimental Psychology*, *35*(4), 253–266.

Spence, K. W. (1950). Cognitive versus stimulus-response theories of learning. *Psychological Review*, *57*(3), 159–172.

Spencer, J. P., Simmering, V. R., Schutte, A. R., & Schöner, G. (2007). What does theoretical neuroscience have to offer the study of behavioral development? Insights from a dynamic field theory of spatial cognition. In J. Plumert & J. P. Spencer (Eds.), *The emerging spatial mind* (pp. 320–361). Oxford: Oxford University Press.

Spencer, J. P., Thomas, M. S. C., & McClelland, J. L. (2009). *Toward a unified theory of development*. Oxford: Oxford University Press.

Thomas, D. R. (1962). The effects of drive and discrimination training on stimulus generalization. *Journal of Experimental Psychology*, *64*, 24–28.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, *12*, 1–24.

Received June 12, 2017; accepted January 6, 2018.