Received: 11 March 2022 | Accepted: 21 December 2022

DOI: 10.1002/jeab.829

THEORETICAL ARTICLE

Journal of the Experimental Analysis of Behavior

SEAB SOCIETY FOR THE EXPERIMENTAL ANALYSIS OF BEHAVIOR

Computational models of stimulus equivalence: An intersection for the study of symbolic behavior

Ángel Eugenio Tovar[1] | **Álvaro Torres-Chávez**[1] | **Asieh Abolpour Mofrad**[2] | **Erik Arntzen**[3]

[1]Facultad de Psicología, Universidad Nacional Autónoma de México, México

[2]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

[3]Oslo Metropolitan University, Oslo, Norway

Correspondence
Ángel Eugenio Tovar, Office 218, Facultad de Psicología, UNAM, Av. Universidad 3004, Mexico City, CP 04510, Mexico.
Email: aetovar@unam.mx

Funding information
Consejo Nacional de Ciencia y Tecnología, Grant/Award Number: CB-285152

Editor-in-Chief: Mark Galizio
Handling Editor: Liz Kyonka

Abstract

Stimulus equivalence is a central paradigm in the analysis of symbolic behavior, language, and cognition. It describes emergent relations between stimuli that were not explicitly trained and cannot be explained by primary stimulus generalization. In recent years, researchers have developed computational models to simulate the learning of equivalence relations. These models have been used to address primary theoretical and methodological issues in this field, such as exploring the underlying mechanisms that explain emergent equivalence relations and analyzing the effects of training and testing protocols on equivalence outcomes. Nonetheless, although these models build upon general learning principles, their operation is usually obscure for nonmodelers, and in the field of stimulus equivalence computational models have been developed with a variety of approaches, architectures, and algorithms that make it difficult to understand the scope and contributions of these tools. In this paper, we present the state of the art in computational modeling of stimulus equivalence. We seek to provide concise and accessible descriptions of the models' functioning and operation, highlight their main theoretical and methodological contributions, identify the existing software available for researchers to run experiments, and suggest future directions in the emergent field of computational modeling of stimulus equivalence.

KEYWORDS
artificial neural networks, computational models, reinforcement learning, stimulus equivalence, symbolic behavior

Since the seminal and significant works on stimulus equivalence by Murray Sidman (1971, 1992, 1994, 2000; Sidman & Tailby, 1982), this paradigm has become central in the experimental analysis of behavior for the study of language, symbolic behavior, and cognition (Barnes-Holmes et al., 2018; Critchfield et al., 2018; Dickins & Dickins, 2001; Dougher et al., 2014; Green & Saunders, 1998). Stimulus equivalence has been mainly studied using behavioral tasks that assess the ability to derive a full set of stimulus relations (e.g., A = A, B = B, C = C, A = B, B = A, B = C, C = B, A = C, C = A) when only a limited number of these relations have been explicitly trained (e.g., A = B, B = C). This simple yet powerful model is used to account for emergent behavior that cannot be explained by primary stimulus generalization (Dougher et al., 2014). In recent years, innovative methods based on computational modeling have been proposed as a promising technology for research in this field. Notably, several computational models have already been used to explore traditional and new theories on stimulus equivalence, and they have provided research tools for behavioral scientists, who can now run simulations with these models to address their research questions and strengthen links with other scientific disciplines such as linguistics and neuroscience. Nonetheless, the influence of these models within the behavior-analytic tradition has been limited, possibly because researchers in this field are less familiar with this approach. In this

paper, we review and discuss the contributions of computational models of stimulus equivalence. We attempt to present a handy description of the functioning of these models and to highlight the bidirectional contributions in this unique theoretical intersection. On one hand, stimulus equivalence provides a solid framework for testing artificial systems' abilities to show human-like symbolic behavior. On the other hand, computational models provide a means of exploring previous and new theories on the question of how a system, whether human or artificial, can acquire these symbolic relations and show complex behavior. Our goal is to present the state of the art in computational modeling of stimulus equivalence to promote the dissemination of these models and motivate future developments in this field.

This paper is organized as follows. The next section briefly describes the main concepts to understand basic research paradigms and applications of stimulus equivalence. Then, a general review of concepts in computational modeling and artificial neural networks is presented. The subsequent sections review the existing computational models of stimulus equivalence grouped into four main sections with contributions and future directions for each group of models, and then a final section with conclusions is presented.

STIMULUS EQUIVALENCE

Sidman and colleagues worked on different research questions within stimulus control both with humans and nonhumans beginning in the mid-1960s. The results of this research gradually led to findings that indicated that some relations emerged without direct training (Arntzen & Sætherbakken, 2021). Sidman and Tailby (1982) introduced the formal definition of stimulus equivalence and used terms from mathematical set theory (Hrbacek & Jech, 1999, pp. 29–32). Properties such as reflexivity, symmetry, and transitivity have been used as the criteria for the definition of stimulus equivalence. Traditional procedures consider the establishment of baseline conditional discriminations among arbitrary relations, commonly arranged in a matching-to-sample format; a sample stimulus (e.g., A1) is presented, and a response to the sample stimulus initiates the presentation of two or more comparisons. Selection of the correct comparison (e.g., B1) is reinforced, whereas selections of the incorrect comparisons (e.g., B2, B3) are extinguished. The sample and comparisons are either presented simultaneously (simultaneous matching to sample) or with a delay between the offset of the sample and the onset of the comparisons (delayed matching to sample). Following the training of a minimum of three members in two classes as AB and BC relations, the equivalence tests are presented under extinction conditions. Reflexivity means that A is related to A, B is related to B, and C is related to C. Symmetry means that B is related to A, and C is related to B. Transitivity means that A is related to C. In addition, a global test implies that C is related to A (Sidman & Tailby, 1982). The CA trials are often called equivalence trials (e.g., Arntzen & Mensah, 2020).

Research has proven the robustness and flexibility of this paradigm because equivalence relations have been described within stimulus sets composed of words, pictures, sounds, abstract concepts, mathematical concepts, and interoceptive stimuli, to name a few. Moreover, this has resulted in a solid development of applied studies with benefits in clinical and educational contexts. The paradigm has been used for teaching word-object mappings, basic language, writing, and naming skills in children (Stromer et al., 1992); complex concepts to college students including inferential statistics and statistical interactions (Fields, Travis, et al., 2009; Fienup & Critchfield, 2010), classes of logical fallacies (Gallant et al., 2021), and biological concepts such as brain–behavior relations (Fienup et al., 2010); developing clinical interventions in adults (Guinther & Dougher, 2015), with particular relevance in therapies such as acceptance and commitment therapy (Tarbox et al., 2020); and training communication abilities, relational responding, and generalization skills in children with neurodevelopmental disorders (Arnall et al., 2021; Gale & Stewart, 2020; Tovar & Torres-Chávez, 2021).

Moreover, research on equivalence relations also motivated exploring other kinds of derived relational responding. Consider for example that an arbitrary stimulus relation can be established under different kinds of *contextual control* (Dougher et al., 2002); the relation between A and B can be taught as one of equivalence, opposition, or difference, to name a few (Barnes & Hampson, 1993). The relational frame theory (Barnes-Holmes & Harte, 2022) has been raised as one strong approach for the study of different kinds of derived stimulus relations; nonetheless, most of the computational models reviewed here are mainly circumscribed to understanding equivalence relations.

Several variables can influence the formation of equivalence classes, and one such variable is training and testing protocol (Arntzen, 2012). Simple-to-complex, complex-to-simple, and simultaneous protocols are used for the arrangement of training and test trials in experiments studying emergent relations (Adams et al., 1993; Imam, 2006). For example, in an arrangement with three members (A/B/C), the protocols will differ as follows: In the simple-to-complex protocol, one relation (AB) is trained and tested (BA) before the next relation (BC and CB). When all relations are trained and tested in separate blocks, all relations are presented in one test block (BA, CB, AC, and CA). In the complex-to-simple protocol, all relations are trained in one block (AB and BC) followed by a test block including equivalence trials only (CA) and then a test block with all trials (BA, CB, AC, and CA). In the simultaneous protocol, all relations are trained in one block (AB and BC) followed by a test block including

all trials (BA, CB, AC, and CA). The simple-to-complex, complex-to-simple, and simultaneous protocols have shown different outcomes on the tests for emergent relations (Imam, 2006).

Three different training structures, linear series (LS), many-to-one (MTO), and one-to-many (OTM), have been used in the training of baseline conditional discriminations (Green & Saunders, 1998). When training three members (A/B/C) in the classes using the LS structure, AB and BC relations are trained; in the MTO structure AC and BC relations are trained, and in the OTM structure AB and AC relations are trained. Several studies have shown differences in test outcomes for emergent relations depending on the training structure. The general finding is that MTO and OTM produce the same yields (number of participants who form equivalence classes), whereas LS produces substantially lower yields (Arntzen, 2012). Class size, number of nodes (i.e., stimuli related to at least two others during training), the distribution of "singles" among nodes, and directionality of training are essential parameters for understanding the structure of stimulus classes (Fields & Verhave, 1987), and this is important because the class structure and the training protocols affect the learnability of the classes.

A vast number of empirical studies on stimulus equivalence have focused on understanding the influence of the abovementioned variables on the formation of equivalence classes. Remarkably, the effect of all these variables can be modeled in different computational architectures of stimulus equivalence.

COMPUTATIONAL MODELING

Computational models are useful tools in science given their power to simulate natural phenomena and laboratory experiments, predict most-likely outcomes under certain circumstances, discover underlying mechanisms, and propose explanations for complex phenomena. In recent years, these models have been massively expanding, covering the psychological sciences (McClelland, 2009; Wilson & Collins, 2019; Zuidema et al., 2020). Because most of the computational models of stimulus equivalence are based on artificial neural networks (also referred as connectionist models in this context), in this section we briefly describe the basic functioning and structure of these networks to facilitate the review of neural network models of stimulus equivalence.

Architecture and function of artificial neural networks

Artificial neural networks are described in architecture and function, with the two components highly interrelated. The architecture describes the number of artificial neurons composing a network and how these are connected. The functional properties of a network describe how activation (i.e., information) flows through the network and how the network learns from training trials. **Artificial neurons** are processing units with two main functions: (1) an input function that adds up incoming information from external stimuli or from other processing units and (2) an activation/output function that transforms input information into an activation value that propagates to other units or is taken as the response of the network.

The artificial neurons are arranged in layers. Models are composed of single or multiple layers (Figures 1 and 2) where processing is usually feedforward, which means that activation values spread in one direction only, from input/stimuli to output/response layers. The networks may have **weighted connections** within and between layers. As in the brain, connections between neurons allow spreading activation through the network, which exerts either excitatory or inhibitory control on the activation of the connected neurons. The stronger a connection is, the more activation it spreads. Activation values and connection weights are mathematically modeled through different algorithms as we describe later.

Stimuli are usually symbolically presented to the network through input vectors (e.g., binary values 0, 1; see Figure 1). When a stimulus is presented to the network, it triggers the input and activation/output functions of the neurons in a cascade mode until reaching the last units of the network. The final pattern of activation values in the neural network is taken as the response of the model to the input pattern. For clarity of the above description,

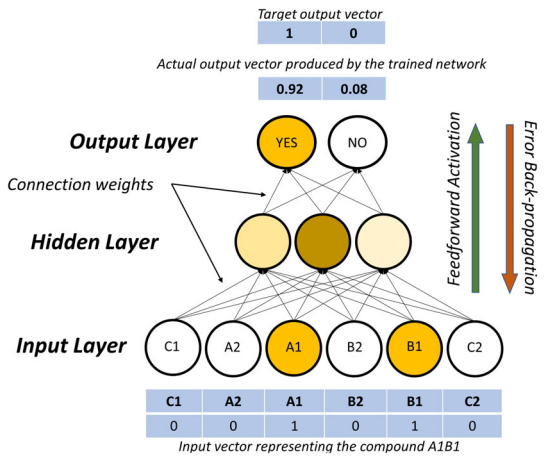

FIGURE 1 A training trial in a three-layer feedforward neural network. A schematic representation of a three-layer feedforward neural network for the learning of the compound stimulus A1B1 with YES/NO responses. Activation flows from the input to the output layer. Error values are computed as the difference between target and actual outputs. Error values are back-propagated to adjust magnitude and direction of connection weights to decrease error for future trials. Colors of neurons represent activation values with yellow units representing highly active units.

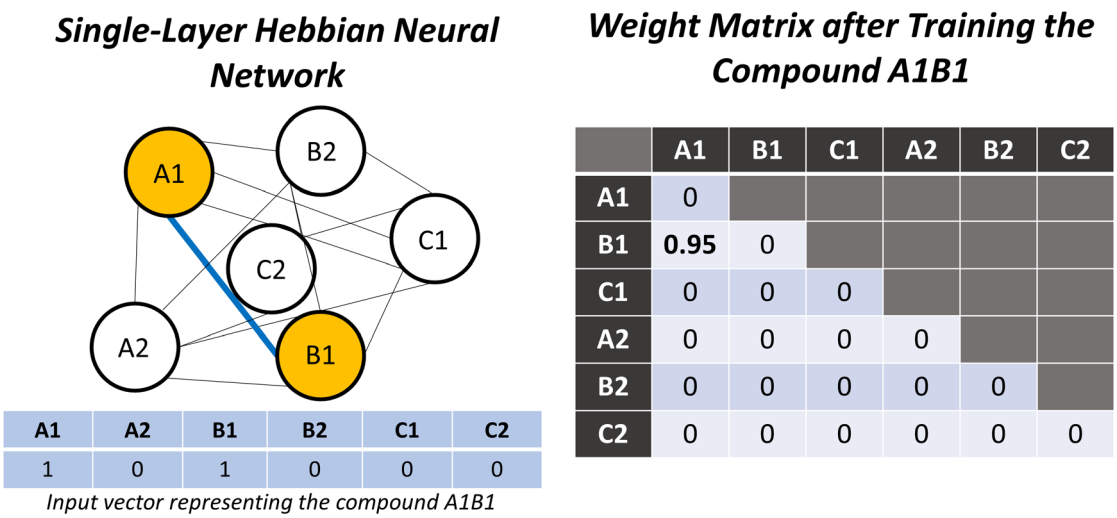

FIGURE 2 A training trial in a single-layer Hebbian neural network. A schematic representation of a single-layer Hebbian neural network for the learning of the compound A1B1. Neurons are fully connected with each other. The coactivation of neurons representing stimuli A1 and B1 leads to strengthening their associative connection, shown as a thicker blue line. The associative strength or relatedness between items is analyzed in the connection matrix after training. The weight matrix on the right shows a strong connection between A1 and B1, the remaining connections are shown in 0 for this example. Note that self-connections (in the matrix diagonal) are possible, which allow for modeling reflexivity in this kind of networks.

Figure 1 shows an example of a multilayer feedforward network for the learning of the stimulus pair A1B1. A binary vector is presented to the input layer that activates neurons A1 and B1. Activation from this layer flows through the weighted connections to the hidden layer and from this one to the output layer where a response pattern is activated. The output activation for this example is 0.92 and 0.08 for the YES and NO response units, respectively, which is close to the target activation of 1 and 0, respectively. In the next section, we describe how these networks produce the expected outcomes from adjusting connection weights.

Learning in artificial neural networks

Artificial neural networks learn. This has been claimed by scientists in the field for many years. A simple but straightforward description of the mechanistic function of connection weights facilitates an understanding of how learning is conceptualized by researchers and accomplished by these networks. The mapping process from stimuli (input vectors) to responses (output vectors) depends on how activation flows through the weighted connections (Figure 1). An *untrained* or *naïve* network has either random or zero connection weight values, which leads to incorrect or inefficient unit activations. However, certain combinations of weight values lead to correct or efficient input–output mappings. The task of the learning algorithms is to find the best combination of connection weights. Remarkably, and highly relevant for stimulus equivalence, many learning algorithms are efficient at finding the best connection weights to solve the input–output mappings of the training phase, and they are also able to show correct responses to new problems (e.g., generalization) during test phases. The ability of artificial neural networks to solve new problems makes them highly suitable for stimulus equivalence research because derived relations can be seen as a case of new problems to solve after training with baseline relations.

Learning algorithms can be classified into many different families. Two relevant families for understanding the neural network models of stimulus equivalence are unsupervised learning and supervised learning.

Supervised learning is provided through labeled examples and is sometimes called "learning with a teacher." Frequently, an external supervisor (i.e., through *teaching* or target vectors as shown in Figure 1) provides the target response for a given input pattern, and the learning process consists of reducing the error (i.e., difference) between the target output and actual output of the network. In neural networks, one of the most popular supervised algorithms is *back propagation* (Rumelhart et al., 1986), which operates by adjusting the weight values across the network layers and gets its name from the fact that error is computed in the output layer and then it is propagated in a direction back to the input layer. This process operates many times (e.g., for many trials, cycles, or iterations; these terms are usually interchangeable) until a small error value is obtained.

On the other hand, **unsupervised learning**, also called "learning without a teacher," refers to those cases where the training examples are presented without labels, expected responses, or target vectors. Instead, learning occurs by detecting stimulus correlations, similarities,

TABLE 1 Computational models of equivalence class formation

Group	Model	Architecture	Learning algorithm	Main experimental procedure simulated	Empirical procedures/data used to evaluate the model	Main contributions
Multilayer feedforward networks	RELNET. Barnes & Hampson (1993); Cullinan et al. (1994); Lyddy et al. (2001); Lyddy & Barnes-Holmes (2007)	Three-layer feedforward neural network	Back propagation	Matching to sample	Arntzen and Holth (1997); Cullinan et al. (1994); Steele and Hayes (1991)	First attempts at simulating derived relational responding and contextual control
	Tovar & Torres-Chávez (2012)	Three-layer feedforward neural network	Back propagation	Compound stimuli with Yes/No responses	Tovar & Torres-Chávez (2012)	First demonstration of responding to derived equivalence relations with back-propagation learning
	Vernucio & Debert (2016)	Three-layer feedforward neural network	Back propagation	Compound stimuli with go/no-go responses	Tovar & Torres-Chávez (2012)	Extends previous models to include Go/No-Go responses in compound stimuli procedures
	EVA. Ninness, et al. (2018); Ninness, Rehfeldt & Ninness (2019); Ninnes & Ninness (2020)	Three-layer, and four-layer (deep) feedforward neural networks	Back propagation	Compound stimuli with yes/no responses, and same/reciprocal responses	Ninnes & Ninness (2020); Ninness, Rehfeldt & Ninness (2019);	First successful simulations of contextual control. Use of deep neural networks for studying derived stimulus relations
Self-organizing maps	Self-Organizing Map Martín et al. (2007); García-Garcia et al. (2010)	Single-layer self-organizing map	Kohonen's learning modulated by reinforcement	Compound stimuli	-	Provides bridges with traditional models of perceptual processing
Biologically inspired neural networks	Lew & Zanutto (2011)	Multi-module neural networks	Based on unsupervised Hebbian learning, supervised learning with reinforcement and Rescorla & Wagner (1972)	Matching to sample, Delayed matching to sample, Visual discrimination	Experimental procedures were based on Devany, Hayes & Nelson (1986), and Sidman & Tailby (1982)	First computational model showing equivalence responding. It bridges behavioral performance with neuronal mechanisms in a biologically realistic architecture
	Tovar & Westermann (2017)	Single-layer Hebbian network	Supervised Hebbian learning with LTP/LTD adjustments	Matching to sample	Devany, Hayes & Nelson (1986); Sidman & Tailby (1982); Spencer & Chase (1996)	Accounts for both equivalence formation and failure through typical and atypical synaptic plasticity. Models relatedness and nodal number effects from training structure

(Continues)

TABLE 1 (Continued)

Group	Model	Architecture	Learning algorithm	Main experimental procedure simulated	Empirical procedures/data used to evaluate the model	Main contributions
Projective simulation	Mofrad et al., 2020	Episodic memory network	Reinforcement learning and max product/ random walk	Matching to sample	Devany, Hayes & Nelson (1986); Sidman & Tailby (1982); Spencer & Chase (1996)	Applies reinforcement learning and projective simulation to equivalence research. Suggests training structures for experimental studies based on simulation results
	Mofrad et al., 2021	Episodic memory network	Reinforcement learning and network enhancement	Matching to sample	Spencer & Chase (1996)	Includes network enhancement in equivalence research. Models nodal number effects in equivalence class formation

and differences in the input patterns. A simple case to understand unsupervised learning is Hebbian learning (Hebb, 1949).

Figure 2 shows an example to describe single-layer networks with Hebbian learning. Note that the network architecture consists of one layer only of fully connected neurons. During a particular trial, two (or more) stimuli activate their corresponding neurons; for example, presentation of the compound A1B1 activates neurons A1 and B1, respectively. The co-occurrent activation of these neurons triggers the strengthening of their connection; *neurons that fire together wire together*. Different from the multilayer network presented in Figure 1, in single-layer Hebbian networks, learning can be analyzed by exploring the connectivity matrix that captures associative strengths (i.e., relatedness) between stimuli processed by each neuron in the network (Figure 2). Hebbian learning can incorporate additional features; for example, weight adjustments for less strong co-activations may actually lead to weakening the connection values (Tovar et al., 2018; Tovar & Westermann, 2017, 2023). And although the Hebbian algorithm is traditionally unsupervised, networks may be sensitive to co-occurrence with programmed reinforcement signals; for example, Hebbian learning can be positive (i.e., strengthening the connection) for reinforced responses to within-class compounds, such as A1B1, and negative (i.e., weakening the connection) for nonreinforced responses to between-class compounds, such as A1B2, combining the influences of regularity-detection and reinforcement on learning.

Finally, the computational simulations are programmed to mimic the general structure of empirical studies. In the field of stimulus equivalence there are two main experimental phases: training and tests. In computational simulations the main distinction between these phases is captured by the fact that learning algorithms are used to adjust connection weights during training trials only, and these algorithms do not operate during tests trials. The network connections are fixed after training baseline relations, and test trials are presented to evaluate performance without further connection weight adjustments, except for the recent reinforcement models reviewed here in the last section, which allow changes in baseline relations during tests.

NEURAL NETWORK MODELS OF STIMULUS EQUIVALENCE

This review is divided into subsections organized chronologically considering the year of the first publication that uses a particular type of model. Each subsection includes follow-up studies that have used the same model or simulation approaches. The primary features and main differences between models are summarized in Table 1.

Feedforward networks

Feedforward networks using matching to sample

Barnes and Hampson (1993) presented the first computational approach to simulating stimulus equivalence. They proposed the use of connectionist models to link interests in both traditional behavior analysis and cognitive science. One important objective of their study was to test whether complex and symbolic emergent behavior could be simulated in artificial systems: a question under vigorous debate during that time.

They presented a three-layer connectionist network called RELNET (Barnes & Hampson, 1993). In this network (Figure 3a), input patterns represent matching-to-sample trials and activations in the output layer indicate the comparison stimulus selected by the network in each trial. Simulations of contextual control were also possible through the mapping of the type of stimulus relation between input and output units: same, different, and opposite (Figure 3a). This model and a group of follow-up models designed with the same principles (Cullinan et al., 1994; Lyddy et al., 2001; Lyddy & Barnes-Holmes, 2007) used supervised learning with the *back-propagation* algorithm.

With this approach, Barnes and Hampson (1993) modeled derived relations under contextual control simulating empirical data from Steele and Hayes (1991). They also compared the effect of linear series versus one-to-many training structures on equivalence class formation (Lyddy & Barnes-Holmes, 2007), simulating empirical procedures and results comparable to those reported by Arntzen and Holth (1997), and they provided mechanistic explanations for why these procedures resulted in different equivalence class formation yields; they described that each training structure provides training on a different sample of statistical regularities and stimulus functions of the class, with the one-to-many structure providing a more readily applicable training structure.

Although this modeling approach seemed very promising as a new tool with which to study equivalence classes, this optimism was later questioned by Tovar and Torres-Chávez (2012), who pointed out a critical network design flaw in RELNET models. To understand this computational flaw, it is important to explain how the input vectors were presented to RELNET during matching-to-sample trials. The network was trained on numerous stimulus sets, and for each trial the network was informed about stimulus functions: particularly, which stimulus was the sample. A training or test trial was presented through different sections of the input vector (Figure 3a). The first section (stimulus identity) indicated what stimuli were used in a particular trial (e.g., A1, B1, B2, and B3). The second section was called the *sample-marking duplicator* and indicated what stimulus served as the sample (e.g., A1). The critical problem entailed by the implementation of this sample-marking duplicator is that it presented the exact same pattern of activations for different trials (i.e., different stimulus sets). For example, the activation of units in the sample-marking duplicator was the same for training either A1B1 or D1E1, with the only difference being that A1B1 was part of the first stimulus set and D1E1 was part of the second stimulus set. Moreover, the duplicator activations during test trials were the same as those during some training trials, which means that RELNET responses during tests of supposedly derived relations were actually directly trained in the sample-marking duplicator. In other words, the sample-marking duplicator can be seen as a *template* for a stimulus relation that dictated which stimulus must be selected as the correct response, whereas the particular stimuli used in training or test trials were interchangeable and presented in an additional section of the input vectors. Consequently, the performance of RELNET networks during tests cannot be considered evidence of emergent behavior in artificial neural networks (see Ninness et al., 2018; and Vernucio & Debert, 2016 for complementary descriptions of problems entailed by the sample-marking duplicator).

Feedforward networks using compound stimuli

Tovar and Torres-Chávez (2012) proposed the next generation of connectionist networks to study stimulus equivalence. Their main interest was in properly documenting whether connectionist networks were useful for simulating responses indicative of derived equivalence relations. To avoid the problems entailed by the sample-marking duplicator needed for matching-to-sample trials, they focused on simulating compound stimuli procedures with YES/NO responses (Debert et al., 2007, 2009; Fields, Doran, et al., 2009; Tovar et al., 2015) because these procedures do not require the specification of stimulus functions.

Tovar and Torres-Chávez (2012) first ran an empirical study with human adults. During training, participants were presented with stimulus pairs (e.g., A1 adjacent to B1), and they responded with either a YES option for within-class stimulus pairs (e.g., A1B1) or NO for between-class pairs (e.g., A1B2). After training of AB and BC relations, participants were exposed to new configurations of compound stimuli representing symmetry, transitive, and equivalence test trials—BA, CB, AC and CA. Four out of six participants formed the stimulus classes according to the criteria. Then, during the second part of their study, the authors presented a three-layer connectionist network using *back-propagation* learning (Figure 3b) that simulated the same training and test structure as in the study with human participants. Their simulation results showed that five out of six runs of the network (i.e., after each "run" all prior learning is deleted in the network and a new simulated participant is

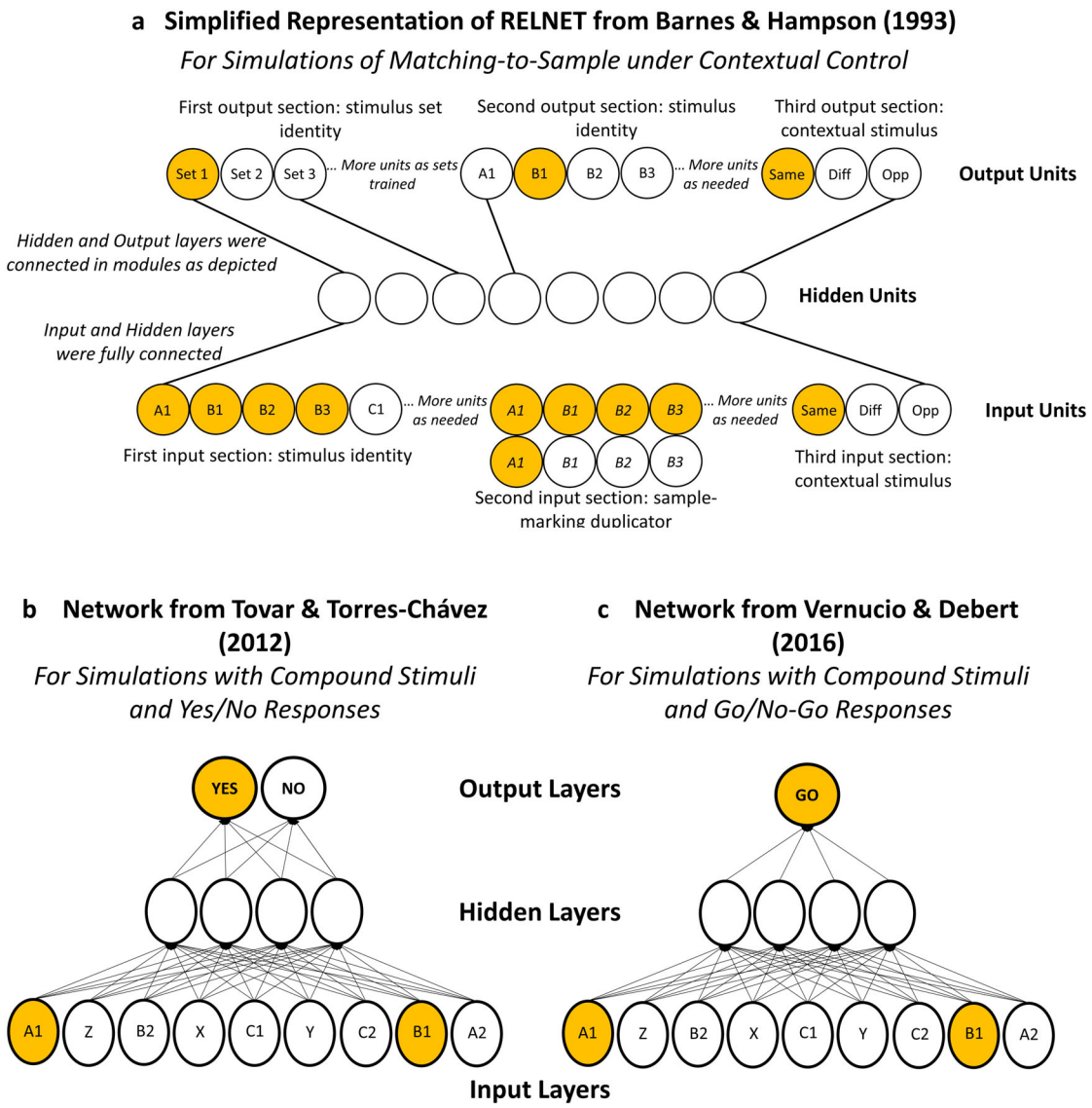

FIGURE 3 Feedforward neural networks for simulations of stimulus equivalence. Panel a: Simplified representation of RELNET as described in Barnes and Hampson (1993), only some units are shown for simplicity. The sample marking duplicator is presented in two rows for space reasons, and the italics indicate that these units change identities depending on the trial, in this case they represent stimuli A and B because these are presented in the first input section. Panel b: Network used in Tovar and Torres-Chávez (2012). Panel c: Network used in Vernucio and Debert (2016). The three architectures show active units in yellow representing a training trial for the relation between A1 and B1. Panels b and c are taken and adapted from the original publications, panel b with permission from the editor, panel c is under License CC BY 4.0.

modeled with new randomized connection weight values) met the equivalence class formation criteria, confirming that connectionist networks were able to simulate trained and derived stimulus relations of the kind documented with human participants.

Vernucio and Debert (2016) later adapted the architecture proposed by Tovar and Torres-Chávez (2012) to simulate go/no-go responses during compound stimuli procedures. They ran simulations using the same training and test protocols as Tovar and Torres-Chávez but with a slight modification to the network architecture; they used only one response unit (Figure 3c), whose activation represented "go" responses, expected for within-class stimulus pairs. Inactivation of the response unit was taken as a proxy for "no/go" expected for between-class trials. This model successfully simulated equivalence class formation.

Building on these models (Tovar & Torres-Chávez, 2012; Vernucio & Debert, 2016), more recently Chris Ninness and his colleagues have been working on developing a computational resource called emergent virtual analytics, or EVA (Ninness et al., 2018). Notably, they have developed a research agenda for extending the application of EVA to analyzing both theoretical and applied aspects of stimulus equivalence and derived stimulus control. For example, they have used EVA to explore the basic training requirements for human participants to derive stimulus relations and generalize to other

training sets in a task directed at establishing stimulus relations between algebraic expressions (Ninness et al., 2019), and more recently they have discussed the implications of simulating more challenging human performances in neural networks (Ninness & Ninness, 2020). Particularly, they have included more hidden layers in the EVA architecture, in a way that the final network includes four layers (one input layer, two hidden layers, and one output layer) instead of the typical three-layer architecture, which provides the network with additional computational power, approaching the methods used in deep neural networks (where "deep" indicates the addition of processing layers).

Ninness and colleagues have modeled the performance of human participants in experiments of contextual control (Ninness & Ninness, 2020), suggesting that responding to symbolic relations under contextual control may require more computational power as provided in their deep neural network model. In this way, they have finally achieved one of the original objectives proposed by Barnes-Holmes and colleagues (Barnes & Hampson, 1993) in the research agenda of connectionist networks and derived responding.

The EVA software is available to researchers interested in running simulation experiments at: http://www.chrisninness.com

Contributions and future directions

Feedforward network models have served as an important bridge between descriptions of stimulus equivalence and studies of symbolic behavior from other areas, including linguistics and cognitive psychology. As a main theoretical contribution, these models have demonstrated that symbolic behavior, as studied in stimulus equivalence paradigms, can be accounted for by the interaction of two components: (a) a learning system with domain-general learning mechanisms, as it is the *back-propagation* algorithm that reduces error during training trials, and (b) the learning history of stimulus regularities. This interaction allows the acquisition of structured representation of stimulus classes.

The notion of stimulus equivalence emerging from domain-general learning mechanisms is theoretically remarkable because it demonstrates that it is unnecessary to use *specific* (i.e., dedicated) functions, instructions, modules, or learning algorithms to learn equivalence relations. This is particularly noticeable in the more recent architectures (Ninness et al., 2018; Ninness & Ninness, 2020; Tovar & Torres-Chávez, 2012; Vernucio & Debert, 2016), where the models did not require explicit instructions or any kind of dedicated computational resources (such as the sample-marking duplicator) to account for derived equivalence relations and contextual control. We highlight this theoretical contribution because it strengthens the empirical view of symbolic behavior as a repertoire that emerges in organisms sensitive to complex stimulus regularities and challenges traditional linguistics, cognitive, and evolutionary theories that propose the need for *specific* (arguably innate) computations to develop symbolic behavior, such as the Chomskyan approach (Berwick et al., 2013; Hauser et al., 2002).

Future work should test the effects of using more realistic representations of stimulus objects and context instead of simple binary representations of them to test how this complexity interacts with the domain-general learning principles included in these architectures. Additionally, slightly more complex architectures can be implemented, such as recurrent networks (Elman, 1990), which are based on the traditional three-layer neural network but include links to feedback activation values of hidden layers to themselves. These recurrent links provide networks with a dynamic memory component that is useful for modeling numerous phenomena and paradigms, such as delayed matching-to-sample and sequence learning.

Self-organizing maps

García and colleagues (García-García et al., 2010; Martín H. et al., 2007) proposed a computational approach to stimulus equivalence using a self-organizing map (SOM). A SOM is a network of artificial neurons arranged in a grid map (Figure 4). In this architecture, each neuron "contains" an internal representation (weight vector). These representations are comparable in size and properties to the stimuli (input vectors) used for training and tests. When one stimulus is presented to the SOM, for example, stimulus A1, the model finds the best matching unit (BMU) as the neuron that has the internal representation with the highest degree of similarity with A1. The BMU and their surrounding units create a neighborhood of active units on the map that respond to the presented stimulus. After finding the BMU, weight adjustments take place; the BMU and its neighboring units adjust their weight vectors to reduce error by means of becoming even closer (i.e., more similar) to the stimulus just presented. Crucially, the BMU's neighboring neurons learn with a reduced amount of error correction. The amount of error correction in each neuron is an inverse function of neighborhood distance to the BMU. Then, the next stimulus is presented to the SOM, and the same process takes place. After several iterations with different stimuli, weight adjustments lead to a topographical organization of the stimuli on the SOM in such a way that similar stimuli will be processed in the same or nearby neurons on the map (Figure 4).

García-García et al. (2010) and Martín H. et al. (2007) proposed a training procedure with single (e.g., A1) and compound stimuli (e.g., A1B1) and a supervised version of the learning algorithm with positive (e.g., for A1B1 trials) or negative (e.g., for A1B2 trials)

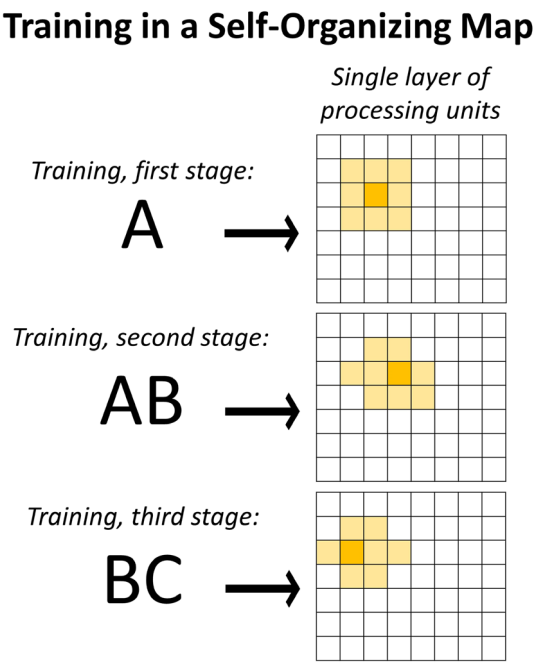

FIGURE 4 Equivalence class formation in a self-organizing map. A Self-organizing map (SOM), as described by García-García et al. (2010) and Martín H. et al. (2007). Colored squares represent active units, with the more saturated square representing the BMU in each map and the light-colored squares representing the BMU's neighboring units. The network is trained sequentially with presentations of A, AB, and BC, which in time results in highly similar activations on the SOM for either A, B, C, AB, BC, BA, CB, AC, and CA, stimuli.

weight adjustments, depending on whether the stimuli in the compound belonged to the same or to different classes. They simulated a traditional arrangement of linear series training by reinforcing AB pairings, followed by reinforcing BC pairings. They argued that the SOM was able to establish equivalence classes because they confirmed that the single and compound stimuli (A, B, C, AB, BA, BC, CB, AC, CA) belonging to the same class were processed within the same cluster in the SOM (Figure 4).

Contributions and future directions

The SOM of García and colleagues (2010; Martín H. et al., 2007) shows that this architecture can be used to model equivalence class formation with stimuli that do not share physical similarities. This is an interesting contribution given that SOMs have been mostly used by psychologists to simulate perceptual processing and categorization where the stimuli in each category maintain physical resemblance (Althaus & Mareschal, 2013; Mayor & Plunkett, 2010; Tovar et al., 2019). Self-organizing maps have shown potential as predictive, descriptive, and theoretical tools in the field of perceptual categorization and language development, but their use in the study of stimulus equivalence is limited to the two studies of García and colleagues reviewed here.

Several limitations of the studies by García and colleagues (2010; Martín H. et al., 2007) should be noted: They have only replicated a generic equivalence experiment, for example, training AB and BC relations before evaluation of symmetry and transitive relations. They did not replicate any empirical studies, did not present detailed data from their simulations, and did not present predictions to be confirmed in future experiments. All these deficits make it difficult to assess the validity and benefits of their modeling approach. Nonetheless, we suggest that one potential field for future research in which these SOMs are useful is in bridging studies of symbolic and perceptual categorization, which is a topic that has received only minimal attention in the field of stimulus equivalence (Fields, 2015). This issue is of interest because SOMs can form clusters of perceptually similar objects, and with the suggestions from the group of García, it will be possible to train SOMs to form classes of objects based on both perceptual and functional similarities to explore in greater detail the integration of perceptual and functional properties during categorization, an area of great interest for cognitive and behavioral scientists. Numerous questions for this field include the following: How are perceptual and functional properties more or less representative of stimulus classes? How are these properties weighted during categorization? Can stimuli have numerous class memberships (e.g., perceptual, symbolic), and how do these properties compete? How do stimulus classes become topographically organized (i.e., relations between classes) as a model of concept development and semantic organization?

Biologically inspired neural networks

The simulations reviewed thus far are built on simple abstract neural networks. The neurons and connections of these models stand for symbolic representations of stimuli and associations between them, respectively. The way the models learn is dictated by the general principles of error reduction. Although this approach offers numerous possibilities to test formal hypothesis of equivalence class formation, the models' functioning hardly relates to specific learning mechanisms implemented by biological neural networks. This lack of correspondence should not be considered a failure of the modeling approach because these models are not intended to provide explanations of brain mechanisms, as has been discussed before for comparable models in other fields of behavior analysis (Burgos, 2007); instead, they provide insight into the main conditions under which equivalence relations are expected to emerge, with the benefit of providing complete control and knowledge of both environmental regularities and learning restrictions.

Nonetheless, one of the most provocative and exciting findings in the literature of stimulus equivalence is that even under the same experimental conditions, only a few organisms can derive equivalence relations, namely, human beings with basic language repertoires. There is no clear evidence of other species forming equivalence classes consistently, despite some controversial results documented in nonhuman animals after very extensive training programs (Kastak et al., 2001; Schusterman & Kastak, 1993), and even humans with limited language repertoires struggle to show derived symmetry and transitivity relations (Devany et al., 1986). One way to account for why human beings are unique in their ability to acquire equivalence classes is by focusing on the role of *processing* and *learning restrictions* beyond studying the effect of training programs and stimulus regularities. It is at this point that stronger links from biological processes to computational implementations became theoretically relevant.

Lew and Zanutto (2011) presented a biologically grounded computational theory for the learning of equivalence relations. Their computational model builds upon a previous model (Lew et al., 2008) that successfully simulated visual discrimination and delayed matching to sample but failed at deriving equivalence relations. Lew and Zautto enriched this previous model with an ambitious set of computational stages and processes to mechanistically explain the emergence of equivalence relations. The resulting model (Lew & Zanutto, 2011) includes response selectivity to stimuli and places, unsupervised associative learning for paired stimuli, reinforcement learning and error reduction for conditional discriminations, and top-down modulation of responses over visual inputs. Notably, each process was included with a biologically rationalized mechanism and the model architecture captures anatomical interactions between structures including the prefrontal cortex, ventro-tegmental area, basal ganglia, and premotor cortex (see original publication for a full schematic representation of the model architecture). The learning algorithms implemented by Lew and Zanutto are both biologically informed and behaviorally relevant, as they are grounded in the Hebbian learning rule (1949) and the Rescorla and Wagner model (1972).

Lew and Zanutto (2011) analyzed the model's performance on three main tasks: visual discriminations, simple conditional relations, and acquisition of equivalence relations. Their focus was on finding the necessary mechanisms for the emergence of equivalence relations beyond learning of visual and conditional discriminations. To do this, they explored the effects of parametric variations and specific lesions to components of their model. Their main findings are as follows: The amount of neural resources affects the equivalence outcomes; they found accuracy in equivalence responding to be a function of the number of neurons in the prefrontal cortex. A minimum number of neurons in this structure was required to show equivalence learning. They programmed three types of lesions to the model: (1) lesions in the dopaminergic system; these mainly disrupted reinforcement learning; (2) lesions in the Hebbian learning process; these mainly disrupted associative learning of paired stimuli; and (3) lesions in the top-down inhibition system, simulating disruptive feedback from frontal functions to processing of visual inputs. To summarize, the three types of lesions impaired the learning of equivalence relations, but they had a lesser or even minimal disruptive effect on the learning of visual and conditional discriminations. These results reveal that a minimal complexity of the prefrontal cortex with an intact dopaminergic system, Hebbian learning, and top-down control are necessary, but none of them is sufficient to explain the emergence of equivalence relations.

To conclude, Lew and Zanutto (2011) asked why, although all the mechanisms proposed as necessary for equivalence class formation exist in other nonhuman primates, only humans learn equivalence relations. In addition to considering the effect of overtraining as a possible explanation, they proposed an additional possibility; in their model, there is a large parametric space of learning mechanisms that results from considering all possible values (e.g., number of neurons, connections), and functions (e.g., variations in learning algorithms), but there might exist a region within this parametric space best suited for the learning of complex stimulus relations such as equivalence. Although learning mechanisms may exist in different organisms, their precise balance and tuning may explain the emergence of complex repertoires in human beings. The idea of this critical region that supports symbolic behavior converges with cognitive approaches to language function; for example, reviewing the relationship between language and its underlying neurobiology, Elizabeth Bates (1999) suggested that the "'language organ' can be viewed as the result of *quantitative adjustments in neural mechanisms that exist in other mammals* [italics added], permitting us to walk into a problem space that other animals cannot perceive much less solve" (p. 10).

Later, Tovar & Westermann (2017) presented a neural network for the simulation of trained and transitive relations. Although simpler and more abstract than the model of Lew and Zanuto (2011), this model (Tovar & Westermann, 2017) uses a biologically inspired learning algorithm as well.

The model was implemented in a single layer of artificial neurons and trained with matching-to-sample trials (Figure 5). This network has localist representations, which means that activation of each artificial neuron in the network stands for one and only one stimulus. Neurons are fully connected through artificial synapses. The connection strength between neurons changes following Hebbian and reinforcement principles. The basic training procedure consists of presenting a sample stimulus (e.g., A1) and two or more comparison stimuli (e.g., B1

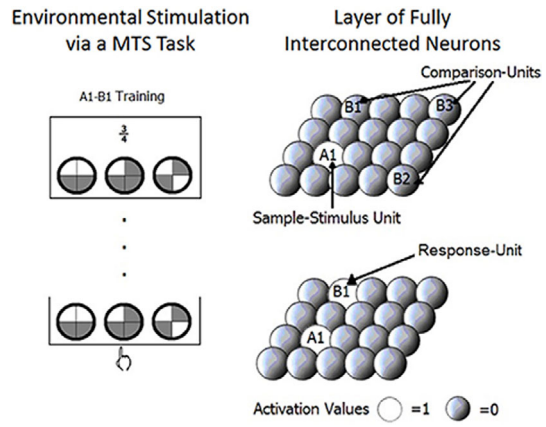

FIGURE 5 Neural network model for equivalence class formation by Tovar and Westermann (2017). The neural network model is presented on the right with active units (in white) representing the processing of the matching-to-sample trial depicted on the left of the figure. All neurons are fully connected with Hebbian connections; however, connections are not visible in the figure. The figure is taken from the original publication under License CC BY 4.0.

and B2). The network selects one comparison (e.g., B1) based on the strongest connection weight from the sample, and if it correctly matches the sample, the connection between these neurons is strengthened; otherwise (i.e., selection of B2), the connection weight is weakened. Strong connections determine the future behavior of the model. Then, the connection weights between all possible stimulus pairs, including trained and derived relations, are analyzed and taken as a proxy for relatedness (i.e., associative strengths) between stimuli. The authors accounted for transitive relations with a simple neuronal process: spreading activation. For example, after several presentations of AB and BC training trials, a functional cell assembly composed of A, B, and C neurons emerged in the network because activation of B during the presentation of BC training also spreads through the trained connection from B to A, recalling the AB relation and allowing activation of all class members in this cell assembly, resulting in the strengthening of the transitive AC relation.

The Hebbian algorithm presented by Tovar & Westermann (2017) includes a continuous function from weakening to strengthening of connections that captures the continuum from long-term depression (LTD) to long-term potentiation (LTP) of biological neural networks (Bienenstock et al., 1982; Bliss et al., 2007; Malenka & Bear, 2004). This Hebbian rule with LTD/LTP provided the possibility of simulating populations with learning disabilities because there is vast neurophysiological evidence of intellectual disability associated with an imbalance favoring LTD at the expense of LTP in synaptic plasticity (Andrade-Talavera et al., 2015; Rueda et al., 2012; Scott-McKean & Costa, 2011). This approach allowed the modeling of biologically relevant variations in the learning mechanisms to test their effect on the acquisition of equivalence classes, and it seeks to explain behavioral differences between populations based on their processing restrictions.

Through parametric variations affecting the LTD/LTP balance, Tovar and Westermann (2017) simulated intellectual disabilities and modeled the classic study of Devany et al. (1986), which was focused on analyzing the acquisition of trained and transitive relations by different groups of children. In the original study, one group of children with learning and language disabilities acquired the trained stimulus relations but failed to show transitive relations, a pattern that was replicated by the model of Tovar and Westermann through the LTD/LTP imbalance. This simulation provided a direct link between realistic neurophysiological variations (i.e., atypical synaptic plasticity) and performance in equivalence class formation. Moreover, the theoretical approach of Tovar and Westermann converges with the hypothesis of the critical region in the space of learning parameters best suited for equivalence learning, which explains important differences between organisms and species in the development of symbolic behavior (Bates, 1999; Lew & Zanutto, 2011).

Tovar and Westermann (2017) also provided a mechanistic explanation for variations in relatedness between members of an equivalence class resulting from class structure and training protocols. This was done through the replication of the training schedules and results of the classic studies by Sidman and Tailby (1982) and Spencer and Chase (1996). The simulations accounted for nodal distance effects and stronger relatedness for trained relations as compared with derived relations. Their results suggest that this model provides a research tool for predicting learning outcomes under different training protocols and structures.

A Matlab implementation of the model by Tovar and Westermann (2017) is available to researchers interested in running simulation experiments at https://osf.io/tx3h4/.

Contributions and future directions

The neural networks reviewed in this section have proposed neurocognitive theories of equivalence class formation that address important questions in stimulus equivalence including the following: What learning mechanisms are required for deriving stimulus equivalence? Are other behavioral repertoires (e.g., basic language skills) necessary for the emergence of equivalence classes? Why do some organisms derive symmetry and transitive relations while others do not? The two models (Lew & Zanutto, 2011; Tovar & Westermann, 2017) converge in demonstrating the emergence of equivalence classes in the absence of language repertoires, and although each of these models emphasizes the relevance of different components and processes for equivalence learning, both suggest the existence of a critical region in the parametric

space of learning processes and restrictions (number of neurons, synaptic thresholds, learning dynamics and modulation) that is optimal for providing the system with the power to derive equivalence relations beyond learning the trained relations.

These models have emphasized the need to better understand the processing restrictions and underlying neurobiology, in addition to the environmental regularities and training history, to account for equivalence class formation. In doing so, they have provided direct links between neuroscience and cognitive behavioral approaches to symbolic behavior.

Several promising directions for future work with biologically inspired models of stimulus equivalence can be considered. One concerns the hypothesis of the critical region in the parametric space of learning processes. What aspects of the best suited learning properties for equivalence class formation may be experience-dependent properties? Answering this question would provide insight into possible interventions to provide such experience and facilitate the acquisition of symbolic repertoires. Additionally, as in the case of the other groups of computational models, using more realistic and informative representations of stimuli (e.g., large vectors that capture the shapes and colors of objects) may be a fruitful area for future research to explore interactions between perceptual processing and stimulus equivalence. Finally, future theoretical work on equivalence class formation should explore the compatibility between descriptions of Hebbian learning models and recent developments in relational density theory (Belisle & Dixon, 2020) because the *density* property of a stimulus relation in the relational density theory may be analogous to the relatedness (i.e., associative strength) captured by the weighted connections in Hebbian networks. This may be indicative of convergent descriptions of stimulus classes arising from complementary perspectives.

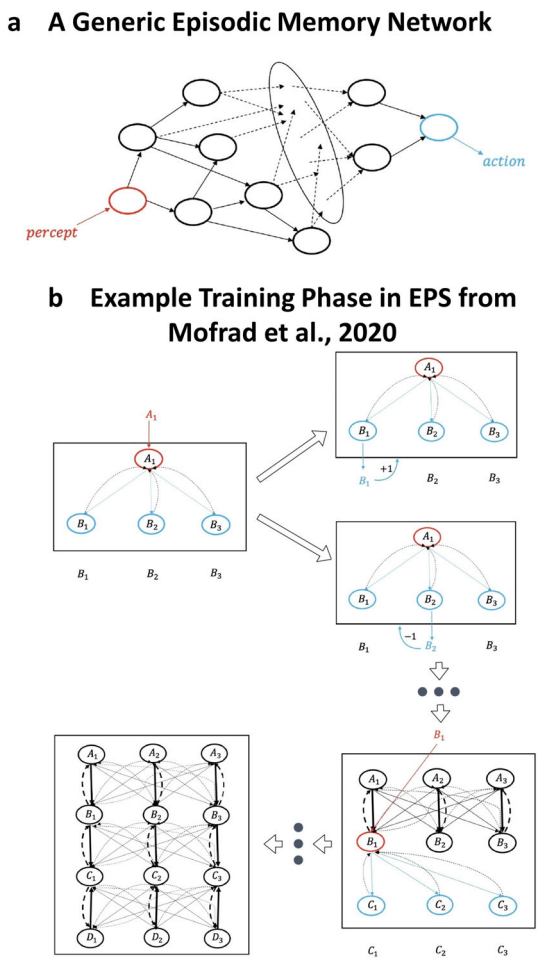

FIGURE 6 Episodic memory network for equivalence class formation by Mofrad et al. (2020). Panel a: Schematic representation of a memory network in a projective simulation model and a random walk on the episodic memory that starts with activation of an episode (in red) and reaches an action episode (in blue). Panel b: Schematic representation of the training phase in EPS that starts by showing the agent a sample stimulus (in red) and three comparison stimuli (in blue). The figure shows that correct selection of B1 is followed by positive feedback and incorrect selection of B2 is followed by negative feedback. Continued lines in each panel show trained relations, and dashed lines show symmetry relations. Thicker lines are used for stronger connections. Panels a and b are taken and modified from (Mofrad et al., 2020).

Reinforcement learning

In this section, we review computational models developed in the framework of **reinforcement learning**, which is the scientific study of how animals, humans, and machines adapt their behavior to maximize the cumulative reward received from the environment (Sutton & Barto, 2018). Reinforcement learning is also referred to as "learning with a critic"; the learner must discover the correct actions through trial and error, and in this sense reinforcement learning is theoretically distinguished from the supervised and unsupervised learning algorithms reviewed above.

Projective simulation

The models of equivalence class formation based on reinforcement learning have been developed in the framework of *projective simulation*. A projective simulator is an agent that learns from interactions with the environment and makes decisions based on its episodic memory network (Briegel & De las Cuevas, 2012; Boyajian et al., 2020; Melnikov et al., 2017). These models are captured as simple graphical networks that are flexible and adaptable, as they can be easily extended when more stimuli, scenarios, and variables are considered within a simulation.

The episodic memory component is a weighted network of *episodes* (Figure 6a). An episode is a unit in the network that represents either percepts (e.g., from stimuli) or actions (e.g., responses), and units are linked through connection weights. Learning in projective

simulation occurs by reconfiguration of the episodic memory network, either by updating the connection weights between units or by adding new units.

During a given trial (Figure 6a), a percept is observed by the agent, the corresponding episode (unit) is created/activated, which triggers a random walk on the episodic memory network. Then, when an action episode is reached, the agent performs this action. Evaluative feedback operates to reinforce or penalize this action. The probabilities for moving between episodes are based on the connection weights between them. When reward is received for a chosen action, the connections navigated for reaching this action are reinforced, and the occurrence of this behavior increases its probabilities.

Equivalence projective simulation

Equivalence projective simulation (EPS) is a modified version of projective simulation, mainly in that it includes symmetry connections and is designed to run simulations of matching-to-sample trials to model equivalence class formation (Mofrad et al., 2020). Additionally, in these memory networks, unit self-connections can be used to model reflexivity. The computational simulations in the EPS model (Mofrad et al., 2020) were designed with two phases: in the training phase the episodic memory is shaped, and in the test phase it is evaluated in its ability to cope with derived relations.

The training phase (Figure 6b) starts by showing a stimulus to the agent, for example, A1. A unit representing A1 is created in the memory space. Three or more comparison stimuli are shown to the agent, and the corresponding memory units are added to the model. At the beginning of training, all comparison stimuli may be selected with equal probabilities. However, during training the connection weights are updated based on reinforcement, and therefore the correct stimulus relations end up with stronger connections in the episodic memory network (Figure 6b).

In the EPS model (Mofrad et al., 2020), the symmetry relations are formed during training, and one assumption is that transitivity and equivalence relations are also acquired during training but their connection weights are calculated on demand upon test trials, which captures the finding that response latencies in transitivity and equivalence tests are typically longer than those of trained relations and symmetry tests (Bentall et al., 1993).

Several methods were proposed by Mofrad et al. (2020) to evaluate derived relations during the test phase, including max product, random walks on the memory network with absorbing action sets, and memory sharpness; these are schematically explained in Figure 7.

The EPS model (Mofrad et al., 2020) successfully simulated the three influential studies in the equivalence literature (Devany et al., 1986; Sidman & Tailby, 1982; Spencer & Chase, 1996) previously modeled by the neural network of Tovar and Westermann (2017). These results confirmed that, with a projective simulation approach, it is also possible to model the typical outcomes observed across a variety of training protocols and the atypical outcomes reported in participants with learning disabilities (Devany et al., 1986).

Moreover, to show the usefulness of EPS, the authors (Mofrad et al., 2020) presented an additional simulation experiment. They asked whether it was possible to obtain better equivalence yields than Devany et al. (1986) while training the same stimulus classes with the same number of trials but with a different training order. In the study by Devany et al. (1986), the group of children with intellectual and language disabilities failed to acquire the derived equivalence relations. The EPS model was run with the learning parameters that replicated learning disability in a variety of training schedules. The new simulation results suggested that there were training sequences that were more efficient for acquiring symmetry relations and consequently the formation of equivalence relations.

A Python implementation of the EPS model by Mofrad et al. (2020) is available to researchers interested in running simulation experiments at https://osf.io/grc2t/.

Enhanced Equivalence Projective Simulation (E-EPS)

In a follow-up study, Mofrad et al. (2021) presented an *enhanced* network with two main upgrading features: one is a computational *enhancement* of the memory network detailed below and the second is that the operation of E-EPS allows modeling the development of derived relations from an update process in the network instead of "producing" them on demand. These procedures allow modeling changes in both baseline and derived stimulus relations during tests.

Network enhancement for EPS. Network enhancement (Wang et al., 2018) is a computational method for *denoising* networks. It converts a noisy weighted network into a network with the same unit structure but adjusted weights (see Figure 8 for a schematic representation). The E-EPS model has a training phase similar to that for the EPS. But during tests, the structure of episodic memory in the agent changes through the network enhancement method. As a result, the E-EPS model actually retrieves the derived relations from its memory.

The E-EPS model (Mofrad et al., 2021) was used to study acquisition of equivalence classes under LS, MTO, and OTM training structures, and it was able to account for the results of prominent studies (Arntzen, 2012; Arntzen et al., 2010; Arntzen & Hansen, 2011), as it yielded better performance in OTM and MTO procedures compared with LS procedures. Finally, from a computational point of view, E-EPS has fewer parameters than EPS and is a much simpler yet accurate computational method.

A Python implementation of the E-EPS model by Mofrad et al. (2021) is available to researchers interested in running simulation experiments at https://osf.io/6czuj/.

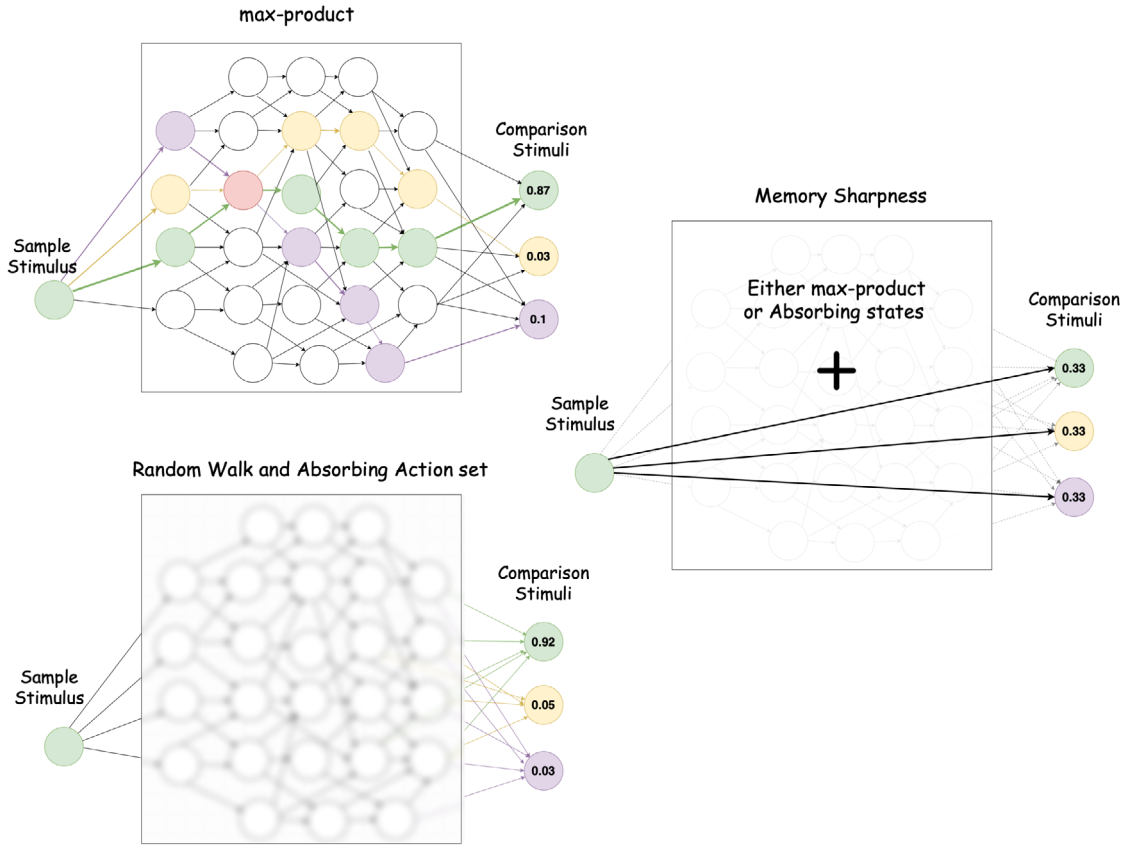

FIGURE 7 Testing methods in equivalence projective simulation. The first method is max product, which finds a path from the sample stimulus to the comparison stimuli with the maximum multiplicative probability. The agent chooses the comparison stimulus based on the calculated probability distribution (shown inside comparison units). The absorbing action sets captures the process of a random walk on the memory network. For this second method, the comparison stimuli are set as absorbing states, which means it is impossible to leave them once visited. The algorithm finds the probability of a random walk ended at each of the comparison stimuli, and like the max-product method selects one unit based on the highest probabilities. The last scenario is called memory sharpness, it combines using both directed connections and memory, as in the previous two methods. Memory sharpness gives the flexibility to model those situations when the baseline relations are acquired but the agent is unable to develop transitive and equivalence relations.

Contributions and future directions

The EPS and E-EPS models (Mofrad et al., 2020, 2021) have strengthened links between the general framework of reinforcement learning, the particular approach of projective simulation, and the study of stimulus equivalence.

As a bidirectional contribution for both equivalence research and projective simulation, Mofrad et al.'s (2020, 2021) approach demonstrates that the variety of computational methods used in the analysis of episodic memory networks during training and tests of equivalence results in the observation of different emergent properties in the memory network and its behavior. Through these methods, it is possible to model empirical results including typical class formation under OTM, MTO, and LS training; nodal distance effects; and atypical acquisition of stimulus relations.

Future research should explore additional methods of developing more advanced versions of projective simulation models. For example, modified versions of these models may be useful to capture other training procedures in addition to matching to sample, such as compound stimuli procedures (Debert et al., 2007; Fields, Doran, et al., 2009; Tovar et al., 2015; Tovar & Torres-Chávez, 2012). A possible approach to modeling compound stimuli is to implement projective simulation with generalization (Melnikov et al., 2017), which considers that each memory episode can be composed of different components; in a similar way, each episode may

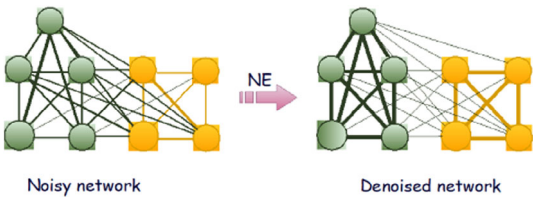

FIGURE 8 Network enhancement. The network enhancement takes a weighted network and then iteratively updates the network using the diffusion process. The thickness of connections represents a higher weight. This figure is a modified version of Figure 1 in (Wang et al., 2018), under License CC BY 4.0.

represent compound stimuli instead of single stimuli and relations between each episode component (i.e., each single stimulus) may emerge and reconfigure as a function of training regularities.

Finally, as in the previous models reviewed in this paper, future work with projective simulation models should seek to include more complex input patterns that can represent perceptual properties and differences between stimuli. These models should be informative on how perceptual and memory systems interact during complex learning, and by doing so, they will also be informative on how reinforcement learning is modulated under different perceptual restrictions.

CONCLUSION

For the last 30 years, computational models have provided tools for the theoretical development and experimental simulation of stimulus equivalence and symbolic behavior. In this review, we have documented several important contributions of these models, summarized here in two main categories. First, are theoretical advances on the core question of what mechanisms underlie equivalence class formation and symbolic behavior. This has been a matter of study since the first computational approach to stimulus equivalence (Barnes & Hampson, 1993). Computational models have now shown that equivalence relations emerge in systems with domain-general learning abilities (Lew & Zanutto, 2011; Mofrad et al., 2021; Ninness & Ninness, 2020; Tovar & Torres-Chávez, 2012; Tovar & Westermann, 2017). Notably, this finding directly contradicts other influential perspectives on symbolic behavior, as it is the Chomskyan approach (Berwick et al., 2013; Hauser et al., 2002), which postulates specific computational abilities underlying such behavior. Through computational simulations, it has been demonstrated that there is no need to use dedicated learning rules, systems, modules, or abilities to learn symbolic stimulus relations such as equivalence; instead, equivalence learners use general abilities such as associative, supervised, unsupervised, and reinforcement learning, but they use it efficiently enough to create internal models (i.e., representations) of complex environmental structure (i.e., stimulus regularities) that underlie symbolic behavior. The biologically inspired models of stimulus equivalence (Lew & Zanutto, 2011; Tovar & Westermann, 2017) have suggested the existence of a critical region in the space of learning parameters best suited for providing this efficiency, and by doing so these models have provided insight into understanding failure of equivalence class formation in other animal species and difficulties in human participants with learning disabilities.

Second, the computational models have extended the range of experimental tools for studying equivalence relations. Although in most areas of behavior and cognition, comparative research with nonhuman animals has yielded valuable insights and provided solid experimental models, stimulus equivalence is mainly an approach to symbolic human behavior, and as such there are no convincing animal models for its study. All disciplines studying human-only abilities deal with this methodological challenge (Marcus & Rabagliati, 2006). Remarkably, we have documented here that it is now possible to use computational models of stimulus equivalence to design experiments and test hypotheses for both describing and predicting human behavior. The models are particularly useful in exploring the effects of preexperimental repertoires, class structures, training protocols, and learning disabilities. Notably, the models of Mofrad, Ninness, and Tovar have provided software and resources for researchers interested in running simulation experiments in this field.

For each family of models reviewed here, we have suggested directions for future research that we believe are relevant and challenging for both equivalence researchers and computational modelers. There are several exciting developments waiting at the intersection of these fields that will promote a better understanding of stimulus equivalence and symbolic behavior in humans and artificial systems.

ACKNOWLEDGMENTS

This work was supported by a CONACYT grant to AET [CB 285152].

CONFLICT OF INTEREST

All authors declare that they have no conflicts of interest.

ETHICS STATEMENT

As this work is a review of past research, ethical approval is not required.

ORCID

Ángel Eugenio Tovar https://orcid.org/0000-0003-3669-5468

Erik Arntzen https://orcid.org/0000-0002-8471-1058

REFERENCES

Adams, B. J., Fields, L., & Verhave, T. (1993). Effects of test order on intersubject variability during equivalence class formation. *The Psychological Record*, *43*(1), 133–152.

Althaus, N., & Mareschal, D. (2013). Modeling cross-modal interactions in early word learning. *IEEE Transactions on Autonomous Mental Development*, *5*(4), 288–297. https://doi.org/10.1109/TAMD.2013.2264858

Andrade-Talavera, Y., Benito, I., Casañas, J. J., Rodríguez-Moreno, A., & Montesinos, M. L. (2015). Rapamycin restores BDNF-LTP and the persistence of long-term memory in a model of Down's syndrome. *Neurobiology of Disease*, *82*, 516–525. https://doi.org/10.1016/j.nbd.2015.09.005

Arnall, R., Garcia, Y., Griffith, A. K., & Spear, J. (2021). Stimulus generalization using nonvisual stimuli with a student who has autism spectrum disorder and visual impairment. *Journal of Visual Impairment & Blindness*, *115*(2), 121–133. https://doi.org/10.1177/0145482X21999503

Arntzen, E. (2012). Training and testing parameters in formation of stimulus equivalence: Methodological issues. *European Journal of Behavior Analysis*, *13*(1), 123–135. https://doi.org/10.1080/15021149.2012.11434412

Arntzen, E., Grondahl, T., & Eilifsen, C. (2010). The effects of different training structures in the establishment of conditional discriminations and subsequent performance on tests for stimulus equivalence. *The Psychological Record*, *60*(3), 437–461. https://doi.org/10.1007/BF03395720

Arntzen, E., & Hansen, S. (2011). Training structures and the formation of equivalence classes. *European Journal of Behavior Analysis*, *12*(2), 483–503. https://doi.org/10.1080/15021149.2011.11434397

Arntzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *The Psychological Record*, *47*(2), 309–320. https://doi.org/10.1007/BF03395227

Arntzen, E., & Mensah, J. (2020). On the effectiveness of including meaningful pictures in the formation of equivalence classes. *Journal of the Experimental Analysis of Behavior*, *113*(2), 305–321. https://doi.org/10.1002/jeab.579

Arntzen, E., & Sætherbakken, P. S. (2021). An overview of key papers preceding Sidman equivalence. *Journal of the Experimental Analysis of Behavior*, *115*(1), 224–241. https://doi.org/10.1002/jeab.663

Barnes, D., & Hampson, P. J. (1993). Stimulus equivalence and connectionism: Implications for behavior analysis and cognitive science. *The Psychological Record*, *43*(4), 617–638.

Barnes-Holmes, D., Finn, M., McEnteggart, C., & Barnes-Holmes, Y. (2018). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *Perspectives on Behavior Science*, *41*(1), 155–173. https://doi.org/10.1007/s40614-017-0124-7

Barnes-Holmes, D., & Harte, C. (2022). Relational frame theory 20 years on: The Odysseus voyage and beyond. *Journal of the Experimental Analysis of Behavior*, *117*(2), 240–266. https://doi.org/10.1002/jeab.733

Bates, E. (1999). Plasticity, localization and language development. In S. H. Broman & J. M. Fletcher (Eds.), *The changing nervous system: Neurobehavioral consequences of early brain disorders* (pp. 214–253). Oxford University Press.

Belisle, J., & Dixon, M. R. (2020). Relational density theory: Nonlinearity of equivalence relating examined through higher-order volumetric-mass-density. *Perspectives on Behavior Science*, *43*(2), 259–283. https://doi.org/10.1007/s40614-020-00248-w

Bentall, R. P., Dickins, D. W., & Fox, S. R. A. (1993). Naming and equivalence: Response latencies for emergent relations. *The Quarterly Journal of Experimental Psychology Section B*, *46*(2), 187–214. https://doi.org/10.1080/14640749308401085

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, *17*(2), 89–98. https://doi.org/doi.org/10.1016/j.tics.2012.12.002

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *2*(1), 32–48.

Bliss, T. V. P., Collingridge, G. L., & Morris, R. G. M. (2007). Synaptic plasticity in the hippocampus. In P. Andersen, R. G. M. Morris, D. G. Amaral, T. V. P. Bliss, & J. O'Keefe (Eds.), *The hippocampus book* (pp. 343–474). Oxford University Press.

Boyajian, W. L., Clausen, J., Trenkwalder, L. M., Dunjko, V., & Briegel, H. J. (2020). On the convergence of projective-simulation–based reinforcement learning in Markov decision processes. *Quantum Machine Intelligence*, *2*(2), 13. https://doi.org/10.1007/s42484-020-00023-9

Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, *2*(1), 400. https://doi.org/10.1038/srep00400

Burgos, J. E. (2007). Autoshaping and automaintenance: A neural-network approach. *Journal of the Experimental Analysis of Behavior*, *88*(1), 115–130. https://doi.org/10.1901/jeab.2007.75-04

Critchfield, T. S., Barnes-Holmes, D., & Dougher, M. J. (2018). Editorial: What Sidman did: Historical and contemporary significance of research on derived stimulus relations. *Perspectives on Behavior Science*, *41*(1), 9–32. https://doi.org/10.1007/s40614-018-0154-9

Cullinan, V. A., Barnes, D., Hampson, P. J., & Lyddy, F. (1994). A transfer of explicitly and nonexplicitly trained sequence responses through equivalence relations: An experimental demonstration and connectionist model. *The Psychological Record*, *44*(4), 559–585. https://doi.org/10.1007/BF03395144

Debert, P., Huziwara, E. M., Faggiani, R. B., De Mathis, M. E. S., & Mcilvane, W. J. (2009). Emergent conditional relations in a go/no-go procedure: Figure–ground and stimulus-position compound relations. *Journal of the Experimental Analysis of Behavior*, *92*(2), 233–243. https://doi.org/10.1901/jeab.2009.92-233

Debert, P., Matos, M. A., & McIlvane, W. (2007). Conditional relations with compound abstract stimuli using a go/no-go procedure. *Journal of the Experimental Analysis of Behavior*, *87*(1), 89–96. https://doi.org/10.1901/jeab.2007.46-05

Devany, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language-disabled children. *Journal of the Experimental Analysis of Behavior*, *46*(3), 243–257. https://doi.org/10.1901/jeab.1986.46-243

Dickins, T. E., & Dickins, D. W. (2001). Symbols, stimulus equivalence and the origins of language. *Behavior and Philosophy*, *29*, 221–244.

Dougher, M., Perkins, D. R., Greenway, D., Koons, A., & Chiasson, C. (2002). Contextual control of equivalence-based transformation of functions. *Journal of the Experimental Analysis of Behavior*, *78*(1), 63–93. https://doi.org/10.1901/jeab.2002.78-63

Dougher, M., Twohig, M. P., & Madden, G. J. (2014). Editorial: Basic and translational research on stimulus–stimulus relations. *Journal of the Experimental Analysis of Behavior*, *101*(1), 1–9. https://doi.org/10.1002/jeab.69

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fields, L. (2015). Stimulus relatedness in equivalence classes, perceptual categories, and semantic memory networks. *European Journal of Behavior Analysis*, *17*(1), 2–18. https://doi.org/10.1080/15021149.2015.1084713

Fields, L., Doran, E., & Marroquin, M. (2009). Equivalence class formation in a trace stimulus pairing two-response format: Effects of response labels and prior programmed transitivity induction. *Journal of the Experimental Analysis of Behavior*, *92*(1), 57–84. https://doi.org/10.1901/jeab.2009.92-57

Fields, L., Travis, R., Roy, D., Yadlovker, E., de Aguiar-Rocha, L., & Sturmey, P. (2009). Equivalence class formation: A method for teaching statistical interactions. *Journal of Applied Behavior Analysis*, *42*(3), 575–593. https://doi.org/10.1901/jaba.2009.42-575

Fields, L., & Verhave, T. (1987). The structure of equivalence classes. *Journal of the Experimental Analysis of Behavior*, *48*(2), 317–332. https://doi.org/10.1901/jeab.1987.48-317

Fienup, D. M., Covey, D. P., & Critchfield, T. S. (2010). Teaching brain-behavior relations economically with stimulus equivalence technology. *Journal of Applied Behavior Analysis*, *43*(1), 19–33. https://doi.org/10.1901/jaba.2010.43-19

Fienup, D. M., & Critchfield, T. S. (2010). Efficiently establishing concepts of inferential statistics and hypothesis decision making through contextually controlled equivalence classes. *Journal of Applied Behavior Analysis*, *43*(3), 437–462. https://doi.org/10.1901/jaba.2010.43-437

Gale, L., & Stewart, I. (2020). Assessing and training comparative relations in children with autism spectrum disorder. *Journal of European Psychology Students*, *11*(1), Article 1. https://doi.org/10.5334/jeps.487

Gallant, E. E., Reeve, K. F., Reeve, S. A., Vladescu, J. C., & Kisamore, A. N. (2021). Comparing two equivalence-based instruction protocols and self-study for teaching logical fallacies to college students. *Behavioral Interventions*, *36*(2), 434–456. https://doi.org/10.1002/bin.1772

García-García, A., Martín-Hernández, J. A., & Gutiérrez-Domínguez, M. T. (2010). Modelo computacional para la formación de clases de equivalencia. *International Journal of Psychology and Psychological Therapy*, *10*(1), 163–176.

Green, G., & Saunders, R. R. (1998). Stimulus equivalence. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 229–262). Springer. https://doi.org/10.1007/978-1-4899-1947-2_8

Guinther, P. M., & Dougher, M. J. (2015). The clinical relevance of stimulus equivalence and relational frame theory in influencing the behavior of verbally competent adults. *Current Opinion in Psychology*, *2*, 21–25. https://doi.org/10.1016/j.copsyc.2015.01.015

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley.

Hrbacek, K., & Jech, T. (1999). *Introduction to mathematical set theory* (3rd ed.). Marcel Deccer.

Imam, A. A. (2006). Experimental control of nodality via equal presentations of conditional discriminations in different equivalence protocols under speed and no-speed conditions. *Journal of the Experimental Analysis of Behavior*, *85*(1), 107–124. https://doi.org/10.1901/jeab.2006.58-04

Kastak, C. R., Schusterman, R. J., & Kastak, D. (2001). Equivalence classification by California sea lions using class-specific reinforcers. *Journal of the Experimental Analysis of Behavior*, *76*(2), 131–158. https://doi.org/10.1901/jeab.2001.76-131

Lew, S. E., Rey, H. G., Gutnisky, D. A., & Zanutto, B. S. (2008). Differences in prefrontal and motor structures learning dynamics depend on task complexity: A neural network model. *Neurocomputing*, *71*(13), 2782–2793. https://doi.org/10.1016/j.neucom.2007.09.010

Lew, S. E., & Zanutto, B. S. (2011). A computational theory for the learning of equivalence relations. *Frontiers in Human Neuroscience*, *5*, Article 113. https://doi.org/10.3389/fnhum.2011.00113

Lyddy, F., & Barnes-Holmes, D. (2007). stimulus equivalence as a function of training protocol in a connectionist network. *Journal of Speech and Language Pathology and Applied Behavior Analysis*, *2*, 14–24. https://doi.org/10.1037/h0100204

Lyddy, F., Barnes-Holmes, D., & Hampson, P. J. (2001). A transfer of sequence function via equivalence in a connectionist network. *The Psychological Record*, *51*(3), 409–428. https://doi.org/10.1007/BF03395406

Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*, 5–21. https://doi.org/10.1016/j.neuron.2004.09.012

Marcus, G., & Rabagliati, H. (2006). What developmental disorders can tell us about the nature and origins of language. *Nature Neuroscience*, *9*(10), 1226–1229. https://doi.org/10.1038/nn1766

Martín H., J. A., Santos, M., García, A., & de Lope, J. (2007). A computational model of the equivalence class formation psychological phenomenon. In E. Corchado, J. M. Corchado, & A. Abraham (Eds.), *Innovations in hybrid intelligent systems* (pp. 104–111). Springer. https://doi.org/10.1007/978-3-540-74972-1_15

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31. https://doi.org/10.1037/a0018130

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

Melnikov, A. A., Makmal, A., Dunjko, V., & Briegel, H. J. (2017). Projective simulation with generalization. *Scientific Reports*, *7*(1), Article 14430. https://doi.org/10.1038/s41598-017-14740-y

Mofrad, A. A., Yazidi, A., Hammer, H. L., & Arntzen, E. (2020). Equivalence projective simulation as a framework for modeling formation of stimulus equivalence classes. *Neural Computation*, *32*(5), 912–968. https://doi.org/10.1162/neco_a_01274

Mofrad, A. A., Yazidi, A., Mofrad, S. A., Hammer, H. L., & Arntzen, E. (2021). Enhanced equivalence projective simulation: A framework for modeling formation of stimulus equivalence classes. *Neural Computation*, *33*(2), 483–527. https://doi.org/10.1162/neco_a_01346

Ninness, C., & Ninness, S. K. (2020). Emergent virtual analytics: Modeling contextual control of derived stimulus relations. *Behavior and Social Issues*, *29*(1), 119–137. https://doi.org/10.1007/s42822-020-00032-0

Ninness, C., Ninness, S. K., Rumph, M., & Lawson, D. (2018). The emergence of stimulus relations: Human and computer learning. *Perspectives on Behavior Science*, *41*(1), 121–154. https://doi.org/10.1007/s40614-017-0125-6

Ninness, C., Rehfeldt, R. A., & Ninness, S. K. (2019). Identifying accurate and inaccurate stimulus relations: Human and computer learning. *The Psychological Record*, *69*(3), 333–356. https://doi.org/10.1007/s40732-019-00337-6

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64–99). Appleton-Century-Crofts.

Rueda, N., Flórez, J., & Martínez-Cué, C. (2012). Mouse models of Down syndrome as a tool to unravel the causes of mental disabilities. *Neural Plasticity*, *2012*, Article 584071. https://doi.org/10.1155/2012/584071

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.1038/323533a0

Schusterman, R. J., & Kastak, D. (1993). A California sea lion (*Zalophus Californianus*) is capable of forming equivalence relations. *The Psychological Record*, *43*(4), 823–839. https://doi.org/10.1007/BF03395915

Scott-McKean, J. J., & Costa, A. C. S. (2011). Exaggerated NMDA mediated LTD in a mouse model of Down syndrome and pharmacological rescuing by memantine. *Learning & Memory*, *18*(12), 774–778. https://doi.org/10.1101/lm.024182.111

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech Language and Hearing Research*, *14*(1), 5–13. https://doi.org/10.1044/jshr.1401.05

Sidman, M. (1992). Equivalence relations: Some basic considerations. In S. C. Hayes & L. J. Hayes (Eds.), *Understanding verbal relations. The second and third international institute on verbal relations* (pp. 15–27). Context Press.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Authors Cooperative.

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, *74*(1), 127–146. https://doi.org/10.1901/jeab.2000.74-127

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. https://doi.org/10.1901/jeab.1982.37-5

Spencer, T. J., & Chase, P. N. (1996). Speed analyses of stimulus equivalence. *Journal of the Experimental Analysis of Behavior*, *65*(3), 643–659. https://doi.org/10.1901/jeab.1996.65-643

Steele, D., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior*, *56*(3), 519–555. https://doi.org/10.1901/jeab.1991.56-519

Stromer, R., Mackay, H. A., & Stoddard, L. T. (1992). Classroom applications of stimulus equivalence technology. *Journal of Behavioral Education*, *2*(3), 225–256. https://doi.org/10.1007/BF00948817

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning* (second ed.). An introduction: MIT Press.

Tarbox, J., Szabo, T. G., & Aclan, M. (2020). Acceptance and commitment training within the scope of practice of applied behavior analysis. *Behavior Analysis in Practice*, *5*, 11–32. https://doi.org/10.1007/s40617-020-00466-3

Tovar, Á. E., Rodriguez-Granados, A., & Arias-Trejo, N. (2019). Atypical shape bias and categorization in autism: Evidence from children and computational simulations. *Developmental Science*, *23*(2), Article e12885. https://doi.org/10.1111/desc.12885

Tovar, Á. E., & Torres-Chávez, A. (2012). A connectionist model of stimulus class formation with a yes/no procedure and compound stimuli. *Psychological Record*, *62*(4), 747–762. https://doi.org/10.1007/BF03395833

Tovar, Á. E., & Torres-Chávez, A. (2021). Teaching symbolic relations in Down syndrome through equivalence-based instruction: A case study. *Mexican Journal of Behavior Analysis*, *47*(2), 368–391. https://doi.org/10.5514/rmac.v47.i2.81165

Tovar, Á. E., Torres-Chávez, A., & Ruiz, A. (2015). Effects of verbal-labeled responses on stimulus class formation in a compound stimulus procedure. *Mexican Journal of Behavior Analysis*, *41*(1), 68–85. https://doi.org/10.5514/rmac.v41.i1.63694

Tovar, Á. E., & Westermann, G. (2017). A neurocomputational approach to trained and transitive relations in equivalence classes. *Frontiers in Psychology*, *8*, Article 01848. https://doi.org/10.3389/fpsyg.2017.01848

Tovar, Á. E., & Westermann, G. (2023). No need to forget, just keep the balance: Hebbian neural networks for statistical learning. *Cognition*, *230*, Article 105176. https://doi.org/10.1016/j.cognition.2022.105176

Tovar, Á. E., Westermann, G., & Torres, A. (2018). From altered synaptic plasticity to atypical learning: A computational model of Down syndrome. *Cognition*, *171*, 15–24. https://doi.org/10.1016/j.cognition.2017.10.021

Vernucio, R. R., & Debert, P. (2016). Computational simulation of equivalence class formation using the go/no-go procedure with compound stimuli. *The Psychological Record*, *66*(3), 439–449. https://doi.org/10.1007/s40732-016-0184-1

Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C. D., Batzoglou, S., & Leskovec, J. (2018). Network enhancement as a general method to denoise weighted biological networks. *Nature Communications*, *9*(1), Article 3108. https://doi.org/10.1038/s41467-018-05469-x

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, Article e49547. https://doi.org/10.7554/eLife.49547

Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O'Donnell, T. J., Sainburg, T., & Gentner, T. Q. (2020). Five ways in which computational modeling can help advance cognitive science: Lessons from artificial grammar learning. *Topics in Cognitive Science*, *12*(3), 925–941. https://doi.org/10.1111/tops.12474

How to cite this article: Tovar, Á. E., Torres-Chávez, Á., Mofrad, A. A., & Arntzen, E. (2023). Computational models of stimulus equivalence: An intersection for the study of symbolic behavior. *Journal of the Experimental Analysis of Behavior*, *119*(2), 407–425. https://doi.org/10.1002/jeab.829