



Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Second-order correlation learning in 2- to 4-year-old children, and its underlying mechanism

Deon T. Benton

Department of Psychology and Human Development, Vanderbilt University, Peabody College, 230 Appleton Place #552, Nashville, TN 37235, USA

ARTICLE INFO

Keywords:

Causal learning
Artificial neural networks
Developmental mechanisms
Computational modeling

ABSTRACT

Second-order correlation learning—or the capacity to infer an indirect relation between features based on separate direct relations—has been studied extensively in children older than 4 years of age. Comparatively less research has examined whether children 4 years of age and younger can engage in this process. One of the few studies that assessed second-order correlation learning in children younger than this age showed that 2- to 3-year-olds can use this form of learning to make causal inferences in a discrimination context (e.g., Benton et al., 2021). The present study extended this work in two key ways. First, it tested whether 2- to 4-year-olds could detect and encode second-order correlations that are embedded in a cognitively more demanding category-like context consisting of multiple objects and then use those correlations to make causal inferences. Second, it explored a possible mechanistic basis of second-order correlation learning through a connectionist computational model. The results of the behavioral study revealed that children can detect second-order correlations among several distinct objects and use them to make causal inferences, and the computational model captured children's performance and elucidated a possible mechanism through which such correlations might emerge in the mind. Taken together, the results demonstrate that children as young as two can use second-order correlations to guide causal inference in more complex contexts and that this ability may emerge from more general processes of the mind based on representational overlap.

Few abilities are more important for navigating the world effectively than causal reasoning. The capacities believed to comprise this ability include distinguishing causes from non-causes, making inferences about potential causes of an effect based on observed data, intervening on causal systems to produce novel effects, and reasoning counterfactually about potential outcomes if the suspected cause is absent. There is now considerable evidence that each of these capacities undergoes a considerable developmental progression throughout the first several years of life. For example, early work showed that by 2 years of age children can categorize and distinguish causes from non-causes (e.g., Gopnik & Sobel, 2000). Between 3 and 4 years of age children can use different patterns of statistical evidence to make causal inferences, both in deterministic (e.g., Gopnik et al., 2001) and probabilistic (e.g., Kushnir & Gopnik, 2005; Griffiths et al., 2011) contexts, as well as in more realistic situations (e.g., Schulz & Gopnik, 2004). By 4 years of age children can reason counterfactually about causes and effects (e.g., Harris et al., 1996; Sobel, 2004; cf., Rafetseder et al., 2010) and between 4 and 8 years of age children can design interventions to bring about novel effects (e.g., McCormack et al., 2016; Schulz et al., 2007). Taken together, this research suggests that the components that comprise causal reasoning undergo a considerable developmental progression during the first several years of life.

One of the proposed cognitive mechanisms through which children are thought to acquire causal knowledge is associative learning

E-mail address: deon.benton@vanderbilt.edu.

<https://doi.org/10.1016/j.jecp.2026.106494>

Received 25 August 2025; Received in revised form 25 February 2026;

Available online 7 March 2026

0022-0965/© 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

(e.g., Benton, 2025; Cohen et al., 2002; Rakison & Lupyan, 2008). The crux of this learning process is that learners will perceive cues that are paired more frequently with an effect as better causal candidates than cues that are paired less frequently with the effect. One of the first attempts to formalize this idea was the Rescorla-Wagner model (Rescorla & Wagner, 1972). The crux of this model is that associative “links” between cues and effects are increased or decreased based on the cue’s ability to predict the presence or absence of an effect; the ability of the cue accurately to predict an effect depends ultimately on the cooccurrence of the cue and the effect. For example, if a cue fails to predict the occurrence of an effect—that is, if the cue is presented but a reasoner fails to anticipate the corresponding effect—then the link between the cue and the effect will be increased in strength. This ensures that the cue will better predict the effect in the future. In contrast, if a cue predicts an effect that does not actually occur, then the strength of the link between it and the effect will be diminished. This ensures that the cue will better predict the *absence* of the effect in the future.

Although this model can explain various effects in the animal conditioning literature (for a review, see Miller et al., 1995; Soto et al., 2023), it has faced criticism in the context of human causal learning. One critique is that the model struggles to explain “retrospective reevaluation” in individuals; that is, it fails to account for how people update their causal beliefs based on new information (e.g., Benton & Rakison, 2023; Lovibond et al., 2003; Griffiths et al., 2011; Sobel et al., 2004). Another limitation is that the model can only encode direct associations and cannot acquire indirect associations whereby causes are either separated in time and space (e.g., Holland, 1992) or are not directly linked to effects but instead are separated by intermediate levels of representation (Rumelhart et al., 1986). This weakness has led to an open question: Do infants and young children have access to more sophisticated forms of associative learning that eschew these limitations and that involve forming “distant” associations among elements? An example of such a mechanism is *second-order correlation learning*. This refers to the capacity to recognize that two features go together or are associated even if those features were never shown together. For example, if children learn that a particular tone of voice (feature A) is associated with surprised facial expressions (feature B) on one series of occasions and then that the same tone of voice (feature A) tends to occur whenever a novel object is present (feature C), a child capable of second-order correlation learning may infer that a surprised facial expression may indicate the presence of a novel object. Perhaps combined with a bias for novelty (e.g., Horst et al., 2011; Kucker et al., 2018), this may drive children to seek out the novel object.

Most of the research to date on second-order correlation learning in children has involved the use of “stem facts” (e.g., Bauer, 2021). The basic procedure involves presenting children with an initial stem fact such as “dolphins live in pods” followed by a brief filler activity. Children are then read a second, related stem fact such as “dolphins communicate by clicking and squeaking” followed by open-ended and forced-choice questions that require them to derive a “second order” fact from both stem facts (e.g., to derive the fact that “pods communicate by clicking and squeaking”). Some notable takeaways from this research include the fact that by 6 years of age, if not earlier (e.g., Bauer & San Souci, 2010), children can engage in this process in laboratory and classroom settings (Esposito & Bauer, 2017, 2019) and can derive and retain second-order facts across dissimilar contexts (Bauer et al., 2012), under high information-processing demands (e.g., Bauer & Larkina, 2017), over a one-week delay (e.g., Varga & Bauer, 2013; for evidence in 4-year-olds see Varga et al., 2016), and when the stem facts are presented in different formats (Esposito et al., 2021) and languages (e.g., Esposito & Bauer, 2019). This capacity has also been shown to predict academic achievement (e.g., Esposito & Bauer, 2018, 2024), to be positively related to verbal comprehension (Esposito & Bauer, 2018) and metacognition (e.g., Dugan & Bauer, 2022), and to underlie knowledge acquisition broadly (Bauer et al., 2024). Finally, children have been shown to use this process “recursively” such that 7- to 9-year-olds can use the derived second-order facts as initial stem facts in yet other second-order relations (Wilson & Bauer, 2021).

Despite this impressive body of research, it is limited in at least three important ways. A first limitation is that this capacity has rarely been tested in children younger than 4 years of age. One reason this research may be limited to children 4 years of age and older is that forming second-order correlations over separate text passages presupposes an ability to encode and comprehend the content that underlies those passages in the first place; children younger than 4 years of age may simply struggle with these processes, even if the ability for second-order correlation learning is fully operational by 4 years of age. This highlights the need to assess whether this capacity is already present in children younger than 4 years of age. Theoretically, this is an important question to address because a positive answer would suggest that the cognitive structure that underlies this capacity is in place before 4 years of age but could potentially be masked by performance limitations in more demanding tasks such as those involving the integration of information across multiple content-rich text passages.

A second limitation is that given that the bulk of the attention in this literature has been on *whether* children can engage in this process, there has been comparatively less focus on *how* second-order correlation learning is achieved; that is, how does this process work and how might it unfold in the human mind to support inferences of different sorts? One way to gain traction on this problem is through computational modeling. Broadly, computational modeling is the practice of implementing a verbal theory into a computer programming language, software package, or bespoke simulator to determine the implications of that theory on behavior. A strength of computational modeling—and the primary reason one is developed here (described in detail below)—is that if the model matches human performance in a simulation analogue of a human behavioral task, then the process through which the model achieved that match can be explicitly and precisely elucidated (McClelland, 2009). Such an exercise can provide valuable insights into possible corresponding mechanisms in the human mind.

A third limitation concerns the fact that it remains largely unanswered whether children can use second-order correlation learning in contexts beyond text-based ones. In the few cases in which second-order correlation learning has been tested in settings that do not rely on text passages, children were explicitly told in a child-friendly way that they would need to derive the second-order relation (e.g., Schlichting et al., 2022; Shing et al., 2019). Thus, these studies cannot shed light on whether children younger than 4 years of age can engage in *unprompted* second-order learning in a non-text-based setting. If children can engage in second-order correlation learning unprompted, then this would suggest that they possess a powerful learning mechanism for acquiring knowledge in the real world where indirect relations must often be inferred without explicit prompting or instruction.

To date, only a small number of studies have examined directly whether infants and young children can engage in unprompted second-order correlation learning in a non-text-based setting (e.g., Rakison & Benton, 2019; Yermolayeva & Rakison, 2016). One particularly relevant example of this comes from research by Benton et al. (2021). This study examined 2- and 3-year-old children's ability to use second-order correlation learning to make causal inferences in the context of two objects. The study consisted of two training phases, a test phase, and a memory check phase. During the *static training phase*, a red cube and a green cylinder were placed next to each other on a table in full view of the child—the red cube had a purple diamond sticker attached to the center of it; the green cylinder had a yellow circle attached to its center. During the *causal training phase*, the first two objects were removed from the table and placed out of sight, and two new objects were placed on the table. These objects were identical to the first two except that neither object had a sticker attached to its center. Children were then introduced to a machine that lights up and plays music (Gopnik & Sobel, 2000). The experimenter then demonstrated that one of the objects (e.g., the green cylinder) activated the machine and that the other object (e.g., the red cylinder) did not.

During the test phase, two new, identically shaped blue blocks were brought out. One of the blocks had the purple diamond sticker attached to it; the other object had the yellow circle attached to it. Children were then simply asked to make the machine go by placing one of the two blocks on the machine. If children detected a direct association between the green cylinder and the yellow sticker during the Static phase as well as a direct association between the stickerless green cylinder and the machine's activation during the Causal phase, then at test they should use the blue block with the yellow sticker to activate the machine and not the other block. Finally, during the *memory check phase*, children were simply asked to put the stickers on their original objects. This phase enabled Benton et al. (2021) to determine whether any difficulties children might have had in detecting the relevant second-order correlation were due to their inability to recall one of the first-order associations (e.g., the association between the green cylinder and the yellow sticker). The results indicated that only the children who passed the memory check were more likely to activate the machine with the block whose internal feature formed a second-order correlation with the machine's activation.

Although Benton et al. (2021) were among the first to examine second-order correlation learning in children younger than 4 years of age, their study suffers from notable weaknesses. One is that it did not examine the nature of second-order correlation learning or how it might be implemented in the mind. Although a model was developed to account for the behavioral findings, the authors did not assess how the model achieved second-order correlation learning. Given that a recognized benefit of computational modeling is that it can elucidate potential mechanisms and processes that might be in operation in the human mind (e.g., Benton, 2022; Mareschal & Thomas, 2007; Yermolayeva & Rakison, 2016), this represents a missed opportunity. A second limitation with this study is that it leaves unanswered whether children can use second-order correlation learning to make inferences in a *category context* that consists of several objects. Instead, Benton et al. (2021) tested whether children can engage in second-order correlation learning in a *discrimination* context. A discrimination context is one that contains only two objects that possess different features. In contrast, a category context consists of at least two groups of objects, where each group is defined by a distinct set of correlated features (e.g., Rakison, 2004; Younger & Cohen, 1986). One category might be $A^1B^1C^1$ and $A^1B^1C^2$; another category might be $A^2B^2C^1$ and $A^2B^2C^2$. Here, the letters A, B, and C refer to features (e.g., the color of an object, its causal status, and its shape), and the superscripts refer to a feature's value (e.g., a value of 2 for feature C could mean that the object is not causal).

Second-order correlation learning would be of limited usefulness as a cognitive mechanism if it only operated in discrimination contexts. This is principally because it is rarely the case in the real world that learning is constrained to single instances of objects or entities; instead, a more common situation is that reasoners must learn about groups or “categories” of objects and entities. Although second-order correlation learning may well be one of the cognitive mechanisms that support category learning and formation in the real world, it is important to be clear here that the current study did not test category learning per se; that is, the current study did not examine whether children can use second-order correlation learning to form categories and generalize to new category members. Instead, the goal was to assess whether children can deploy this mechanism to make inferences about unseen causal relations in a context in which different objects with a shared feature are presented across phases.

One potential drawback with testing whether children can extract second-order correlations that are embedded in a category context is that such contexts are inherently more demanding than discrimination contexts (cf., Younger & Cohen, 1983). This stems from the fact that category contexts necessarily include more objects than non-category contexts and thus may stretch children's processing abilities. One study that supports this contention showed that 22-month-olds infants failed to encode direct correlations between the body of an object and its moving parts or its global motion trajectory when those correlations were embedded in a category (Rakison, 2004). Equally plausible, however, is that the additional objects in a category context could facilitate young children's ability to encode second-order relations. For example, this could occur if these additional objects are treated as “noise” and are ignored. This contention is supported by a study by Rakison and Benton (2019). They showed that 20- and 26-month-olds successfully engaged in second-order correlation learning in a category context but not in a discrimination context. Given that Benton et al. (2021) showed that 2- and 3-year-olds can use second-order correlation learning to make causal inferences in a discrimination context, a failure to do so in the current study would suggest that the category context may have taxed children's processing capacities. In contrast, if children succeed in the current study, then this would suggest that they possess robust second-order correlation abilities.

Current study

The present behavioral study and computational model were designed with two goals in mind. First, the behavioral study assessed whether 2- to 4-year-olds could form second-order correlations over groups or categories of objects to make causal inferences. These ages were chosen for two reasons. First, this is when explicit causal reasoning begins to emerge in earnest (e.g., Gopnik & Sobel, 2000; Sobel et al., 2004). This contrasts with earlier forms of causal understanding that emerge during early infancy. One early form of causal

understanding is causal *perception*. This refers to the ability to distinguish causal events from non-causal events (e.g., Leslie & Keeble, 1987; Oakes & Cohen, 1990) and to attribute agency and reciprocity to different objects in the events (Cohen & Oakes, 1993; Cohen et al., 1999; Rakison et al., 2016). Another early form is causal *inference*. This refers to the ability to make inferences about the nature of a causal relation from incomplete or indirect information (e.g., Begus & Bonawitz, 2024; Kominsky & Carey, 2024; Kominsky et al., 2022; Muentener & Carey, 2010; Muentener & Schulz, 2014; Saxe et al., 2007; Saxe et al., 2005; Sobel & Kirkham, 2006, 2007; Spelke et al., 1995). What distinguishes these forms of causal learning from causal reasoning is that the tasks used to measure causal perception and inference do not require the kind of explicit, effortful, and hypothesis-driven reasoning that causal reasoning tasks tend to require. Second, the computational model explored whether a connectionist artificial neural network could capture children's responses as well as provide novel insight into the nature of second-order correlation learning. Specifically, a connectionist model was developed to simulate children's performance in the current experiment and to understand better how the model achieved second-order correlation learning. Connectionist models are computer programs that are based loosely on the architecture of the brain and that consist of layers of simple, neuron-like processing units. In a standard connectionist model, adjacent layers of units are connected to each other via synapse-like weights (see Singh et al., 2025 for a gentle introduction). Given the loose parallel between these models and the human brain as well as the fact that one of the goals of this study was to elucidate how second-order correlation learning might operate in children's minds, a connectionist model represented an ideal modeling formalism. Crucially, if it can be determined how second-order correlation learning is carried out in a model, this could clarify how such learning might unfold in children's minds. This kind of demonstration would go beyond a mere demonstration proof that a computational model can capture existing behavioral data to offering meaningful insight into how second-order correlation learning might be instantiated in the children's minds.

Behavioral study

Methods

Participants. Participants were 92 2- to 4-year-olds (54 boys and 38 girls; $M = 42.36$ months, range = 24.63–59.60 months, $SD = 10.58$). Sample size was determined based on a power analysis. This analysis—which used the “pwr” package in R—determined that 17 participants were needed for a one-tailed, intercept-only logistic regression to detect an effect size equivalent to Benton et al. (2021; $p = .68$, odds ratio = 2.125). Thus, the experiments reported here had more than sufficient power. Participants were 1.09% Asian/Asian American, 5.43% Black/African American, 15.22% Interracial/Mixed Race, 5.43% unknown, and 72.82% White/Caucasian. Seventeen additional children were tested but not included in the analysis: 4 were excluded for having a developmental delay, 6 for refusing to participate, 1 for being too fussy, 2 for generating uncodable data, and 4 for being not fully engaged/distracted.

This study was approved by the IRB at the institution at which this work was carried out. This study was not preregistered. All data can be found here: https://osf.io/7emg3/?view_only=2d036d5e3a9844a9ac461d0c99bef40e.

Materials. The experiment used a 5.98 cm × 28.79 cm computer-animated version of the blinket detector (Benton et al., 2024), four differently shaped objects, and two differently shaped and colored internal features. The shapes were a 3.88 cm × 4.36 cm green star, a 3.88 cm × 4.36 cm green circle, a 3.93 cm × 3.78 red square, and a 4.19 cm × 4.49 cm red diamond. The internal features were

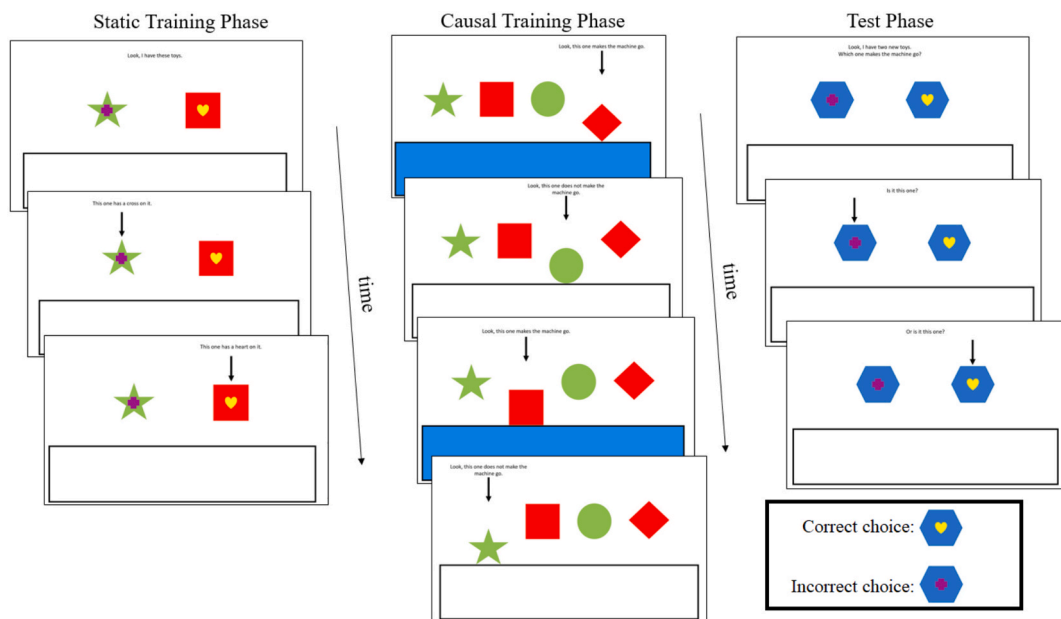


Fig. 1. An example sequence of events shown to participants. Children first participated in the static training phase (left), then the causal training phase (middle), and finally the test phase (right). The correct choice for this particular sequence of events is shown at the bottom right.

1.24 cm × 1.24 cm purple cross and a 1.25 cm × 1.25 cm yellow heart.

Procedure. Children were tested in a quiet room either at a local children’s science center or in a small, quiet, softly lit laboratory room. The experiment consisted of a pretraining phase, static and causal training trials, a test phase, and a memory check phase. Fig. 1 below shows the latter three phases in schematic form. Fig. 2 shows the final phase.

Pretraining phase

A pretraining trial preceded the main training phases to ensure that participants understood the purpose of the experiment. Instructional text—which the experimenter read as the trial unfolded—appeared on the screen in synchrony with the visual events to ensure that all participants received identical wording. Such text also appeared in the remaining phases. At the beginning of this trial, experimenters said, “We’re going to play a game with my machine. This is a very special machine. It’s my blicket machine. Bliackets make the machine go. So, let’s find all the blickets.” The machine itself was a rectangular box—in its “inactive” state, the box had a black border with a white interior. The experimenter then said, “Look, I have these toys. Let’s find the blickets. Watch what happens.” The toys—which were positioned above the machine—were a gray triangle (on the left) and a gray pentagon (on the right).

After approximately 8 s, the triangle descended until it contacted and activated the machine (i.e., the white region turned light blue). At this point, the experimenter said, “Look, this one makes the machine go”. The triangle then returned to its starting position above the machine by ascending upward. The pentagon then descended until it contacted and failed to activate the machine. The experimenter then said, “Look, this one does not make the machine go!”, at which point the object returned to its starting position above the machine. Both objects then descended together until they contacted and activated the machine. The experimenter then said, “Look, these make the machine go”. Both objects then returned to their starting positions.

The experimenter then asked, “Is this one a blicket?” while a downward facing arrow appeared above both objects in succession. The video was paused after each question to ensure that participants had sufficient time to answer. Although children did receive explicit prompting in the pretraining phase to identify the blicket, neither the blicket label nor any explicit prompting about the goal of the main task (i.e., to form a second-order link) was provided to children. This was critical given that, again, one goal of this study was to assess *unprompted* second-order correlation learning in 2- to 4-year-old children.

Training and test phases.

Static training phase. The static training trial began with a red square and a green star that were positioned next to each other above the rectangle. One object had a purple cross attached to its center (e.g., the red square); the other object had a yellow heart attached to its center (e.g., the green star). The specific object-feature pairings were counterbalanced across participants, as was the left–right position of the red square and green star. At the start of the static training phase, the experimenter said, “Look, I have these toys”. The experimenter then pointed to one object and said, “This one has a cross on it”. At the same time, a downward facing arrow appeared

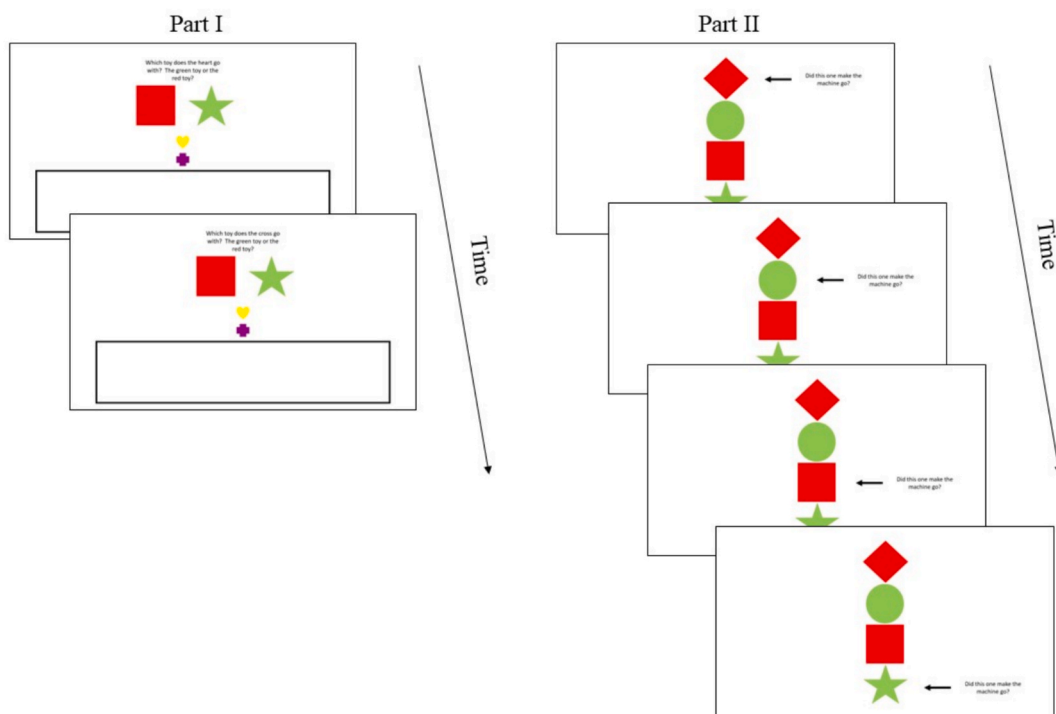


Fig. 2. An example sequence of events shown to participants during Part I (left) and Part II (right) of the memory check phase. Children first participated in Part I and then in Part II.

above the object (e.g., the one on the left in Fig. 1). The experimenter then pointed to the other object and said, “This one has a heart on it” while the downward facing arrow appeared above it (e.g., the one on the right in Fig. 1). To reinforce each object’s internal feature, the experimenter said, “Remember, this one has a cross on it” and “Remember, this one has a heart on it”. The same downward facing arrow appeared sequentially above each object at the time that it was referenced.

Causal training phase. Participants were then shown a new video with four objects, two of which were novel: the former green star, a novel green circle, the former red square, and a novel red diamond. These objects were positioned next to each other above the machine. Crucially, none of these objects possessed internal features and children were explicitly told that these objects differed from those presented in the Static phase. The green star then descended until it contacted and either activated or failed to activate the machine. This star then returned to its starting position above the machine, after which the red square descended until it contacted and either immediately activated or failed to activate the machine. This process continued until the red diamond descended onto the machine and returned to its starting position. For objects that activated the machine, the experimenter said, “Look, this one makes the machine go” (e.g., the red objects in Fig. 1); for objects that failed to activate the machine, the experimenter said, “Look, this one does not make the machine go” (e.g., the green objects in Fig. 1). Regardless of whether an object was causally efficacious, the same downward facing arrow appeared above each object while it was on the machine. For half of the participants, the red objects activated the machine; the green objects activated the machine for the remaining half. A Latin square was used to counterbalance the position of each object.

Test phase. Two new, identical blue hexagons then appeared next to each other above the machine. One of the hexagons had the original purple cross attached to its center, whereas the other hexagon had the original yellow heart attached to its center. At the beginning of this trial, the experimenter said, “Look, I have two new toys. Which one makes the machine go?”. The experimenter then inquired about the causal status of each object by pointing to each object and asking, “Is it this one?”. The object that was being asked about flashed briefly to guide children’s focus to the correct side of the screen.

Memory check phase. Children then participated in a two-part memory check phase. A depiction of this phase is shown in Fig. 2. Part I (left side of Fig. 2) began with the red square, green star, purple cross, and yellow heart. The red square and green star were positioned next to each other. The yellow heart and purple cross were positioned vertically, one (the purple cross) located directly above the other (the yellow heart), and both located below the star and square. The positions of each object were counterbalanced across participants. During this phase, participants were asked, “Which toy does the heart go with? The green toy or the red toy?”. Upon answering the question, participants were asked, “Which toy does the cross go with? The green toy or the red toy?”.

During Part II (right side of Fig. 2), the four objects that were shown during the Causal phase were arranged vertically in the center of the video. Their positions within this vertical arrangement were counterbalanced using a Latin square. For each object, the experimenter asked, “Did this one make the machine go?”.

Results

Analyses were conducted in RStudio 4.4.0 (Posit team, 2025) and p -values were supplemented with Bayes Factors (BFs) to quantify the relative support for the alternative hypothesis (that children engaged in second-order correlation learning) compared to the null hypothesis (that children were at chance); that is, a Bayes Factor indicates how much more likely the data are under the alternative hypothesis than under the null hypothesis. Bayes Factors were computed using the “BayesFactor” package in RStudio (Morey & Rouder,

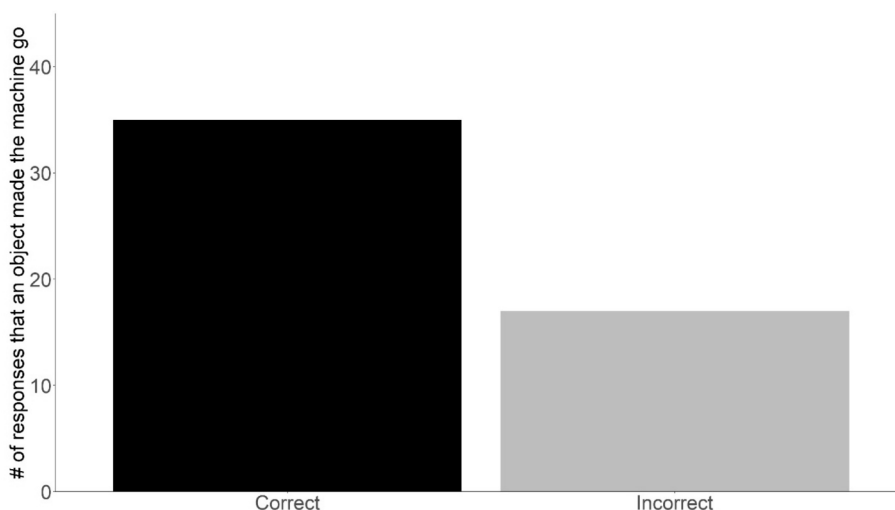


Fig. 3. Participants’ responses to the question of whether the Correct (i.e., the object with the feature that was indirectly associated with the machine’s activation) or Incorrect (i.e., the object with the feature that was indirectly associated with the machine’s non-activation) test object was causally effective.

2023). In addition, all analyses reported below focused on children who correctly answered both Phase-I memory check questions ($N = 52$); that is, it focused on only those children who correctly recalled both object-feature pairings from the first phase. The rationale for this was threefold. First, Benton et al. (2021) found that successful use of second-order correlation learning to make causal inferences was limited to children who recalled the original object-feature pairings. Second, success during the test phase requires, at minimum, that children encode these pairings. If a child who fails to encode this information also performs at chance during the test phase, then it becomes difficult, if not impossible, to determine why they failed. One possibility is that the child lacked the ability to engage in second-order correlation learning. Another is that they were simply off task or looked away during one of the training phases. In this latter case, it would have been impossible to form an indirect relation based on information presented across phases. Third, restricting the analysis to children who passed both Phase-I memory checks avoided the need for arbitrary decisions about inclusion thresholds. For example, if inclusion were based on children's performance during the second memory check phase alone (maximum score = 4), it would be unclear whether to focus on children who answered at least three questions correctly or those who answered all four correctly. Likewise, if inclusion were based on children's combined performance across both phases (maximum score = 6), there would be no principled justification for focusing on children who answered at least four questions correctly, at least five correctly, or all six questions correctly. Focusing on those who passed both Phase-I checks circumvented this problem and provided a clear, justified, and conservative inclusion criterion.

Preliminary analyses examined whether any of the counterbalancing variables—that is, participants' sex, the positions of the objects during the static and causal training phases, and the position of the objects during the main test—as well as participants' pretest performance related to their test choices. None of these factors were related to participants' test choices. All subsequent analyses therefore collapsed over these variables.

Fig. 3 shows participants' yes-no responses to the two test objects, and Table 1 shows the distribution of correct and incorrect choices. The main analysis assessed whether participants were more likely to choose the correct test object than the incorrect one. A second coder blind to hypotheses and conditions recoded participants' test choices in a random 25% of the video clips; inter-rater agreement was almost perfect ($\kappa = .83$). A one-tailed, logistic model showed that participants were more likely to choose the correct test object than the incorrect test object, odds ratio = 2.06, $SE = 0.30$, $z = 2.44$, 95% CI[1.17, 3.76], $p = .007$. The BF for this analysis indicated that the data were 5.48 times more likely under the alternative hypothesis than under the null hypothesis. As expected, participants who either were at chance or answered both questions incorrectly during the first memory check phase ($N = 40$) were as likely to choose the correct test object as they were to choose the incorrect test object, odds ratio = 0.90, $SE = 0.32$, $z = -0.32$, 95% CI[0.48, 1.69], $p = .75$, BF = 0.38.

A final analysis examined whether participants' (mean-centered) age related to their test performance. This analysis indicated that participants' age was not related to their test performance, odds ratio = 1.03, $SE = 0.04$, $z = 0.69$, 95% CI[0.95, 1.11], $p = .49$.

Discussion

This experiment examined whether 2- to 4-year-old children could use second-order correlation learning to make causal inferences in a category context. The results showed that children who remembered the original locations of the internal features were also more likely to choose the correct test object than the incorrect test object. This finding extends that by Benton et al. (2021) by demonstrating that children can use second-order correlation learning to make causal inferences in both discrimination and category contexts alike. Additionally, the fact that children were successful even when more objects were shown suggests that their second-order correlation learning is robust to greater contextual complexity.

Although the present experiment extends what is known about how children use second-order correlation learning to make causal inferences, outstanding questions remain. One especially important question concerns the mechanistic basis of second-order correlation learning; that is, it remains unresolved how second-order correlation learning might have operated in children's minds to support successful second-order correlation learning in this study as well as in that by Benton et al. (2021). By addressing the cognitive mechanism through which second-order correlation learning supports successful causal inferences, it is possible to move existing research on this topic beyond establishing developmental timetables such as *whether* and *when* children can engage in this process to elucidating *how* mechanistically this process might work.

One way to address this issue is via computational modeling. Specifically, if it can be shown that a model can engage in second-order correlation learning in a model analogue of the present behavioral study, then it can be taken apart to determine the basis of its success. In other words, the model's internal dynamics can be examined as they unfold over time to determine how they gave rise to second-order correlations. I built a connectionist computational model to address this issue. It should be noted that although Benton et al. (2021) also built a connectionist model and showed that it was capable of second-order correlation learning, they did not examine

Table 1

Frequency of children choosing the test object consistent with registering the second-order correlation. Top row shows data for children who recalled the location of the internal features during the first memory-check phase. The bottom row shows data for children who did not recall the location of such features during the first memory-check phase.

Correct Choices	Incorrect Choices	Total
35	17	52
19	21	40

how such correlations arose in the model. This means that Benton et al.'s (2021) simulations were limited to reproducing children's second-order correlation performance. The simulations presented next were carried out to confirm that a connectionist model can engage in second-order correlation learning and to explore how such correlations might arise.

Computational Model.

Method.

Network architecture. I built a three-layer autoencoder neural network that was trained using backpropagation in cross-entropy error. The learning rate ranged from 0.05 to 0.1 in increments of 0.025. In contrast, the number of training epochs in the Static and Causal training phases—which captured the number of times that the network weights were updated—ranged from 30 to 50 in increments of 10. Different values were used for the learning rate and the number of training epochs to ensure that the model's second-order correlation learning performance was not idiosyncratic to particular parameter values. Momentum and the number of hidden units were held constant at 0.9 and 15, respectively.

The input layer of the network consisted of patterns of activity across three different groups of input units. The groups were Shape, Color, and Feature (Fig. 4). All three groups used a localist coding scheme. The Shape group consisted of 20 units—different sets of four units were used to represent each object. Specifically, activation of the four units in row 1 corresponded to a square-shaped object, row 2 to a diamond-shaped object, row 3 to a star-shaped object, row 4 to a circle-shaped object, and row 5 to the test object. The color group consisted of three units. Each unit corresponded, respectively, to red, green, or blue. Finally, the Feature group consisted of two

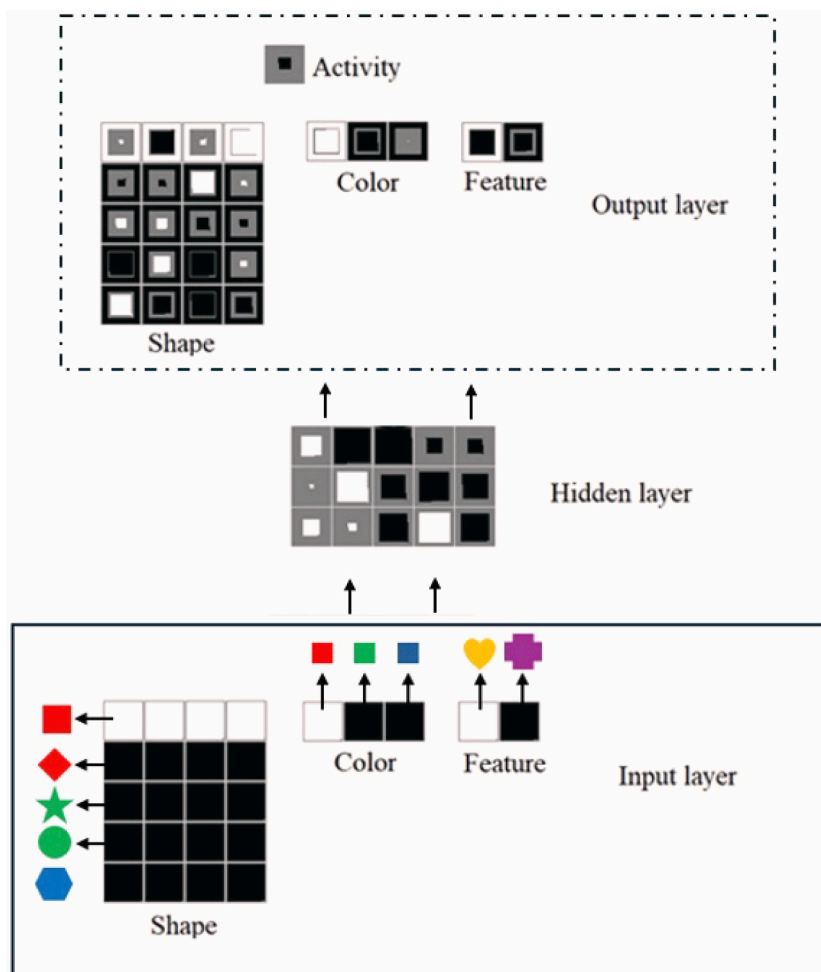


Fig. 4. Schematic of the version of the model that contained 20 units in the Shape group. The three input groups were connected to one hidden layer. The hidden layer, in turn, was connected to four output groups, three of which mirrored the three input groups. The task was to recreate the input pattern of activity along the three input groups on each of the corresponding output groups as well as to respond with a 0 or 1 in the Activity group if a causally ineffective or effective object was presented, respectively. The colored borders surrounding each output unit correspond to the target activation value for that unit. Black borders correspond to target activations of “0” (i.e., the unit should be “off”). White borders correspond to target activations of “1” (i.e., the unit should be “on”). For units in either the hidden or output layers, the size of the colored region within these borders corresponds to the actual activation value of that unit. Larger inner regions correspond to greater activations. Darker colored inner regions for the output units correspond to activation values ≤ 0.5 , whereas lighter colored regions correspond to activation values > 0.5 .

units. Each unit corresponded to the heart feature and cross feature, respectively. The output layer consisted of copies of the input groups along with an additional, one-unit Activity output group whose target was 1 whenever a causally effective object was presented as input to the model and 0 otherwise. The model architecture used here mirrored that in Benton et al. (2021) with one key exception: the current Shape group consisted of more units than that in Benton et al.'s (2021) model. This was necessary because the current study implemented a category context, whereas Benton et al.'s (2021) study instantiated a discrimination context. However, because the Shape group in both models contained many more units than either the Color or Feature groups, it may have a disproportionate influence on the model's overall second-order correlation performance than the remaining two groups. To address this possibility, the performance of the current model was compared to a second version of the model in which the Shape group consisted of five units: the first four units corresponded to the four training objects, and the fifth unit corresponded to the test object. If the 20-unit Shape group had an outsized influence on the model's behavior relative to the remaining groups, then the second-order correlation learning performance of the two models—that is, the original model with 20 units in the Shape group and the comparison model with 5 units in

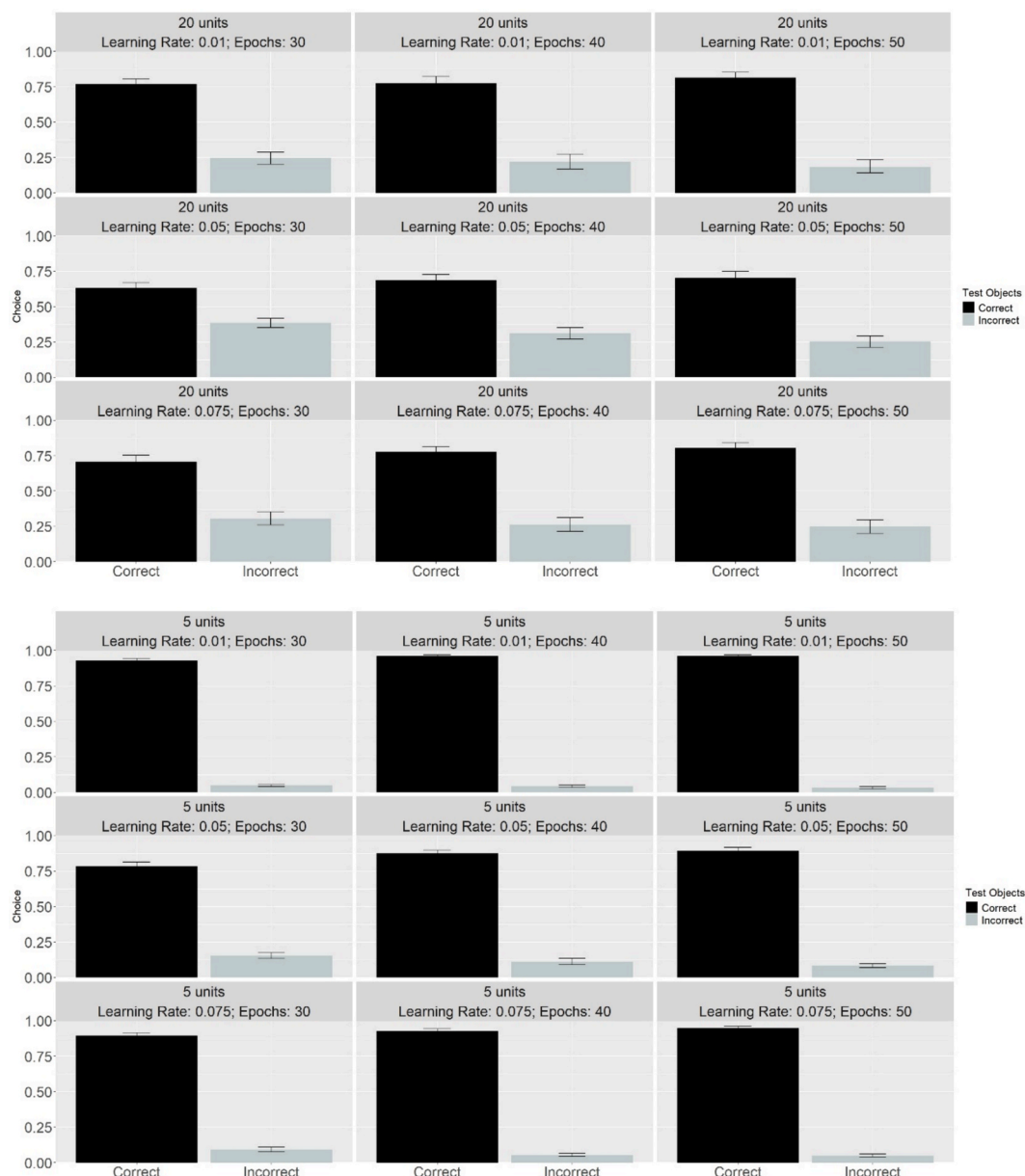


Fig. 5. Mean activation of the single Activity unit to the presentation of the Correct (i.e., the object with the feature that was indirectly associated with the machine's activation) and Incorrect (i.e., the object that was indirectly associated with the machine's non-activation) test objects across different combinations of learning rate and training epochs and for networks with 20 units in the Shape group (Top) and networks with 5 units in the Shape group (Bottom). Shaded regions represent ± 1 standard deviation.

the same group—should diverge qualitatively. All other aspects of this model were identical to the original one.

Training. Fifty networks were simulated for each combination of learning rate, training epoch, and model type (i.e., 5-unit Shape group vs. 20-unit Shape group; $50 \times 9 \times 2$ for a total of 900 networks). Each model run—which corresponded to a new participant—involved initializing a fresh set of small, random weights (distribution range = ± 0.8). During the Static training phase, networks learned that a yellow heart was appended to a red square and that a purple cross was appended to a green star. Critically, the Activity group, which encoded the causal efficacy of different objects, was functionally missing from the network during this phase; that is, the incoming weights to this group remain unchanged throughout the Static training phase. This design choice mirrored the behavioral study: Given that children were not presented with causal information during the Static training phase, the model also did not receive such information during the model analogue of this phase. During the causal training phase, networks learned that a featureless red square and red diamond (the first two rows of the Shape group) were causally effective, whereas a featureless green star and green circle (the next two rows) were causally ineffective. As with the Activity group during the Static training phase, the Feature group was functionally missing during the Causal training phase. Thus, networks faced the same task as children: they had to learn that feature information was related to causal information. However, because this information was presented in separate phases, networks had to integrate across these phases to discover the second-order correlation between object features and causal efficacy.

Testing. Immediately following the training phase, all networks were shown two identically shaped objects. This new object was instantiated by activating the units in the final row of the Shape group (and deactivating the remaining units). One of the test objects possessed the feature that formed an indirect association with the machine's activation across the Static and Causal training phases (i.e., the correct test object); the other object possessed the feature that formed an indirect association with the machine's non-activation (i.e., the incorrect test object). If networks could engage in second-order correlation learning in a category-like context with multiple objects, then the Activity group was predicted to be more active (i.e., produce greater activity) for the correct test object than for the incorrect test object. In contrast, if networks were unable to use second-order correlation learning to make causal inferences, then the activity of the single activity unit was expected to be statistically equivalent for the two test objects.

Results

Fig. 5 shows the mean activation of the Activity unit in response to the Correct and Incorrect test object across the different combinations of learning rate and training epochs and for different network types (i.e., networks with 20 units in the Shape group and those with 5 units in this group). The main takeaway from this figure is that networks produced more activity in the Activity group in response to the correct test object than in response to the incorrect test object for all parameter combinations, although the effect appears to be stronger for networks with 5 units in the Shape group than those with 20 units in this group. Thus, networks, like the children in the behavioral experiment, successfully engaged in second-order correlation learning. This finding not only replicates the main behavioral result but also indicates that networks' ability to engage in second-order correlation does not depend on a particular value for learning rate, particular number of training epochs, or the number of units in the Shape group.

These qualitative observations were also supported by quantitative analyses. The analysis assessed whether, across the different combinations of learning rate and training epochs, mean activation of the Activity unit was greater when the correct test object was shown than when the incorrect test object was shown. Data were entered into a linear model with Object Type (Correct vs. Incorrect) as the within-subjects factor and Network Type (networks with 5 units in the Shape group vs. networks with 20 units in this group) as the between-subjects factor. This analysis yielded a significant main effect for Object Type, $F(1, 898) = 20661.11$, $p = 0$. This result reflected the fact that the Correct object ($M = .82$, $SD = 0.15$) produced greater activation in the Activity unit than the Incorrect object ($M = .17$, $SD = 0.16$). In addition to this main effect, there was a significant interaction between Network Type and Object Type, $F(1, 898) = 1565.64$, $p = 0$. As shown in Fig. 5, this interaction reflected the fact that the difference in activation of the Activity unit between the Correct and Incorrect objects was greater for the networks with 5 units in the Shape group ($M = .83$, $SD = .11$) than for networks with 20 units in that group ($M = .47$, $SD = .16$), $t(802.59) = 39.57$, $p = 0$. This difference arose because networks with 20 units in the Shape group had more active input units overall (25 total input units) than networks with 5 units (10 total input units). As a consequence, the net input to each to hidden unit was larger for the former networks than the latter ones. The increased activity at the hidden layer for the 20-unit networks, in turn, produced larger net inputs to the output units than the 5-unit networks. Despite this difference in overall activation levels, both types of networks—like participants in the behavioral experiment—successfully engaged in second-order correlation learning.

Discussion

The computational model was designed with two goals in mind. First, it assessed whether a connectionist computational model could engage in second-order correlation learning in a context with multiple objects. Second, the model provided an opportunity to shed light on how, mechanistically, second-order correlations arose in the model. The current simulation results presented thus far speak only to the first goal. Specifically, the results indicated that a connectionist model could robustly engage in second-order correlation learning. However, it remains to be seen how second-order correlations emerged in the model and supported successful causal inferences. The next section addresses this issue.

How—or through what mechanism—did second-order correlations arise in the model?

One way to answer this question is to track the model's behavior at different points in training. During the Static training phase, two objects (e.g., a red square object with a yellow heart and a green star object with a purple cross) are presented to the model as a pattern of activity across the Shape, Color, and Feature input groups (Fig. 5A). This pattern, in turn, evokes a corresponding pattern of activity

across the hidden units via the weights that connect the input layer to the hidden layer. This pattern of activity then elicits a pattern of activity across the units that comprise the Shape, Color, and Feature output groups via the second set of weights that connect the hidden layer to the output layer. At this point, the model compares this pattern of activity to the desired or “target” pattern of activity. In this case, the target pattern of activity is just a copy of the input pattern of activity. At the beginning of training, the observed output pattern of activation necessarily differs from the target output pattern of activity. This can be seen clearly in Fig. 5A by noting that the internal region of each output unit (the actual activation of the unit) differs markedly from the outer region (the target activation of the unit). The model treats this discrepancy as error, which it attempts to minimize by iteratively adjusting the weights throughout the model via a learning procedure.

This process will tend to produce a pattern of activity across the hidden units (in response to input to the model) that is sufficient to reproduce that input activity at the output layer for both objects. This can be seen above in Fig. 5B; that is, it should be clear from the figure that, following the static training event, the activation of the output units mirrors the pattern of activation along the input groups. It should also be clear from Fig. 5B that the model has discovered largely non-overlapping patterns of activation at the hidden layer for both objects (i.e., the hidden units that are “on” for the red square object with a yellow heart [left] are largely off for the green star with a purple cross [right] and vice versa).

Such disparate patterns derive from the fact that model is trained to produce non-overlapping patterns of activation along the output groups for each object. In other words, because the pattern of error—which is operationalized as the difference between a given unit’s target activation and its actual activation—differs for both objects, the weights to and from the units that denote each object move in different directions during training. These differential weight changes lead to an alternating pattern of weights in the Feature, Color, and Shape groups. This can be seen in Fig. 6B and 6C. Consider the Shape group, for example. In this group, the weights that are positive for the red square with a yellow heart (input units 1–4) are largely negative for the green star with a cross (input units 5–8), and vice versa. It should also be clear from the figure that this pattern is largely absent before any training has taken place and the network has learned anything (Fig. 6A), Fig. 7..

Because the input pattern of activity presented to the model during the Causal training phase (e.g., a red square *without* a yellow heart) will be similar to that during the initial Static training phase (e.g., a red square *with* a yellow heart), the pattern of activation evoked across the hidden layer by this new object will be similar to that produced by the same object with the feature attached to it. This can be seen by comparing the input to the model between the Static (Fig. 5B) and Causal phases (Fig. 5C) and noting that the input differs in only one respect: feature information is present during the Static but absent during the Causal phase. Given that there is

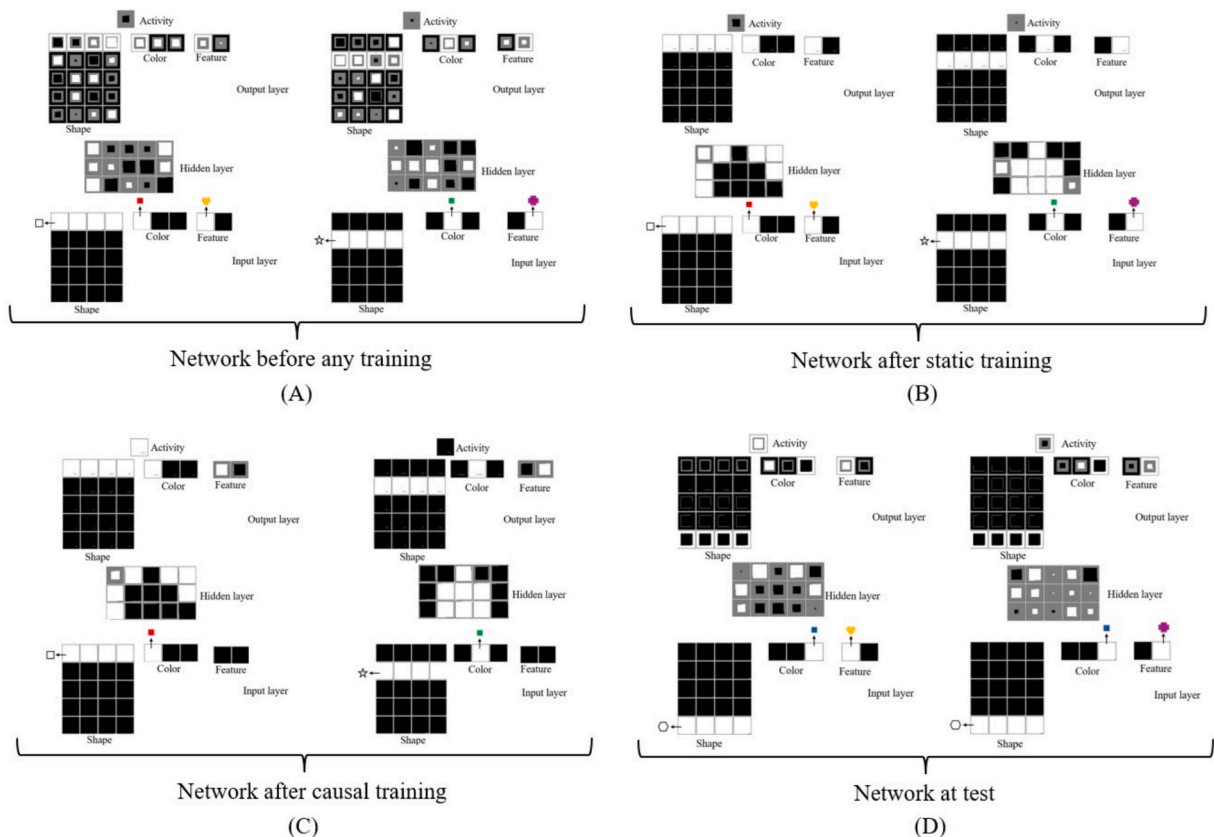


Fig. 6. A depiction of the model at various points during training. (A) The model before any training; (B) the model following the static training phase; (C) the model following the dynamic training phase; and (D) the model at test.

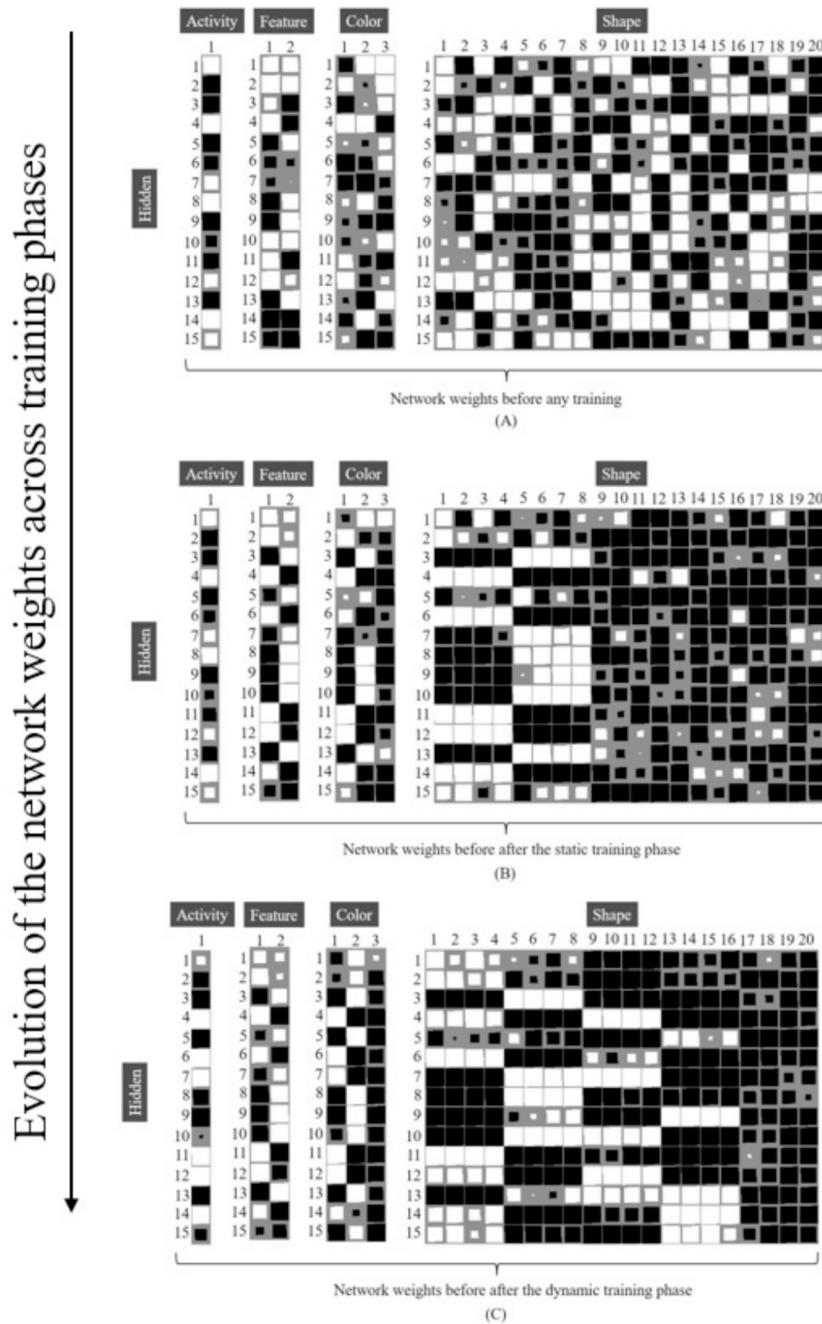


Fig. 7. Visualization of the weights connecting the hidden layer to the four output groups (i.e., Activity, Feature, Color, and Shape). Hidden units are listed vertically; individual units within a group are listed across the top. Each cell corresponds to the weight on the connection between a particular hidden unit and a unit in a given output group. For example, the top-left unit in the Shape group corresponds to the connection between the first hidden unit and the first unit in the Shape group. The size of the colored region corresponds to the magnitude of a weight; larger regions reflect larger weights. The color of the region corresponds to a weight's sign: darker regions correspond to negative weights; lighter colored regions correspond to positive weights.

significant overlap in the information presented during the two phases, this means that the target pattern of output activity will be similar to that in the Static training phase with one key exception: to the network now must activate the single activity unit in the Activity output group if a causal object is presented to the model (e.g., left side of Fig. 5C); if instead a non-causal object is presented to the model, its task is to deactivate this unit (right side of Fig. 5C). In Fig. 5C, for example, the model learns to activate the Activity unit in the presence of the red square because it is causal. However, the model learns to deactivate the Activity unit when shown the green

star because it is non-causal. The fact that the network's task during the Causal phase is largely consistent with its task during the Static phase—to reproduce at the output layer the input pattern of activation along the Color, Feature, and Shape groups—means that the error signal at the output layer produced by the presentation of a red square and green star will also be similar between the two phases. Additionally, because there is considerable overlap between the information presented to the model as input during the Static and Causal phases, the pattern of activity at the hidden layer that both objects elicit will also be similar. However, because the network will have discovered a set of weights that generates a pattern of activation across the hidden units during the Static phase that is sufficient to reproduce each object at the output layer, any error that remains during the Causal training phase will mostly be coming from the single activity unit in the Activity group. Instead of adjusting the weights radically to discover a whole new hidden pattern of activity that is sufficient both to reproduce the input activity across the Shape, Color, and Feature groups at the output layer and to activate the single activity unit, the network will tend to make small adjustments to the weights that connect the hidden layer to the Activity group; the weights between the Shape, Color, and Feature input groups and the hidden layer as well as those between the hidden layer and the Shape, Color, and Feature output groups will remain largely unchanged. This can be seen in Fig. 6 by comparing the network weights after the Static training phase (6B) to those following the Causal training phase (6C). As should be clear from the figure, the pattern of weights in the Color and Feature groups overall remains unchanged between both phases, and any changes that are made are quantitative rather than qualitative. The pattern of weights in the Shape group also remains mostly unchanged for the red square (units 1–4 along the top) and green star (units 5–8 along the top), although it can be seen in the figure that a new alternating pattern of weights has emerged across some of the columns. This is due to the fact that two additional objects—namely, a red diamond (units 9–12) and a green circle (units 13–16)—are introduced during the Causal phase; these objects are not shown in Fig. 5. The explanation for why the pattern of weights for the two new objects alternates is identical to that above. In contrast, the pattern of weights in the Activity group changes substantially between both phases. The network will continue to make these adjustments until it learns to activate the Activity unit for the causal objects (i.e., the red ones) and to deactivate this same unit for the non-causal objects (i.e., the green ones).

A consequence of this learning process is that when a blue hexagon with a yellow heart is subsequently presented as input to the model, the heart will evoke the pattern of activity at the hidden layer that this unit produced during the Static phase. It should be clear from Fig. 5D that the pattern that is evinced at the hidden layer resembles that produced by the red square during the Static and Causal phases. This suggests that the model “sees” this object as similar to the red square. This pattern, in turn, will activate the single activity unit and the yellow heart unit. It should also be clear from the figure that the elicited pattern of hidden activation activates the output units that correspond to the color of the formerly trained red square. This stems from the fact that, because *red* and *heart* were co-present during the Static training phase, they both became linked to the same pattern of hidden activation. As a consequence, when this pattern is re-instantiated, both features are activated. In contrast, presenting the model with a blue hexagon with a purple cross elicits a different hidden activation pattern that causes the activity output unit to turn off and the output unit corresponding to the cross feature to become active. Although the overall pattern of activation of hidden activation bears less of a resemblance to that produced by the green star during the Static and Causal training phases, many of the key units that reproduce the pattern of activation that corresponds to the green star and its causal status are activated.

To summarize, the model establishes a second-order correlation between two features by transforming what initially is an indirect association between an object's internal feature and its causal effectiveness into a direct one; that is, the model links feature information *and* causal information to a common hidden representation. This enables the model effectively to reduce a second-order problem into a first-order one; that is, activating the hidden representation via one cue (e.g., a yellow heart) is sufficient to trigger the other cue (e.g., causal efficacy) because both cues are linked directly to that pattern. This transformation is the crux of how second-order correlations emerged in the model and may reflect how such relations emerge in the human mind, although this latter point will require additional empirical work that is beyond the scope of the current paper.

Before closing, it is worth noting that this is not the only proposal for how second-order correlation learning might work. For example, Kumaran and McClelland (2012) introduced a model called REMERGE (for recurrency and episodic memory results in generalization) that provided a process-level account of second-order correlation learning. The model was based on a study by Bunsey and Eichenbaum (1996). To my knowledge, this is the first study to investigate second-order correlation learning, albeit in rats. In the study, rats first learned that a cue odor A predicted a rewarded choice odor B and an unrewarded choice odor Y, whereas a cue odor X predicted a rewarded choice odor Y and an unrewarded choice odor B. Then rats learned that odor B, now the cue odor, predicted a rewarded choice odor C and an unrewarded choice odor Z; they also learned that the now cue odor Y predicted a rewarded choice odor Z and an unrewarded choice odor C. Rats were then presented with cue odor A and required to choose between odors Z and C. The results indicated that intact rats chose odor C presumably because it was indirectly related to A.

Kumaran and McClelland's (2012) model of this task consisted of a six-unit “feature” layer. These units represented one of the six possible odors, each corresponding to one of the six odors in Bunsey and Eichenbaum (1996; i.e., A, B, C, X, Y, and Z). This layer was bidirectionally connected to a four-unit “conjunctive” layer. These units represented the associations learned during training (AB, BC, XY, and YZ). This layer, in turn, was connected to a four-unit response layer—each unit represented one of the four possible choice options that rats were presented with in Bunsey and Eichenbaum (1996; i.e., B, C, Y, and Z). The mechanism by which Kumaran and McClelland's (2012) model captured the behavior of the rats (i.e., the choice of C over Z) was that when the feature units A, C, and Z are activated, their activity causes the conjunctive units AB, BC, and YZ to activate. Activity from the conjunctive units in turn activates the feature units A, B, C, Y, and Z due to their bidirectional connectivity with the feature layer. Given that feature unit B is now more active than unit Y—because it is receiving input from the AB and BC conjunctive units, whereas feature unit Y only receives input from conjunctive unit YZ—the BC conjunctive unit will become more active than the YZ conjunctive unit on the next cycle. A corollary of this is that when activity then propagates from the conjunctive layer to the response layer, response unit C will be more active than

response unit Z, thereby simulating rats' choice of odor C over odor Z.

An important difference between the current account of second-order correlation learning and [Kumaran and McClelland's \(2012\)](#) account of it is that REMERGE's ability to encode indirect relations depended at least in part on the presence of the conjunctive *localist* units; in other words, these associations were stipulated rather than learned. In contrast, the current model's ability to encode indirect relations depended crucially on *learned* distributed representations. The model that the current one perhaps bears the strongest resemblance to is [Zhou et al.'s \(2023\)](#) C-HORSE (for complementary hippocampal operations for representing statistics and episodes) model. This model was designed to capture second-order correlation learning in human adults. Although the details of this model are beyond the scope of this paper, the success of the current model and C-HORSE depended largely on learned distributed representations. Yet, this model also differed from the current one in a crucial way. This concerns the fact that C-HORSE's ability to encode indirect associations depended crucially on the order in which the stimuli were presented to the model; no such constraint applied to the current model. That said, the current model may represent a synthesis between REMERGE and C-HORSE. For example, by representing each object locally (at the input and output layers), the current model captures a key feature of REMERGE. Yet, because the current model's ability to encode indirect relations depended on learned distributed representations, it has much in common with C-HORSE.

Given the scarcity of research on second-order correlation learning in children under 4 years of age, much less research into the nature of that learning in this population, more work is ultimately needed to determine two open questions. First, how is second-order correlation learning implemented in the minds of children younger than 4 years of age and how are the indirect relations that emerge from this learning represented (i.e., locally, in a distributed fashion, or some combination of both). One way to answer this question is to combine functional brain imaging with techniques for measuring similarity among representations (e.g., RSA). Second, does the mechanistic basis for second-order correlation learning change across development, and if so, in what ways?

General Discussion

The current study tested whether 2- to 4-year-olds could use second-order correlation learning to make causal inferences in a category-like context. The results indicated that children could engage in this process, but only if they remembered which objects possessed which features. A connectionist computational model was subsequently developed to provide mechanistic insight into how second-order correlations might emerge in the mind. An analysis of the model's internal dynamics revealed that such correlations emerged through a common representation that was linked to an object's feature and its corresponding causal status. Taken together, these results indicate that children younger than 4 years of age can use second-order correlation learning to discover causal relations.

More generally, the current results extend [Benton et al. \(2021\)](#) and the work of Bauer and colleagues (see [Bauer, 2021](#)). Specifically, the present study demonstrates that children can extract second-order correlations in a causal, category-like context beyond the traditional stem-fact paradigm and that such correlations may arise when their components are linked through a shared representation. Additionally, the current findings suggest that the cognitive machinery required to engage in this process may be in place before children's fourth birthdays. The findings also shed light on the very nature of this ability by showing that similar processes can support stem-fact derivation and causal judgments. Together with evidence that children use this process to learn about object motion ([Rakison & Benton, 2019](#)), object properties ([Yermolayeva & Rakison, 2016](#)), and word meanings ([Savic et al., 2023](#); [Unger et al., 2024](#)), these results suggest that second-order correlation learning may well be a domain-general mechanism.

Before concluding, several potential criticisms of the current study are worth noting. The first potential criticism is that because two of the four objects shown during the Causal phase were also present during the Static phase, children may have selectively focused only on these shared objects and ignored the two novel ones during the second phase. If this were the case, this would entail that the current study is more akin to a discrimination context than to a category context and thus represents a replication rather than an extension of [Benton et al. \(2021\)](#). One concern with this account is that it leaves open the question of why children would be compelled to attend to the shared objects over the novel ones in the first place. One possibility is that because the shared objects were spatially adjacent during the Static phase, children were more likely to track them across phases because they treated them as a unit. If this were the case, then children for whom the shared objects remained spatially adjacent during the Causal phase should have shown superior second-order correlation performance relative to those for whom this spatial adjacency was disrupted during the Causal phase. However, a mixed-effects logistic regression, with Adjacency Type as the sole fixed effect and subjects as a random effect, revealed no reliable difference between the groups, odds ratio = 1.41, $SE = 0.67$, $z = 0.51$, 95% CI [0.36, 5.16], $p = .61$. This suggests that spatial proximity did not drive children's performance.

Although the results above rule out a spatial-adjacency account, they do not rule out the possibility that children may have treated two of the four objects shown during the Causal phase as the same, but now rotated, objects that were presented during the Static phase; that is, children may have erroneously believed that they were looking at the backs of the two objects that were shown during the Static phase. There are two reasons to question this interpretation. First, children were explicitly told that the objects shown during the Causal phase were different toys from those shown during the Static phase. The rationale for this was to encourage children to treat Causal-phase objects as new instances rather than as previously seen objects. Although it remains possible that some children failed to encode this information, the fact that they were made aware that different objects were used across phases reduces the plausibility of a simple identity-tracking account.

Second, even if some children treated the Causal-phase objects as the same ones from the Static phase but viewed from a different angle, success in the task would still require them to infer the presence of the internal feature in its absence. The reason is that participants were shown featureless objects during the Causal phase. Thus, any representation of an internal feature at that point must be internally rather than externally generated. Crucially, recalling the internal feature of an object when it is not present in the input is entirely consistent with the mechanism through which second-order correlations emerged in the model. Recall that after the model learns during the Static phase that a particular object possesses an internal feature, presenting that object without the feature during the Causal phase activates a representation that contains both the object's appearance and its learned internal feature. Because this

representation is active while the model learns that the object is causally effective, its causal efficacy becomes linked to that representation. A corollary of this for children's performance in the behavioral study is that even if they perceived the Causal-phase objects to be turned around, successful second-order correlation learning at test would still reflect integration of information across phases via a shared internal representation—precisely the mechanism proposed by the model.

More broadly, the current study differs from Benton et al. (2021) in several important respects. First, it tested a wider age range than that in Benton et al. (2021). Specifically, Benton et al. (2021) tested 2- to 3-year-olds, whereas the 2- to 4-year-olds were tested in the current study. The rationale for widening the tested age range was that it was unclear whether 2- and 3-year-olds would be able to engage in second-order correlation learning in a dynamic context that contained more objects than that used in Benton et al. (2021). This approach also made it possible to assess whether, under the current conditions, this process underwent a more protracted developmental progression. Second, although Benton et al. (2021) showed that a connectionist computational model could capture the performance of the 2- and 3-year-olds in their study, they did not discuss how second-order correlations emerged in the model. Thus, Benton et al.'s (2021) presentation of their model provided little insight into the possible mechanistic basis of second-order correlation learning.

In sum, the current experiments and computational model not only reveal that children between 2 and 4 years of age can use second-order correlation learning to make causal inferences within a category context, but they also suggest a potential mechanism for how this learning might be implemented in the mind. These findings contribute to a growing body of research demonstrating that second-order correlation learning is a powerful mechanism by which children represent higher-order, indirect relations and generalize these relations to novel features.

CRedit authorship contribution statement

Deon T. Benton: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I thank Rebecca Beaton and members of the Computational Cognitive Development Lab at Vanderbilt University for assisting with data collection and participants and their caregivers for kindly agreeing to participate in this research. I also thank Amy Needham and Veronica Hearne for helpful comments on an earlier version of this manuscript. Finally, I would like to give special thanks to the Adventure Science Center in Nashville, TN for their generosity in allowing us to collect data on site. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. This study was not preregistered. The data necessary to reproduce the analyses presented here are publicly accessible at: https://osf.io/7emg3/?view_only=2d036d5e3a9844a9ac461d0c99bef40e. The analyses presented here were not preregistered.

Data availability

I have shared the data at the Attach file step.

References

- Bauer, P. J. (2021). We know more than we ever learned: Processes involved in accumulation of world knowledge. *Child development perspectives*, 15(4), 220–227. <https://doi.org/10.1111/cdep.12430>
- Bauer, P. J., Dugan, J. A., Cronin-Golomb, L. M., Lee, K. A., Del Solar, B., Hanft, M., & Miller, A. G. (2024). Development of self-derivation through memory integration and relations with world knowledge. *Memory*, 1–15. <https://doi.org/10.1080/09658211.2024.2373898>
- Bauer, P. J., King, J. E., Larkina, M., Varga, N. L., & White, E. A. (2012). Characters and clues: Factors affecting children's extension of knowledge through integration of separate episodes. *Journal of Experimental Child Psychology*, 111(4), 681–694. <https://doi.org/10.1016/j.jecp.2011.10.005>
- Bauer, P. J., & Larkina, M. (2017). Realizing relevance: The influence of domain-specific information on generation of new knowledge through integration in 4- to 8-year-old children. *Child development*, 88(1), 247–262. <https://doi.org/10.1111/cdev.12584>
- Bauer, P. J., & San Souci, P. (2010). Going beyond the facts: Young children extend knowledge by integrating episodes. *Journal of experimental child psychology*, 107(4), 452–465. <https://doi.org/10.1016/j.jecp.2010.05.012>
- Begus, K., & Bonawitz, E. (2024). Infants evaluate informativeness of evidence and predict causal events as revealed in theta oscillations and predictive looking. *Commun Psychol*, 2, 77. <https://doi.org/10.1038/s44271-024-00131-3>
- Benton, D. T. (2022). The Elusive “Developmental Mechanism”: What they are and how to study and test them. *Developmental Review*, 65, Article 101034. <https://doi.org/10.1016/j.dr.2022.101034>
- Benton, D. T. (2025). An associative-learning account of how infants learn about causal action in animates and inanimates: A critical reexamination of four classic studies. *Journal of Experimental Psychology: General*, 154(2), 497–521. <https://doi.org/10.1037/xge0001656>
- Benton, D. T., Kamper, D., Beaton, R. M., & Sobel, D. M. (2024). Don't throw the associative baby out with the Bayesian bathwater: Children are more associative when reasoning retrospectively under information processing demands. *Developmental Science*, e13464. <https://doi.org/10.1111/desc.13464>
- Benton, D. T., & Rakison, D. H. (2023). Associative learning or Bayesian inference? Revisiting backwards blocking reasoning in adults. *Cognition*, 241, Article 105626. <https://doi.org/10.1016/j.cognition.2023.105626>

- Benton, D. T., Rakison, D. H., & Sobel, D. M. (2021). When correlation equals causation: A behavioral and computational account of second-order correlation learning in children. *Journal of Experimental Child Psychology*, 202, Article 105008. <https://doi.org/10.1016/j.jecp.2020.105008>
- Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, 379(6562), 255–257.
- Cohen, L. B., Chaput, H. H., & Cashion, C. H. (2002). A constructivist model of infant cognition. *Cognitive Development*, 17(3–4), 1323–1343. [https://doi.org/10.1016/S0885-2014\(02\)00124-7](https://doi.org/10.1016/S0885-2014(02)00124-7)
- Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental psychology*, 29(3), 421–433.
- Cohen, L. B., Rundell, L. J., Spellman, B. A., & Cashion, C. H. (1999). Infants' perception of causal chains. *Psychological Science*, 10(5), 412–418. <https://doi.org/10.1111/1467-9280.00178>
- Dugan, J. A., & Bauer, P. J. (2022). Putting the pieces together: Cognitive correlates of self-derivation of new knowledge in elementary school classrooms. *Journal of experimental child psychology*, 221, Article 105441. <https://doi.org/10.1016/j.jecp.2022.105441>
- Esposito, A. G., & Bauer, P. J. (2017). Going beyond the lesson: Self-generating new factual knowledge in the classroom. *Journal of experimental child psychology*, 153, 110–125. <https://doi.org/10.1016/j.jecp.2016.09.003>
- Esposito, A. G., & Bauer, P. J. (2018). Building a knowledge base: Predicting self-derivation through integration in 6-to 10-year-olds. *Journal of experimental child psychology*, 176, 55–72. <https://doi.org/10.1016/j.jecp.2018.07.011>
- Esposito, A. G., & Bauer, P. J. (2019). Self-derivation through memory integration under low surface similarity conditions: The case of multiple languages. *Journal of Experimental Child Psychology*, 187, 1–21. <https://doi.org/10.1016/j.jecp.2019.07.001>
- Esposito, A. G., & Bauer, P. J. (2024). Self-derivation through memory integration: A longitudinal examination of performance and relations with academic achievements in elementary classrooms. *Cognitive development*, 69, Article 101416. <https://doi.org/10.1016/j.cogdev.2024.101416>
- Esposito, A. G., Lee, K., Dugan, J. A., Lauer, J. E., & Bauer, P. J. (2021). Relating a picture and 1000 words: Self-derivation through integration within and across presentation formats. *Cognitive development*, 60, 1–14. <https://doi.org/10.1016/j.cogdev.2021.101099>
- Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71(5), 1205–1222. <https://doi.org/10.1111/1467-8624.00224>
- Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37(5), 620–629. <https://doi.org/10.1037/0012-1649.37.5.620>
- Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive science*, 35(8), 1407–1455. <https://doi.org/10.1111/j.1551-6709.2011.01203.x>
- Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61(3), 233–259. [https://doi.org/10.1016/S0010-0277\(96\)00715-9](https://doi.org/10.1016/S0010-0277(96)00715-9)
- Holland, P. C. (1992). *Occasion setting in Pavlovian conditioning*. In *Psychology of learning and motivation* (Vol. 28., 69–125.
- Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children prefer novelty in referent selection. *Cognition*, 118(2), 234–244. <https://doi.org/10.1016/j.cognition.2010.10.015>
- Kominsky, J. F., & Carey, S. (2024). Infants' representations of michottean triggering events. *Cognition*, 250, Article 105844. <https://doi.org/10.1016/j.cognition.2024.105844>
- Kominsky, J. F., Li, Y., & Carey, S. (2022). Infants' attributions of insides and animacy in causal interactions. *Cognitive Science*, 46(1), Article e13087. <https://doi.org/10.1111/cogs.13087>
- Kucker, S. C., McMurray, B., & Samuelson, L. K. (2018). Too much of a good thing: How novelty biases and vocabulary influence known and novel referent selection in 18-month-old children and associative learning models. *Cognitive science*, 42, 463–493. <https://doi.org/10.1111/cogs.12610>
- Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, 119(3), 573–616. <https://doi.org/10.1037/a0028681>
- Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological science*, 16(9), 678–683. <https://doi.org/10.1111/j.1467-9280.2005.01595.x>
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3), 265–288. [https://doi.org/10.1016/S0010-0277\(87\)80006-9](https://doi.org/10.1016/S0010-0277(87)80006-9)
- Mareschal, D., & Thomas, M. S. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, 11(2), 137–150. <https://doi.org/10.1109/TEVC.2006.890232>
- 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>
- McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of experimental child psychology*, 141, 1–22. <https://doi.org/10.1016/j.jecp.2015.06.017>
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3), 363–386. <https://doi.org/10.1037/0033-2909.117.3.363>
- Morey, R., & Rouder, J. (2023). BayesFactor: Computation of Bayes Factors for Common designs. *R package version*, 9.
- Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive psychology*, 61(2), 63–86. <https://doi.org/10.1016/j.cogpsych.2010.02.001>
- Muentener, P., & Schulz, L. (2014). Toddlers infer unobserved causes for spontaneous events. *Frontiers in psychology*, 5, 1496. <https://doi.org/10.3389/fpsyg.2014.01496>
- Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, 5(2), 193–207. [https://doi.org/10.1016/0885-2014\(90\)90026-P](https://doi.org/10.1016/0885-2014(90)90026-P)
- Posit team. (2025). *RStudio: Integrated Development Environment for R*. PBC, Boston, MA: Posit Software. <http://www.posit.co/>.
- Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of “nearest possible world”. *Child development*, 81(1), 376–389. <https://doi.org/10.1111/j.1467-8624.2009.01401.x>
- Rakison, D. H. (2004). Infants' sensitivity to correlations between static and dynamic features in a category context. *Journal of Experimental Child Psychology*, 89(1), 1–30. <https://doi.org/10.1016/j.jecp.2004.06.001>
- Rakison, D. H., & Benton, D. T. (2019). Second-order correlation learning of dynamic stimuli: Evidence from infants and computational modeling. *Infancy*, 24(1), 57–78. <https://doi.org/10.1111/inf.12274>
- Rakison, D. H., & Lupyan, G. (2008). Developing object concepts in infancy: An associative learning perspective. *Monographs of the Society for Research in Child Development*, 1–127.
- Rakison, D. H., Smith, G. T., & Ali, A. (2016). Who is the dynamic duo? how infants learn about the identity of objects in a causal chain. *Developmental psychology*, 52(3), 355–363.
- Rescorla, R. A., & Wagner, A. R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, Classical Conditioning II, A.H. Black & W.F. Prokasy, Eds., pp. 64–99. Appleton-Century-Crofts.
- Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536. <https://doi.org/10.1038/323533a0>
- Savic, O., Unger, L., & Sloutsky, V. M. (2023). Experience and maturation: The contribution of co-occurrence regularities in language to the development of semantic organization. *Child development*, 94(1), 142–158. <https://doi.org/10.1111/cdev.13844>
- Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology*, 43(1), 149–158. <https://doi.org/10.1037/0012-1649.43.1.149>
- Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological science*, 16(12), 995–1001. <https://doi.org/10.1111/j.1467-9280.2005.01649.x>
- Schlichting, M. L., Guarino, K. F., Roome, H. E., & Preston, A. R. (2022). Developmental differences in memory reactivation relate to encoding and inference in the human brain. *Nature Human Behaviour*, 6(3), 415–428. <https://doi.org/10.1038/s41562-021-01206-5>
- Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental psychology*, 40(2), 162–176. <https://doi.org/10.1037/0012-1649.40.2.162>

- Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental science*, 10(3), 322–332. <https://doi.org/10.1111/j.1467-7687.2007.00587.x>
- Shing, Y. L., Finke, C., Hoffmann, M., Pajkert, A., Heekeren, H. R., & Ploner, C. J. (2019). Integrating across memory episodes: Developmental trends. *PLoS One*, 14(4), Article e0215848. <https://doi.org/10.1371/journal.pone.0215848>
- Singh, S. E., Benton, D. T., & Lei, R. F. (2025). Using artificial neural networks to simulate social category learning: A tutorial. *Social Cognition*, 43(3), 152–166. <https://doi.org/10.1521/soco.2025.43.3.152>
- Sobel, D. M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *British Journal of Developmental Psychology*, 22(1), 37–58. <https://doi.org/10.1348/026151004772901104>
- Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, 42(6), 1103–1115. <https://doi.org/10.1037/0012-1649.42.6.1103>
- Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental science*, 10(3), 298–306. <https://doi.org/10.1111/j.1467-7687.2007.00589.x>
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive science*, 28(3), 303–333. https://doi.org/10.1207/s15516709cog2803_1
- Soto, F. A., Vogel, E. H., Uribe-Bahamonde, Y. E., & Perez, O. D. (2023). Why is the Rescorla-Wagner model so influential? *Neurobiology of Learning and Memory*, 204, Article 107794.
- Spelke, E., Phillips, A., & Woodward, A. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 44–78). Oxford: Clarendon Press/Oxford University Press.
- Unger, L., Chang, T., Savic, O., Bergen, B. K., & Sloutsky, V. M. (2024). When is a word in good company for learning? *Developmental Science*, 27(5), Article e13510. <https://doi.org/10.1111/desc.13510>
- Varga, N. L., & Bauer, P. J. (2013). Effects of delays on 6-year-old children's self-generation and retention of knowledge through integration. *Journal of experimental child psychology*, 115(2), 326–341. <https://doi.org/10.1016/j.jecp.2013.01.008>
- Varga, N. L., Stewart, R. A., & Bauer, P. J. (2016). Integrating across episodes: Investigating the long-term accessibility of self-derived knowledge in 4-year-old children. *Journal of experimental child psychology*, 145, 48–63. <https://doi.org/10.1016/j.jecp.2015.11.015>
- Wilson, J. T., & Bauer, P. J. (2021). Prompt-facilitated learning: The development of unprompted memory integration and subsequent self-derivation. *Memory & Cognition*, 49(7), 1473–1487. <https://doi.org/10.3758/s13421-021-01155-4>
- Yermolayeva, Y., & Rakison, D. H. (2016). Seeing the unseen: Second-order correlation learning in 7-to 11-month-olds. *Cognition*, 152, 87–100. <https://doi.org/10.1016/j.cognition.2016.03.012>
- Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development*, 54, 858–867. <https://doi.org/10.2307/1129890>
- Younger, B. A., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child development*, 57, 803–815. <https://doi.org/10.2307/1130356>
- Zhou, Z., Singh, D., Tandoc, M. C., & Schapiro, A. C. (2023). Building integrated representations through interleaved learning. *Journal of Experimental Psychology: General*, 152(9), 2666–2684. <https://doi.org/10.1037/xge0001415>