Journal of Experimental Child Psychology 141 (2016) 1–22

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

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

Children's use of interventions to learn causal structure

Teresa McCormack [a,*], Neil Bramley [b], Caren Frosch [c], Fiona Patrick [b], David Lagnado [b]

[a] School of Psychology, Queen's University Belfast, Belfast, Northern Ireland BT7 1NN, UK
[b] Division of Psychology and Language Sciences, University College London, London WC1E 6BT, UK
[c] School of Psychology, University of Leicester, Leicester LE1 9HN, UK

ARTICLE INFO

Article history:
Received 20 October 2014
Revised 29 April 2015

Keywords:
Causal learning
Bayesian modelling
Scientific learning
Causal structure
Active learning
Self directed learning

ABSTRACT

Children between 5 and 8 years of age freely intervened on a three-variable causal system, with their task being to discover whether it was a common cause structure or one of two causal chains. From 6 or 7 years of age, children were able to use information from their interventions to correctly disambiguate the structure of a causal chain. We used a Bayesian model to examine children's interventions on the system; this showed that with development children became more efficient in producing the interventions needed to disambiguate the causal structure and that the quality of interventions, as measured by their informativeness, improved developmentally. The latter measure was a significant predictor of children's correct inferences about the causal structure. A second experiment showed that levels of performance were not reduced in a task where children did not select and carry out interventions themselves, indicating no advantage for self-directed learning. However, children's performance was not related to intervention quality in these circumstances, suggesting that children learn in a different way when they carry out interventions themselves.

* Corresponding author.
E-mail address: t.mccormack@qub.ac.uk (T. McCormack).

http://dx.doi.org/10.1016/j.jecp.2015.06.017
0022-0965/

Introduction

Most developmental studies of causal learning (e.g., Bullock, Gellman, & Baillargeon, 1982; Gopnik, Sobel, Schulz, & Glymour, 2001; Shultz, 1982) have required children to judge whether an event is causally efficacious. However, in learning about the world, what is at issue is often not just whether a specific variable has a particular causal power but also the structure of the causal relations between a set of variables. You might observe that Events A, B, and C tend to co-occur (e.g., that when you feel stressed you are likely to drink more heavily and also that your blood pressure is raised). This co-occurrence is consistent with a variety of different causal structures; for example, the structure may be a common cause in which A independently causes both B and C, B ← A → C (stress causes heavier drinking and also independently causes raised blood pressure), or a causal chain in which A causes B which causes C, A → B → C (stress causes heavier drinking which raises blood pressure). How do we distinguish between these possibilities? Intervening on a causal system potentially provides very important information (Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Observing what happens if we intervened on B only would allow us to distinguish between the two suggested causal structures; if, assuming no background causes, we make B occur on its own (engage in heavy drinking when not stressed), and C does not occur (blood pressure is not elevated), we can rule out the A → B → C causal chain.

A variety of studies have required adults to infer the structure of the relations between sets of variables (e.g., Fernbach & Sloman, 2009; Kushnir, Gopnik, Lucas, & Schulz, 2010; Lagnado & Sloman, 2004, 2006; Sobel & Kushnir, 2006; Steyvers et al., 2003). In several of these studies, participants learned the causal structure by deciding what interventions to make on elements in the system, carrying out these interventions and observing their effects (Bramley, Lagnado, & Speekenbrink, 2015; Lagnado & Sloman, 2004, 2006; Sobel & Kushnir, 2006; Steyvers et al., 2003). Such interventions are assumed to reveal conditional dependencies or independencies between variables, and adults' success on these tasks has been interpreted as being consistent with the causal Bayes net approach to causal learning (e.g., Glymour, 2001; Gopnik et al., 2004) that captures causal learning in terms of the construction of causal models based on conditional probability information. A major advantage of this approach is that it specifies how and why interventions on a system yield richer information about the causal (in)dependencies between variables than that which is available through observation of patterns of covariation (Hagmayer et al., 2007; Steyvers et al., 2003; Waldmann & Hagmayer, 2005).

The causal Bayes net approach has been extensively adopted by developmental psychologists interested in explaining children's learning about causation (Gopnik, 2012; Gopnik & Wellman, 2012; Gopnik et al., 2004). The majority of studies in this tradition have involved children learning whether an object possesses a particular causal power, usually on the basis of observing the experimenter's actions (e.g., whether an object makes a box light up and play a tune; Gopnik & Sobel, 2000; Kushnir & Gopnik, 2007; Sobel, Tenenbaum, & Gopnik, 2004). Relatively few studies have used tasks in which children themselves decide which interventions to carry out in order to discover the causal structure of a system (e.g., whether it has a causal chain or common cause structure). Such studies are particularly important because they can be used to assess young children's effectiveness in generating and testing hypotheses about the causal relations between sets of variables. Moreover, a key advantage of the causal Bayes net approach over most other accounts of causal learning is that it can capture this more complex type of learning, distinguishing between different causal paths as well as identifying variables' ultimate effects.

One study that did examine children's ability to learn causal structure by means of making interventions on a system is that of Schulz, Gopnik, and Glymour (2007, Experiment 3), in which 4- and 5-year-olds intervened on a causal system involving a box with two gears. Children needed to decide whether each gear moved by itself or whether one of the gears caused the other one to move. Children could remove each gear in turn from the box to examine whether the other gear worked on its own when the box itself was switched on. They gave their answers about the relations between the gears by selecting from a set of anthropomorphized pictures of the two gears that depicted different possible relations between them. Performance on this task was mixed. Children did not all reliably generate the

right interventions to distinguish between the different possible relations that might hold between the gears. Even among those who did make appropriate interventions, children were not successful at identifying instances in which one of the other gears caused the other one to move.

Schulz and colleagues' (2007) study provides limited support for the claim that children will generate informative interventions and use this information to distinguish between different causal structures. Not only was performance relatively weak, but children were only required to make judgments about the dependencies between pairs of variables rather than to distinguish among three-variable causal structures. Children gave their responses by pointing to pictures of the two gears that showed whether they turned themselves or one turned the other. There was a third variable—a switch—that was important in the system, but children did not need to represent how its relation to the gears varied for the different causal systems, and it did not feature in the pictures depicting causal relations between the gears. Thus, this study does not allow us to draw firm conclusions about whether children can use interventions to distinguish between, for example, common cause and causal chain structures.

However, the findings of some other studies suggest that we should expect even very young children to be good at crafting appropriate interventions and using them to learn about causal systems (Bonawitz, van Schijndel, Friel, & Schulz, 2012; Cook, Goodman, & Schulz, 2011; Schulz & Bonawitz, 2007). Indeed, Schulz (2012) argued that the ability to select appropriate interventions and use the evidence generated from such interventions may be developmentally basic. In various studies, she and her colleagues showed that young children will appropriately explore a causal scenario when given ambiguous information (Cook et al., 2011; Schulz & Bonawitz, 2007). In these scenarios, children's behavior did suggest that children were trying to figure out whether an object possessed a certain causal property. However, children did not need to make interventions to disambiguate the structure of the relations between different variables and then use this information to decide, for example, whether a system was a common cause or causal chain.

A further study by Sobel and Sommerville (2010) tried to address this specific issue. Children viewed a box with four colored lights—A, B, C, and D—and were told that some of the lights could make other lights turn on. The box was configured so that the relations between the lights took the form of either a common cause, $B \leftarrow A \rightarrow C$, or a causal chain structure, $A \rightarrow B \rightarrow C$. Children could interact freely with the box by switching on lights and observing their effects. They were then asked a series of questions about the relation between pairs of lights. Sobel and Sommerville found that children performed above chance on these questions, which could be interpreted as indicating that they were able to use the information generated from their interventions to decide on the structure of the causal relations. There are, however, two difficulties with this interpretation. First, before children answered the questions, the experimenter pressed each of the buttons in turn and narrated what it did; arguably, this provided children with the answers to the test questions (indeed, children performed above chance, although less accurately, when given just this narration). Second, it was not clear that to answer correctly children needed to have an integrated representation of how the three variables in the system were related to each other rather than just knowledge of pairwise relations. Indeed, Sobel and Sommerville did not include in their analyses the answers that children gave to the question of whether A makes C go in the case of the causal chain, arguing that answers to this question are hard to interpret. However, by questioning children only about the other pairwise relations between A and B and between B and C, it is impossible to know whether children actually understood the nature of the overall causal structure.

The general point here is that we can distinguish between learning structurally local pairwise links and integrating such links to form a representation of causal structure. This distinction is important because learning localized pairwise relations is likely to be easier than learning global structure (Fernbach & Sloman, 2009). Moreover, learning local pairs only is often liable to lead to the wrong global model; for example, when one connection "explains away" the dependence between two others, a pairwise learning strategy would still attribute a connection between these two variables, whereas a global strategy would not. A number of other studies of children's causal learning can also be interpreted as studying children's learning of pairwise relations rather than global causal structure (e.g., Schulz, Goodman, Tenenbaum, & Jenkins, 2008; Sobel & Sommerville, 2009), meaning that we still have limited evidence about children's ability to learn causal structure.

Uncertainty as to whether children are adept at appropriately generating interventions and using them to learn causal structure comes from two sources. First, research on children's scientific learning has for many years suggested that younger children may have great difficulty in generating appropriately informative interventions and learning the nature of relations between variables from the evidence generated by these interventions (e.g., Klahr & Dunbar, 1988; Klahr, Fay, & Dunbar, 1993; Kuhn, 1989; Schauble, 1996; Zimmerman, 2000, 2007). On the face of it, this body of findings seems at odds with recent findings from the causal Bayes net tradition. One possible explanation of the differing findings lies in the role of the knowledge base in scientific learning studies. For example, preexisting, and sometimes erroneous, beliefs can hamper children's ability to generate appropriate interventions and interpret statistical data (e.g., Amsel & Brock, 1996; Kuhn et al., 1988). Indeed, Schulz and colleagues suggested that this type of factor, along with task complexity, may mask children's basic learning skills (Bonawitz, van Schijndel, Friel, & Schulz, 2012; Cook et al., 2011), which may be better demonstrated in the tasks used in the causal Bayes net tradition where domain-specific knowledge is of limited importance and statistical evidence is simple.

However, the findings of a recent study by McCormack, Frosch, Patrick, and Lagnado (2015) provide a second reason for being unsure about children's ability to learn from interventions on a causal system. This study, like most of those in the causal Bayes net tradition, involved children learning about a novel mechanical system. The only relevant data for causal learning were supplied by two types of domain-general cues: statistical information provided through interventions on the system and the temporal patterns of event occurrence. Children needed to learn the causal structure of the system—a box with three separate shapes (A, B, and C) on its surface that rotated. Across two experiments, children watched while the experimenter intervened on components of the system. In one experiment, the experimenter carried out interventions in which she disabled one of the shapes by preventing it from moving before moving each of the other shapes in turn. Children did not find it straightforward to use the patterns of evidence provided by these interventions to discriminate between causal structures even when the system operated deterministically. Although 6- and 7-year-olds were able to use the evidence from the more complex interventions to accurately infer when the system was one of the causal chains, children younger than this could not do so, and even 7- and 8-year-olds were unable to use information from these interventions to accurately judge when the system was a common cause. McCormack and colleagues argued that children's difficulties may stem from integrating pieces of evidence provided across a number of separate observations of the causal system.

At first sight, McCormack and colleagues' (2015) findings seem to be more consistent with the conclusions stemming from research on children's scientific learning that has emphasized its limitations. It might be argued, however, that this study did not provide children with an optimal opportunity to demonstrate their abilities. Children watched while the experimenter made a series of interventions rather than making the interventions themselves. Sobel and Kushnir (2006) argued that participants find it easier to learn causal structure when they decide what interventions to conduct, largely because this provides an opportunity for them to engage in more active hypothesis testing (but see Lagnado & Sloman, 2004). In particular, they suggested that when participants craft interventions themselves, they obtain evidence in a structured way that makes it more apparent whether it supports specific hypotheses. Moreover, children might be particularly likely to benefit from being allowed to explore how a system operates in that hands-on interventions may ensure they stay engaged with the task.

In this study, we used a task very similar to that of McCormack and colleagues (2015) in which children needed to decide whether a three-element causal system was a common cause, B ← A → C, an A → B → C causal chain, or an A → C → B causal chain. Children intervened on the system themselves in order to learn its structure. Shapes on top of a box rotated when children moved them by hand, or shapes could be moved by rotating another shape that was causally connected to them. For example, for the A → B → C causal chain, spinning A initiated the rotation of both B and C, and spinning B rotated C; all of the shapes always moved simultaneously in the tasks to minimize temporal cues. Children needed to select and carry out a series of interventions; these could be simple interventions in which they made one of the three shapes spin, or they could be more complex interventions in which children prevented one of the three shapes from moving by disabling it and then spun one

of the other two shapes. Note that we were not attempting to faithfully recreate a free-play situation because it was important for our analyses that we were able to exhaustively categorize children's actions on the system. Although they were completely free to choose their interventions, the only actions children could carry out were interventions on the system. Furthermore, it was made clear to children that their job was to learn the causal structure of the system and that they could not make an unlimited number of interventions. This allowed us to look in our modeling work at the efficiency with which children produced informative interventions.

We examined two aspects of performance: the nature of the interventions that children selected and children's causal structure choices. Not all interventions provided useful information to discriminate among the three possible causal structures, which allowed us to examine whether the tendency to choose informative interventions changes with age. We also examined whether there was any relation between the quality of children's interventions and the likelihood that children chose the correct causal structure at test. It is possible to try to examine these issues without formal modeling (see Sobel & Kushnir, 2006) by, for example, simply distinguishing between two broad classes of informative and non-informative interventions. However, we chose to model children's learning in a Bayesian framework. Doing so has two key advantages. First, it allows us to properly assess whether there are developmental changes in the extent to which children resemble idealized Bayesian learners. This is important because, within the currently dominant causal Bayes net tradition, young children's learning is often characterized as approximating to such an ideal, particularly with regard to causal learning from statistical information (e.g., Gopnik, 2012; Gopnik & Wellman, 2012). Formal modeling allows us to assess the extent to which this characterization is appropriate by assessing children's performance against the standards set by the Bayesian tradition itself.

Second, although in this study we can (and do) classify interventions broadly as informative or non-informative, the learning task itself is sequential. This means that how informative an intervention is depends on what children have already observed and what they can remember about such observations. However, figuring out the informativeness of each intervention that a participant makes on a trial-by-trial basis would be a formidable task without a formal model. Indeed, without such a model, it is hard to see how one would operationalize the notion of informativeness under such circumstances. Our Bayesian model allowed us to capture the sequential nature of the learning task by assuming that the most informative interventions were those that maximally reduced uncertainty about which was the correct hypothesis at any particular point in the learning sequence given some level of forgetting.

Experiment 1

Method

Participants

Children were from three different school years: 21 5- and 6-year-olds (M = 72 months, range = 64–80), 31 6- and 7-year-olds (M = 86 months, range = 80–93) and 25 7- and 8-year-olds (M = 98 months, range = 93–103). Children were tested individually in their schools.

Materials

The study used a wooden box, 41 cm (long) × 32 cm (wide) × 20 cm (high), which had an on/off switch at the front. There were three different colored lids for the box. Two of these had three colored/patterned shapes (e.g., circle, rectangle, star) inserted on their surface that rotated independently on the horizontal plane; a separate lid was used in pretraining and had only two shapes (see Fig. 1). The colors and shapes of the components were varied across participants and causal structures. On each of the two lids used in testing, the three shapes formed an equilateral triangle of 24-cm sides. Each shape had a small hole that aligned with a hole in the lid of the box. There was a miniature red-and-white "Stop" sign affixed to a metal rod that could be inserted through the hole on any shape into the corresponding hole in the box, preventing it from moving. Each of the shapes could be rotated by hand; the rotation of the other shapes was controlled by a laptop hidden inside the apparatus that

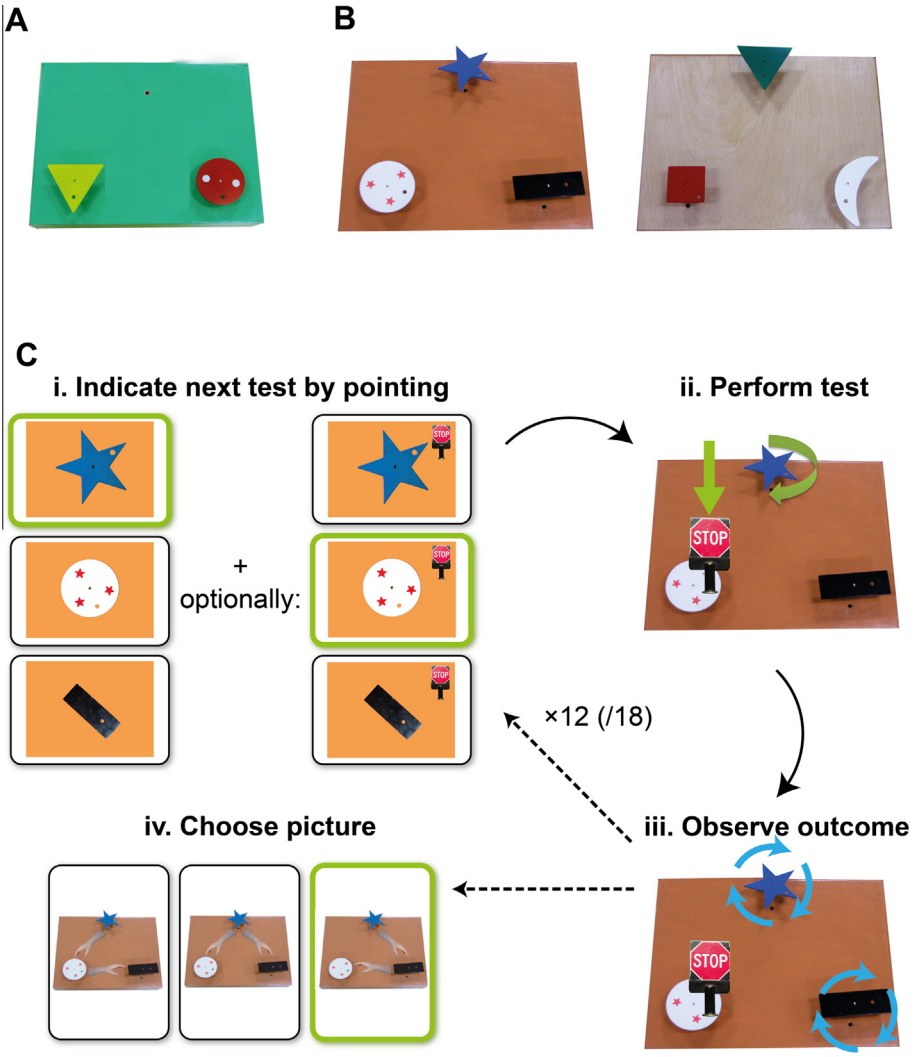

Fig. 1. (A,B) Apparatus: box lid used in training (A) and two box lids used during testing (counterbalanced between common cause and chain trials) (B). (C) Procedure: (i) participants indicate (optionally) which shape to block with the stop sign and which shape to spin; (ii) participants perform the action(s) they chose; (iii) participants observe which shapes spin as a result of their test; (iv) after 12 (or 18) tests, participants point to the card showing how they think the machine works. Green arrows and highlighting show participants' actions on an example trial. (For interpretation of the reference to color in this figure legend, the reader is referred to the web version of this article.)

participants were unaware of. A set of photographs was used during the learning phase that participants used to indicate which intervention they were going to make; these photographs depicted each shape on the box, and in addition there were photographs of each of the shapes alongside the stop sign. Photographs of the whole box with its shapes depicting three possible causal structures were used at test for children to indicate their judgment of the causal structure: one common cause and two causal chains (i.e., depicting B ← A → C, A → B → C, or A → C → B). The photographs for use at test were overlaid with pictures of hands to indicate causal links (following Frosch, McCormack, Lagnado, & Burns, 2012).

Procedure

Children completed two test trials: one common cause and one causal chain (order counterbalanced). There was a pretraining phase to ensure that children knew what their task was and how to give their answer at test. The pretraining procedure used a lid on the box that had only two colored shapes inserted on its surface; its purpose was to demonstrate that some shapes caused others to move but that the stop sign could be used to prevent a shape from moving. Children were initially asked to name the colors of the shapes to ensure that they would know which shapes the experimenter was referring to, and the experimenter drew children's attention to the on/off switch at the front set at the "off" position. She then switched the box on and manually rotated one of the two shapes (X). This had no effect on the other shape (Y), which remained stationary, and the experimenter pointed this out to children. She then rotated the other shape (Y), which resulted in the first shape (X) simultaneously rotating. She explained to children, "Some shapes are made to move by others." The experimenter then switched the box off and introduced children to the stop sign, which she inserted into X to stop it from moving, saying, "See this stop sign, it can be used to stop a shape from moving. See the [color of X] one cannot move now." She then switched the box on again and rotated Y, which this time had no impact on the movement of X because it was prevented from moving by the stop sign. Following this, the lid was removed from the box and replaced by a different colored lid with three different shapes for the first test trial.

Children were asked to name the colors of the three shapes and were told that their job was to figure out how the box worked. They were introduced to the three test pictures depicting the three different causal structures, with the experimenter saying, "In a moment I will ask you to figure out how the box works, but first I want to show you some pictures of the box which show different ways in which the box may be working. Only one of them is right, and you've got to work out which is the right one. It won't change halfway through, and it is definitely only one of the pictures. You'll have to use your detective skills to work out which picture shows what the box does."

The experimenter described each of the three pictures (e.g., "In this picture, the red one makes the blue one go, and the blue one makes the white one go, and the hands show that"). Following these three descriptions, children were then asked a set of three comprehension questions. For each causal chain picture, the experimenter asked, "Can you show me the picture where the [color of A] one makes the [color of B/C] one go and the [color of B/C] one makes the [color of B/C] one go?" For the common cause picture, the experimenter asked, "Can you show me the picture where the [color of A] one makes *both* the [color of B] one and the [color of C] one go?" The majority of children answered these questions correctly the first time, but if they did not answer all three questions correctly, the experimenter repeated the initial descriptions and asked the comprehension questions again. This procedure was repeated again if necessary.

Following this pretraining, the experimenter said, "I am going to switch the box on now, and I want you to figure out how the box works." Children were told that they could do one of two things (order counterbalanced): either "You can move a shape to see if it makes other shapes move" or "You can stop a shape from moving by putting the stop sign in and then see what happens when you move another shape." It was explained to children that before they carried out each intervention, they needed to point to a card indicating what they intended to do. The experimenter said, "Before you try anything on the box, I want you to point to one of these cards. This card means you want to *spin* the [color] one, and you point to this card if you want to *stop* the [color] one. See, we also have the cards for spinning and stopping the [color and color] ones. So, each time you want to do something, you point to one of these cards first."

Children were told that they had 12 "goes" to start with and that each time they moved a shape counted as 1 "go." It was made clear that using the stop sign did not count as a go by itself; children needed to then in addition move one of the other shapes. The procedure with cards was used to ensure that children interacted with the box in a controlled way and to make clear that they could not make an unlimited number of interventions. It also ensured that all children made a fixed minimum number of interactions before attempting to answer the test question. Children were told that they did not need to keep track of the number of goes that they had with the box because the experimenter would count this for them.

Before children began, the experimenter said, "Remember, you've got to figure out which picture shows how this box really works." She then demonstrated what happened when the A shape was moved, which was that the other two shapes also moved simultaneously, and pointed out that they did not know yet "which ones make other ones go." Participants were subsequently allowed to make interventions on the box by first selecting the appropriate card and then making the intervention. So, for example, if they wanted to see what happened when C was moved if B was disabled, they needed to point to the card depicting B with the stop sign in it and then to the card depicting C. They then carried out their intervention.

After participants completed 12 interventions, the experimenter said, "You have had your 12 goes now. Do you want to choose which picture you think shows what the box did, or do you want to have another 6 goes?" The majority of participants opted to choose after 12 interventions. Children completed a short filler task (a paper-and-pencil maze) in between the common cause/causal chain trials. It was made clear that the second box might work in the same way as the first box or it might work in a different way. The second box always had a lid of a different color and different shapes.

Results

In both trials, 69 of the 77 participants stopped after 12 interventions. The remaining 8 opted for an additional 6 interventions in one or the other trial. Of these, 4 participants opted for the additional 6 interventions on both trial types. Initial data analyses examined participants' responses for each of the two trial types. Fig. 2 shows the percentage of participants who chose each response type for each trial type. The majority of participants in each group, except for the youngest group, chose the correct answer for the causal chain trial. The majority of participants in all groups chose the common cause response for the common cause trial. Binomial analysis showed that each group of participants chose the correct response more often than chance (all $ps < .01$) except for the 5- and 6-year-olds who did not select the correct causal chain more often than chance. This group tended to select the common cause response for both structures. Performance on the causal chain structure was associated with age, $\chi^2(2) = 6.91$, $p < .05$, with the number of correct responses improving with age. Performance on the common cause structure was marginally significantly associated with age, $\chi^2(2) = 5.66$, $p = .056$, although in this case the 6- and 7-year-olds gave more correct responses than each of the other groups.

Analysis of interventions

Subsequent analyses examined the nature of participants' interventions on the system. We initially discriminated between whether an intervention was informative or not given the three possible causal structures. There were three interventions that were never informative: A+, B+C–, and B–C+, adopting the notation of "+" to mean that a particular shape was moved by the participant and "–" to mean that a shape was disabled. Potentially informative interventions were A+B–, A+C–, B+, C+, A–B+, and A–C+. We also classified interventions as simple or complex; A+, B+, and C+ were classified as simple, and those involving initially disabling one of the components before moving another component were classified as complex. Table 1 shows the percentage of times that participants in each age group chose each of these interventions. The most popular intervention tended to be A+, which, although it was uninformative, did make all of the three shapes spin. Propensity to select a complex intervention increased significantly with age, $F(2, 74) = 7.22$, $p < .002$, $\eta^2 = .16$, with 7- and 8-year-olds being the most likely to pick the complex interventions (65% of the time vs. 46% for 5- and 6-year-olds and 45% for 6- and 7-year-olds). We examined whether participants chose informative interventions more often than chance by conducting a one-sample t-test with a test value of .67 given that two thirds of the nine possible interventions were informative. Only the 7- and 8-year-olds were significantly more likely than chance to select informative interventions, $t(24) = 2.83$, $p < .01$, both $ps > .10$ for the younger groups. A logistic regression showed that the proportion of informative interventions significantly predicted the probability of a participant getting the chain trial correct ($z = 2.73$, $p < .01$) (see Table 2), but this was not the case for the common cause trial ($z = -0.62$, $p > .50$). One potential explanation for the latter finding is that the children were overall more likely to select the common cause, doing so 56% of the time. Thus, some of the correct responses on the common cause test trial are likely to have been made by the weaker participants purely by virtue of their favoring the common cause structure.

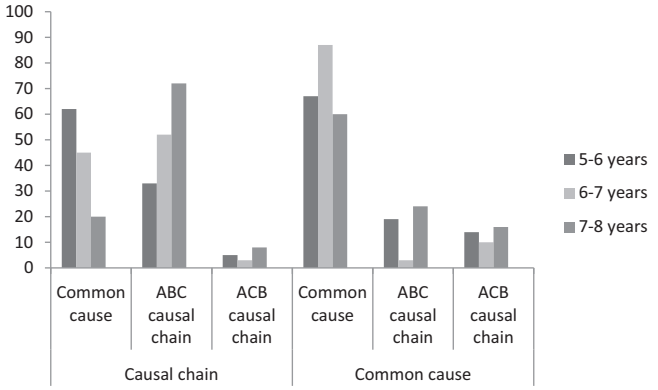

Fig. 2. Percentage of response choices for each causal structure as a function of age group. Correct responses to the causal chain are denoted as ABC.

Table 1
Percentage of times that participants chose each intervention collapsed across common cause and causal chain trials.

	Informative						Uninformative		
	B+	A–B+	C+	A–C+	A+C–	A+B–	A+	B–C+	B+C–
5–6 years	17.9	9.5	12.9	5.1	9.6	8.3	23.3	8.2	5.4
6–7 years	17.5	9.2	14.2	8.2	8.2	9.0	21.8	6.4	5.6
7–8 years	10.8	11.3	9.4	11.1	14.8	13.5	14.2	6.9	8.0

Table 2
Regression analyses from Experiment 1.

Dependent	Parameter	Estimate	*SE*	Odds ratio	*z*	$p(>\|z\|)$
$P_{correct}$(chain)	Intercept	–4.77	1.80		–2.65	.008**
	% Informative	7.37	2.70	1587	2.73	.006**
$P_{correct}$(common cause)	Intercept	1.87	1.48		1.27	.21
	% Informative	–1.32	2.15	0.267	–0.62	.54
$P_{correct}$(chain)	Intercept	–0.81	0.75		–1.09	.278
	Efficiency	1.22	0.92	3.39	1.32	.186
$P_{correct}$(common cause)	Intercept	3.39	0.99		3.39	.0007***
	Efficiency	–3.31	1.27	0.037	–2.62	.009**
$P_{correct}$(chain)	Intercept	–4.43	1.74		2.53	.012*
	Quality	7.98	3.06	2921	2.61	.009**
$P_{correct}$(common cause)	Intercept	1.46	1.40		1.04	.30
	Quality	–0.78	2.24	0.46	–0.35	.73

Note: The table shows three separate analyses of predictors of performance on the causal chain and common cause trials: percentage informative interventions, efficiency, and intervention quality.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Modeling interventions

So far, we have looked at proportion of informative intervention choices without considering the sequential nature of the task or whether and how efficiently children produced a set of informative interventions sufficient to discriminate between causal structures. A child who did not produce such a set but repeatedly produced a single informative intervention would score 100% on this measure. Moreover, how useful an intervention is depends on what learners already know (in this case what they have already learned from their previous interventions). For example, A+B– and C+ are both informative interventions in this task provided that you do not know anything yet. But suppose that you have already performed A+C– and observed that this made B spin. This evidence effectively rules out the ACB chain, leaving only the ABC chain and the common cause as possibilities. Now, on subsequent trials, performing C+, or repeating A+C–, will not tell you anything new because both of these interventions simply distinguish the ACB chain from the other two. To capture how efficiently children's intervention choices allow them to home in on the true structure, we can analyze the interventions sequentially by looking at how effectively these interventions reduce uncertainty, assuming that initially children are perfectly able to remember past outcomes and integrate new information.

To do this, we defined participants' subjective uncertainty about the true structure at a given time point as the *information entropy* $H(S)$ (Shannon, 1948) of their posterior distribution over the three possible structures $s \in S$ given the data they had seen so far[1] (see Eq. (1)):

$$H(S) = -\sum_{s \in S} P(s) \log_2 P(s). \tag{1}$$

Every time children observed new evidence, this distribution was updated using Bayes rule and the likelihoods $P(o|s,i)$ for observing that outcome $o \in O$ given the different structures g and the intervention $i \in I$ giving posterior probabilities $P(s|o,i)$ (see Eq. (2)):

$$P(S|i,o) \propto P(o|i,S)P(S). \tag{2}$$

Because the box worked in a deterministic way, the likelihood was always 0 (if the outcome was impossible given that structure and intervention) or 1 (if the outcome was to be expected given that structure and intervention). If an outcome had a zero likelihood under one structure, that structure's posterior probability would go to zero once participants saw that outcome. By doing this, we were able to compute the *expected reduction in information entropy* $E[H(S|i)]$ for each intervention chosen participants (see Eq. (3)),

$$E[H(S|i)] = \sum_{o \in O} H(S|i,o)P(o|S,i), \tag{3}$$

and to rescale this by the maximum achievable expected increase in information over the different interventions at that time point. This gave a measure of the overall *efficiency* of the intervention choices made by each child for facilitating their identification of the true structure (see Eq. (4)):

$$\text{Efficiency}_i = \frac{E[H(S|i)]}{\underset{r \in I}{max} E[H(S|r)]}. \tag{4}$$

Because of the deterministic nature of the task, in fact all of the children generated enough information with their interventions for their uncertainty to go to zero before the end of the trial, so intervention efficiency was simply calculated for the interventions up until the point that their posterior

[1] Shannon entropy is not the only possible criterion value for modeling active learning. Others include probability gain (Baron, 2005), impact (Wells & Lindsay, 1980), and diagnosticity (Good, 1950). Although in most situations the measures make very similar predictions, Nelson (2005) found that when they differ Shannon entropy is at least as good as any other, both as a normative criterion and as a descriptive account of human active query selection. In the domain of causal learning, Bramley, Lagnado, et al. (2015) found that adults are better described as choosing interventions that minimize expected Shannon entropy rather than maximize expected probability gain or utility and that sequentially minimizing Shannon entropy provides better long-run accuracy than these other two measures. Finally, Bramley, Nelson, Speekenbrink, Crupi, and Lagnado (2014) showed that Shannon entropy performs at least as well as any of a broader family of generalized entropy functions in the domain of active causal learning. For these reasons, we report modeling using Shannon entropy. Nevertheless, we also repeated these analyses using probability gain, finding qualitatively the same results.

uncertainty reached zero. Fig. 3 shows an example of how the model worked using a real set of interventions; it also depicts how these interventions were categorized. We established chance level interventional efficiency by simulating the task 1000 times with randomly selected interventions, finding average chance efficiency levels of .48 for the chain structure and .43 for the common cause.[2]

For all age groups, for both structures, children's interventions were significantly much more efficient than the chance level (mean efficiencies for chain and common cause, respectively: 5- and 6-year-olds = .71 and .66, 6- and 7-year-olds = .79 and .63, and 7- and 8-year-olds = .82 and .80; all ts > 6, ps < .0001). Children's efficiency for the common cause changed significantly with age, $F(2,74) = 4.47$, $p \leqslant .02$, $\eta^2 = .11$, but there was no effect of age on efficiency on the causal chain trials, $F(2,74) = 1.14$, $p = .32$, $\eta^2 = .03$. Unlike *proportion informative interventions,* efficiency did not predict accuracy on the chain (see Table 2) and was in fact negatively related to accuracy on the common cause ($z = -2.62$, $p < .01$).

Whereas proportion informative interventions did not take into account the sequential nature of the task, arguably interventional *efficiency* has the opposite shortcoming. By assuming, implausibly, that children have a perfect memory for the outcomes of their previous interventions and perfect ability to make inferences from this information, it ignores what they do on subsequent interventions once they have, in principle, enough information to potentially identify the correct structure. An inspection of the modeled data found that children obtained sufficient information for certainty after an average of only 2.75 interventions; this means that our measure of efficiency ignores a large proportion of the data and makes no allowances for noise, forgetting, or uncertainty in learning. A more balanced way to assess the quality of participants' interventions is achieved by adding some noise, encapsulating the idea that learning is likely to be somewhat leaky or error prone.

We augmented our Bayesian learning model so that, after each test, some proportion of what was learned previously was "forgotten."[3] This was achieved by mixing a uniform distribution in with the posteriors, with the proportion determined by a forgetting rate γ (see Eq. (5)),

$$P(S|i,o) \propto P(o|i,S)\left[(1-\gamma)P(S) + \gamma\frac{1}{3}\right], \tag{5}$$

and using this as the prior for next intervention. This procedure was carried out for each learner's 12 (or 18) tests. This means that previously ruled out alternatives gradually regained some probability mass, whereas more likely options became a little less favored. The quality of each intervention was then calculated based on the extent that it reduced uncertainty across these distributions compared with an intervention that would have maximally reduced uncertainty. This method captures the idea that continually repeating a particular intervention is less useful than selecting a complementary mixture of different interventions while also allowing that real-world learners are likely to forget, ignore, or make mistakes about the evidence they have seen previously, meaning that revisiting previous interventions is not useless.

The exact level of "forgetting" in the model turned out not to be particularly important. We found qualitatively the same results setting it to 10%, 25%, 50%, 75%, and 90%, although the results were clearer for the lower levels of forgetting. Here we report results assuming 25% forgetting after each test. We established chance levels of intervention quality, again through simulation over 1000 trials. The 5- and 6-year-olds' intervention quality was not significantly above chance on either the chain or common cause; the 6- and 7-year-olds were above chance on the chain, $t(30) = 2.88$, $p < .01$, and marginal on the common cause, $t(30) = 1.73$, $p = .09$; and the 7- and 8-year-olds' intervention quality was above chance for both trials (chain: $t(24) = 2.46$, $p < .02$; common cause: $t(24) = 4.87$, $p < .001$). Averaged over trial types, we found that intervention quality improved with age, $F(2,74) = 4.03$,

[2] This corresponds to getting enough information to identify the true structure after an average of 3 random interventions when the true structure is a chain and 4.5 random tests when the true structure is the common cause. The chain is somewhat quicker to be identifiable by chance because sometimes it can be identified from a single intervention (e.g., B+ allows identification of the ABC chain), whereas the common cause always requires a minimum of two interventions.

[3] There are numerous ways in which to model forgetting (e.g., Wixted, 2004), but a reasonable high-level approach is to assume that children forget random aspects of their priors, leading to a net "flattening" of their subjective priors going into each new intervention. We remain agnostic about whether this parameter captures cognitive forgetting or more generalized sources of error and uncertainty in children's integration of information.

Test	Intervention	Outcome	Informative	Prior	Efficiency	Prior (25%)	Quality
1	B+ A-	C+	✓		1		1
2	B+ A-	C+	✓		0		.42
3	A+ C-	B+	✓		1		1
4	B+		✓		/		1
5	C+ A-		✓		/		1
6	C+ A-		✓		/		.66
7	A+ B-	C+	✓		/		1
8	C+		✓		/		1
9	A+ C-	B+	✓		/		1
10	A+	B+ C+	✗		/		0
11	A+	B+ C+	✗		/		0
12	B+ C-		✗		/		0
	Proportion *complex*			Proportion *informative*		Mean *efficiency*	Mean *quality* (25%)
	.66			.75		.66	.64

Fig. 3. Interventions selected by a 6- or 7-year-old in the common cause trial. From left to right, columns show test order, selected intervention, which shapes (if any) spun as a result, whether the intervention was generally *informative*, a learner's prior given perfect memory and integration of previous tests, the corresponding *efficiency* of the intervention in allowing identification of the common cause, a learner's prior given 25% forgetting, and the corresponding *quality* of each intervention.

$p < .03$, $\eta^2 = .10$, with 7- and 8-year-olds significantly more efficient that 5- and 6-year-olds ($p < .01$), but no significant difference between 5- and 6-year-olds and 6- and 7-year-olds. Breaking this into responses for the two structures, regardless of forgetting rate, intervention quality was a significant predictor for correct identification of the chain structure ($z = 2.61$, $p < .01$), but not for the common cause structure (see Table 2).

Discussion

Our findings provide important information about developmental changes in children's ability to learn causal structure through intervention. Children's ability to learn a causal chain structure improved developmentally, with the youngest children not managing to learn this structure at above-chance levels. However, we need to consider why the 5- and 6-year-olds correctly identified the common cause structure as accurately as the 7- and 8-year-olds. Our view is that the good performance on this second trial type is due to a tendency even among the youngest children to assume that, when events happen simultaneously, the underlying structure is a common cause. Previous studies have found that both children and adults make use of this simple temporal heuristic when they observe a three-variable system with this sort of temporal schedule (Burns & McCormack, 2009; Fernbach & Sloman, 2009; Lagnado & Sloman, 2006; McCormack et al., 2015). Indeed, McCormack et al. (2015) demonstrated that children will use this type of temporal heuristic even when faced with contradictory statistical information provided either through observing the operation of a probabilistic causal system or through observing the effects of interventions on a deterministic system. Thus, the good performance of the younger children on the common cause structure is likely to reflect use of this heuristic rather than use of statistical information derived from interventions on the system. This would also straightforwardly explain the lack of a relation between intervention quality and performance on the common cause structure.

The analyses of children's intervention choices provide insights into why performance improved developmentally on the causal chain trial. Interventions could be initially classified as informative or non-informative given the three possible causal structures. Over all trials, unlike the oldest group, younger groups of children did not choose informative interventions more often than chance. It proved to be fruitful, however, to further examine intervention choices and how these related to performance in more detail by means of our modeling work. The initial analysis of how efficient participants were at producing a set of interventions that could in principle discriminate between the different hypotheses showed that all groups of children produced such a set more quickly than would be expected if they were simply choosing between interventions at random. This means that even the youngest children had the evidence available to them to make the appropriate causal inferences. However, intervention efficiency was not a predictor of performance. This demonstrates that good performance does not hinge on simply initially choosing interventions that are as a matter of fact disambiguating. Children may forget or fail to make use of what they have observed, and the subsequent interventions they make may also influence their judgments.

Our modelling work suggested that this was indeed the case because our measure of the quality of children's interventions that took into account the complete sequence of interventions predicted performance on the causal chain structure, under the assumption that there was some degree of forgetting. Moreover, unlike efficiency, intervention quality improved with age, with older children being more likely to consistently choose interventions that would help to disambiguate the causal structures given what they had already observed. These results indicate that with development children become more discerning in their choice of interventions, and this has an impact on their causal structure learning.

How do our findings fit with what is already known about developmental changes in children's use of interventions to learn about causal systems? In designing our study, we sought to ensure that domain-specific knowledge was not relevant for task performance. However, this did not rule out children exploiting a type of preexisting, albeit domain-general, heuristic about the nature of the causal system, namely that when multiple events occur immediately following an intervention, the underlying structure is likely to be a common cause (McCormack et al., 2015). When the evidence generated from interventions was consistent with this assumption (i.e., in the common cause trial), even young children performed well. However, younger children had difficulty in discarding this assumption on the basis of the contradictory evidence provided by their interventions. This is consistent with evidence from the scientific learning literature indicating that children have difficulty in discarding a preexisting hypothesis and may routinely ignore statistical evidence that fails to support such a hypothesis (Amsel & Brock, 1996; Kuhn et al., 1988).

Furthermore, an inspection of developmental changes in the pattern of children's intervention choices (Table 1) yields some further interesting additional parallels with findings from the scientific learning literature. Our task is very different from those used in research on children's scientific learning; it is simpler, and the children we tested are younger than those typically used in such studies (but see Koerber, Sodian, Thoermer, & Nett, 2005; Piekny, Grube, & Maehler, 2014). Nevertheless, some of our findings confirm broad developmental patterns that are well established in that research. Younger children tended to prefer making the A+ intervention and did so repeatedly. This intervention is the most causally effective (it makes all of the events happen) but does not discriminate among the three available hypotheses. However, it reinforces any existing hypothesis that the causal structure is a common cause by providing the temporal pattern of all events happening simultaneously. Young children's preference for this intervention has parallels with demonstrations in the scientific learning studies showing that children attend most to the variable already believed to be causal, focus more on producing an effect than on generating disambiguating evidence, and produce evidence that is consistent with their existing hypothesis rather than seeking to disconfirm it (e.g., Klahr & Dunbar, 1988; Klahr et al., 1993; Kuhn, 1989; Schauble, 1990, 1996).

Although younger children's patterns of interventions led to poorer performance, it is interesting to note that recent formal analyses have demonstrated that whether their type of approach should be viewed as inefficient depends on the learning context. First, the tendency to intervene on variables already believed to be causal in order to confirm an existing hypothesis is not necessarily always the wrong strategy. This type of strategy has been shown to be rational under the assumption that causal connections in the world are sparse (Navarro & Perfors, 2011), meaning that competing causal hypotheses do not generally share the same effect variables. In such circumstances, "positive tests," operationalized as intervening on the variable thought to be the root cause (Coenen, Rehder, & Gureckis, 2015), are highly diagnostic. Hence, younger children's pattern of interventions could be interpreted as due to a tendency to act in a way that has proved to be an effective general-purpose method for learning causal relationships in the past but is not appropriate given the specific learning context in which they find themselves.

Second, we also found that younger children were less likely than older children to produce the more complex interventions that involved disabling one of the components in the system. This type of intervention can be particularly informative because it can be used to exclude a variable as being necessary for production of an effect. However, separate Bayesian modeling work with adults has demonstrated that producing simple rather than complex interventions is not always an inefficient strategy. Bramley, Lagnado, et al. (2015) showed that simple interventions tend to be more informative than complex interventions with respect to a broader hypothesis space (e.g., all possible three-variable causal models), with more complex interventions becoming more useful once the space of possibilities narrows to favor only a few overlapping causal hypotheses. In our task, children needed to discriminate among just three competing hypotheses, so it is one in which complex interventions are likely to be useful. In summary, the observed developmental changes can be interpreted as supporting the idea that whereas younger children used simple strategies that may have proved to be effective in other contexts, older children were more able to adjust their learning strategy in a way that was appropriate for the task—that is, to use a control of variables strategy (Chen & Klahr, 1999; Dean & Kuhn, 2007) in which confounding variables are experimentally controlled.

Experiment 2

Experiment 1 showed that by 6 or 7 years of age children can generate informative interventions and use these to derive the structure of a three-variable causal system and that the quality of their interventions predicts their performance. What we have not yet shown, however, is that children benefited from generating interventions themselves. Following Sobel and Kushnir's (2006) study with adults and Sobel and Sommerville's (2010) related study with children, we might predict that being able to self-generate interventions facilitates active hypothesis testing. In our second experiment, we tested additional groups of children using a "yoking" procedure similar to Sobel and Kushnir's procedure. Children did not select interventions themselves; rather, each child was individually matched

with one of the children from Experiment 1 and saw the outcomes of the set of interventions made by that child. There were two conditions; children in the yoked–self condition carried out a series of interventions on the system but did not select which interventions to make, and children in yoked–observe condition simply watched interventions being carried out by the experimenter. If children benefit from selecting interventions themselves, we would expect performance to be worse in the yoked conditions compared with the self-generated interventions in Experiment 1. If there are benefits from simply carrying out interventions oneself, even if one has not selected them, we would expect performance in the yoked–self condition to be better than performance in the yoked–observe condition.

In addition to comparing performance across conditions, we examined whether our measures of intervention quality, taken from the interventions made by the yoking participants from Experiment 1, were predictive of performance even in the yoked conditions. It is important to note that all participants in Experiment 2 received sufficient information to disambiguate the causal structures. Because all children in Experiment 1 obtained sufficient information to make correct inferences, the relation between intervention quality and performance found in that experiment can be explained in two potentially independent ways. It may be that this relation was obtained because higher quality interventions resulted in a larger quantity of useful information that was obtained in a sequence that facilitated learning. Alternatively, it may be that this relation was obtained because children who were engaged in productive hypothesis testing selected useful interventions that allowed them to assess these hypotheses. If the former is the case, we might expect to see a relation between intervention quality and performance in the yoked conditions of Experiment 2 because children in those conditions received identical sequences of information to the children who selected the interventions themselves. However, if this relation hinges on the role of active hypothesis testing, it might not be obtained in circumstances where children do not select interventions themselves.

Method

Participants

Three groups of children took part in this part of the study: 28 5- and 6-year-olds (M = 75 months, range = 70–81), 62 6- and 7-year-olds (M = 87 months, range = 81–93), and 50 7- and 8-year-olds (M = 99 months, range = 91–105), who were recruited from the same year groups and from schools in the same area as the child participants in Experiment 1. Half of the children were assigned to the yoked–self condition and half to the yoked–observe condition.

Materials

These were identical to those used in Experiment 1.

Procedure

The procedure for each of the conditions was identical to that of Experiment 1 up until the learning stage of the task. In the yoked–self condition, the experimenter explained that she was going to ask children to do things to the shapes in the box in order to figure out how the box worked, using cards to give instructions. She explained that when she pointed to a picture of a specific shape, children needed to move that shape to see what happened to the other shapes, and when she pointed to a picture of a shape with a stop sign in it, children needed to initially put the stop sign in that shape and then see what happened when they moved whichever shape was depicted in the next picture that she pointed to. Children were told that the experimenter would give them 12 instructions of this sort to start with. During the learning phase, the experimenter then instructed children to carry out the interventions in the same order as their yoked participants from Experiment 1. For those children who were yoked to children from Experiment 1 who completed 12 interventions (the majority of children), after those interventions were completed the experimenter said, "You have had your 12 goes now. Which picture do you think shows how the box works?" For the remaining children, after 12 interventions the experimenter said, "We have 6 more things to try before you give your guess about how the box works," and then showed the remaining 6 interventions before asking the test question. The procedure for the yoked–observe condition was very similar except that the experimenter explained that

she would point to the pictures of the shapes herself before moving them. During the learning phase, the experimenter then pointed to the appropriate pictures before each intervention and carried out all interventions herself with children observing.

Results

Fig. 4A shows the distribution of responses for each trial type for the yoked–self condition, and Fig. 4B shows the distribution for the yoked–observe condition. As can be seen from the figure, correct responses were relatively similar to those in Experiment 1, although the 7- and 8-year-olds seemed to perform less well on the causal chain. We compared performance in these conditions with performance from the yoked participants from Experiment 1; Table 3 shows the percentages of correct responses for the causal chain and common cause trials as a function of condition. There was no significant association between condition and numbers of correct responses for the causal chain trial, $\chi^2(2) = 1.48$, $p = .48$, or for the common cause trial, $\chi^2(2) = 3.89$, $p = .14$; the association between condition and numbers of correct responses for each trial type was also not significant if each age group was examined separately or if each condition was compared separately with the other two conditions. We then examined whether the factors that were predictive of performance in Experiment 1 predicted performance in the yoked conditions. Neither the proportion of informative interventions presented to participants, the efficiency of the sequences of interventions for identifying the true structure, nor their overall quality (assuming 25% forgetting) predicted performance for either the yoked–self or yoked–observe groups on either trial type (all $ps > .05$). The relation between these measures of interventions and performance on the causal chain identified in the 77 participants in Experiment 1 were still significant for the 70 whose data were yoked to the yoked–self and yoked–observe groups (informative interventions: $z = -2.69$, $p < .005$; intervention quality: $z = -2.50$, $p < .01$), indicating that the fact that the measures do not predict performance in the yoked groups is not a problem of slightly reduced statistical power.

General discussion

Experiment 1 examined children's causal structure learning under circumstances in which they selected and carried out interventions on a simple three-variable causal system. To the best of our knowledge, the analyses reported here of its data constitute the first attempt to model the quality of children's interventions when learning causal structure within a Bayesian framework. Our findings regarding children's interventional learning varied depending on whether children were learning a causal chain or a common cause structure. With regard to the former, there were clear developmental improvements not only in terms of accuracy of structure learning but also in terms of the quality of the interventions that children produced, as assessed in our Bayesian modeling. The key advantage of the modeling is that it provided us with a quantitative measure of the quality of children's interventions, allowing us to formally examine the extent to which children's interactions with the system were optimal. This Bayesian measure of intervention quality predicted performance. Put simply, the findings suggest that with development children increasingly resemble idealized Bayesian learners, although we note that the best predictor of performance from our modeling results was a measure of interventional quality that assumed some degree of noise in the Bayesian learning process.

The same pattern of findings was not obtained for the common cause structure, and the most plausible interpretation of this is that younger children's inferences in this task were based on a simple temporal heuristic ("assume a common cause if effects happen simultaneously") rather than on use of statistical information provided from interventions. Use of such temporal heuristics is widespread in both children's and adults' causal structure learning (Burns & McCormack, 2009; Fernbach & Sloman, 2009; Lagnado & Sloman, 2004, 2006; White, 2006), with McCormack et al. (2015) demonstrating that younger children's causal structure inferences are highly influenced by the temporal pattern of events. Their findings are consistent with those from the current study insofar as those authors also found no developmental improvements in the likelihood that children would give a common cause judgment under circumstances in which all events happened simultaneously. Children's

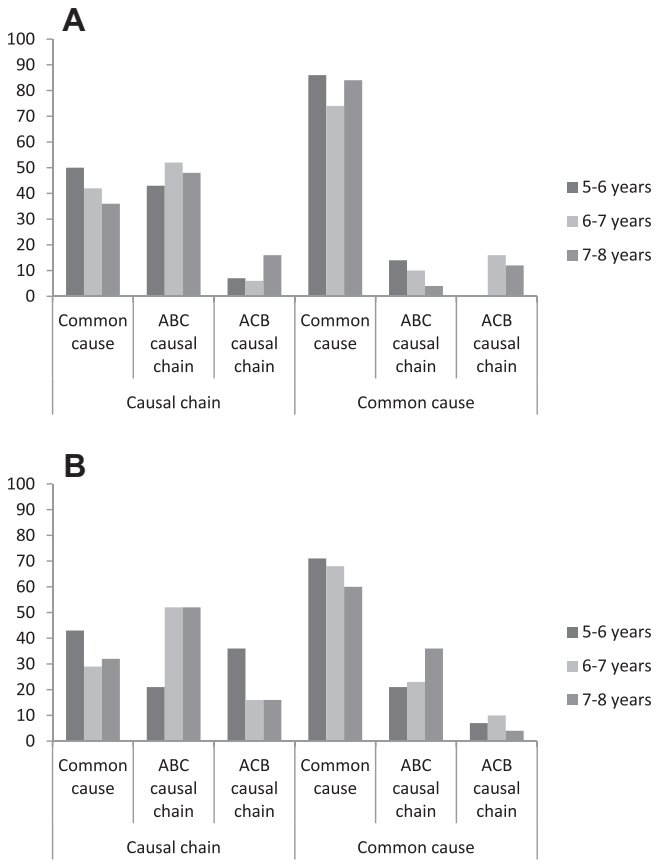

Fig. 4. (A) Percentage of response choices for each causal structure as a function of age group for the yoked–self condition. Correct responses to the causal chain are denoted as ABC. (B) Percentage of response choices for each causal structure as a function of age group for the yoked–observe condition. Correct responses to the causal chain are denoted as ABC.

Table 3
Percentage correct responses for each trial type as a function of condition.

	Causal chain	Common cause
Self-select (Experiment 1)	55.7	75.7
Yoked–self	48.6	80.0
Yoked–observe	45.7	65.7

tendency to recruit temporal heuristics is likely to be due to the heuristics' low demands on information processing in comparison with using statistical information (Fernbach & Sloman, 2009). For example, in the current study, use of such a heuristic would have been based on the observation of a single intervention—the temporal pattern of events following A+. This intervention was the most common one made by the younger groups; we interpreted this as suggesting that these children focus on producing an effect rather than systematically testing the competing hypotheses and in doing so are provided with evidence (i.e., the temporal pattern of events) that they take to be consistent with their existing hypothesis. Younger children were also less likely to disable components in the system, suggesting that they were less likely to try to exclude any variables.

Although younger children's interventions in the system had these characteristics, all children produced a set of interventions that could in principle have allowed them to correctly judge the causal structure. However, the Bayesian analysis proved to be useful in establishing that simply initially producing interventions that could potentially disambiguate the causal structure was not predictive of good performance. Rather, children's performance was related to how informative their interventions were as they moved through the task sequentially, with the Bayesian modeling making it possible to operationalize the informativeness of sequential intervention choices.

In general, as we argued above, our developmental findings are broadly consistent with some well-established findings in the literature on scientific learning. One of the aims of the scientific learning literature is to explore the wide variety of cognitive and metacognitive processes that are important at each stage of the reasoning process, from formulating initial questions, generating and evaluating evidence, to constructing and revising theories. Our aim in this study was not to match such studies in the breadth of cognitive processes that they explore or in the depth of analysis that they typically provide regarding, for example, the range of strategies that participants recruit. For a start, although we have described children as deciding between hypotheses, we do not believe that it is helpful to consider children to be engaged in theory construction in our task. Nevertheless, our findings are valuable in terms of what they add to our knowledge specifically about basic aspects of causal structure learning.

As already emphasized, in many instances scientists must uncover not just which variables are causal but also the overall structure of a causal system. Most studies of children's causal and scientific learning have focused on their ability to learn whether specific variables are causally efficacious. Although in scientific learning studies such variables sometimes have additive or interactive effects (e.g., Kuhn & Pease, 2008; Kuhn, Pease, & Wirkala, 2009), typically participants do not need to distinguish between, for example, common cause and causal chain structures. It is an important strength of the causal Bayes net approach that, unlike more traditional models of causal learning, it can model this sort of learning of structure. Our modeling work demonstrates the utility of examining such learning within a broadly Bayesian framework, albeit to allow us to conclude that younger children might not be the idealized Bayesian learners they are sometimes assumed to be within the causal Bayes net tradition.

Active versus passive learning

Our use of a yoking procedure in Experiment 2 provided insights into whether active participation assists children's causal learning. We found no clear learning benefits of either selecting or carrying out interventions oneself, a finding that contrasts with that of Sobel and Kushnir's (2009) adult study but is consistent with the findings of Lagnado and Sloman (2004). We note that there is a long-standing debate in the literature on children's scientific learning over whether "hands-on" learning is more beneficial than more passive learning (e.g., Bruner, 1961; Kirschner, Sweller, & Clark, 2006; Mayer, 2004). The debate regarding active versus passive learning is multifaceted, encompassing both cognitive and motivational issues, but in the context of children's scientific learning the central question has been whether children show benefits in terms of the amount, depth, and persistence of knowledge gained when allowed to engage in self-guided discovery learning. So, for example, Klahr and Nigam (2004) compared how well children learned a control-of-variables strategy in two conditions: either direct instruction or a condition in which they designed and carried out experiments themselves. On the basis of their findings, these authors argued that direct instruction is more beneficial than discovery learning. Unsurprisingly, such a conclusion has attracted much attention because it is believed to have important consequences for how children should be taught science (Hmelo-Silver, Duncan, & Chinn, 2007; Kuhn, 2007).

In interpreting the contribution of our findings to this debate, it is important to remember that our task is much simpler than those typically used in studies of the benefits of active learning. Moreover, children did not need to learn a theory or generalize from their learning, and the task was not one in which we could contrast hands-on learning with, for example, teacher instruction. Nevertheless, our task provided a context in which we could examine whether there were measurable benefits to actively selecting one's own interventions in circumstances where children need to decide between

different hypotheses. In the self-select condition, unlike in the other two conditions, children freely decided what information to sample and when to sample that information; this freedom with regard to information sampling can be seen as the most important difference between active and passive learning conditions (Gureckis & Markant, 2012). We found no evidence that such freedom yielded benefits. This suggests that, at least on the face of it, in learning causal structure, children do not benefit from being able to actively generate interventions in order to test specific hypotheses that they may be entertaining, as opposed to either just observing such interventions or carrying out interventions that were decided by another person.

However, the results also indicate a more nuanced conclusion because we examined not just whether levels of performance differed depending on learning condition but also whether the quality of interventions predicted performance even in yoked conditions. Interestingly, we found that this predictive relation did not hold in our yoked conditions. This finding is reminiscent of the results of correlational analysis conducted by Sobel and Kushnir (2006) on their adult data. They calculated the proportion of interventions that participants self-generated (or were exposed to in yoked conditions) that would provide information that was critical for learning a particular causal structure (i.e., a simple measure of intervention quality). They found that the correlation between the proportion of critical interventions and accuracy in causal structure judgments was significant only in conditions where participants generated the interventions themselves, not in yoked conditions. Their interpretation of this finding is that participants who generate critical interventions are better placed to interpret the outcome that results from an intervention and use it to distinguish between hypothesized structures.

Although we did not find that self-generating interventions boosted learning, our findings are at least compatible with Sobel and Kushnir's (2006) suggestion that learning may proceed in a different way under such conditions. When selecting interventions themselves, children who were testing hypotheses in an effective way were able to select informative interventions specifically to test particular hypotheses (consistent with Sobel & Kushnir's characterization of learning under this condition), and this may have underpinned the relation between quality of interventions and performance. Gureckis and Markant (2012) argued that when participants are selecting for themselves, they can preferentially select information that reduces their current uncertainty. Indeed, the measure of quality of interventions produced in our Bayesian modeling can be seen as a formal measure of the extent to which participants are adept at selecting interventions to reduce uncertainty, and it is this that predicted performance in the self-select condition.

However, this mode of learning was not available to children in the yoked conditions, which may be why there was no relation between quality of interventions and performance in these conditions. Note, however, that the unavailability of this mode of learning in the yoked conditions did not significantly impair performance. It is possible that the benefits of active hypothesis testing were outweighed in our study by the information processing demands of selecting and implementing interventions, demands that were reduced in the yoked conditions. Such a possibility is consistent with suggestions that there may be disadvantages associated with self-directed learning (for discussions, see Gureckis & Markant, 2012; Kirschner et al., 2006; Markant & Gureckis, 2014). Moreover, as Gureckis and Markant (2012) pointed out, active learning might not necessarily be beneficial under circumstances where learners are biased in their information seeking and focus, for example, on confirming preexisting erroneous hypotheses. As we have discussed, there is evidence that the younger children adopted such an approach to their intervention choices. There may be other circumstances, however, where children's causal learning would benefit from the opportunity to engage in active hypothesis testing, but the broader literature on self-directed learning has yet to clearly identify the situations where any advantages may outweigh potential disadvantages (Gureckis & Markant, 2012).

Conclusions

The findings of this study point to two clear directions for future work in this area. First, the fact that children become increasingly Bayes-efficient information seekers in their causal learning raises

the question of what cognitive changes underpin this developmental shift. There has been considerable discussion over the role of Bayesian modeling not only in a developmental context but also in providing an explanatory framework within cognitive psychology more generally (e.g., Bowers & Davis, 2012; Jones & Love, 2011; Schlesinger & McMurray, 2012). Here, we have used a Bayesian approach simply to provide a more formal analysis of how the quality of children's interventions improves developmentally. Questions regarding the developmental changes that account for these improvements cannot be directly addressed by the modeling reported here, which does not model cognitive processes (although we note that directly inspecting the patterns of children's interventions suggested that some developmental changes have parallels to those shown in the scientific learning literature). It is possible, however, that in the future models based on approximate Bayesian inference that attempt to be more psychologically plausible (Bramley, Dayan, et al., 2015; Kemp, Tenenbaum, Niyogi, & Griffiths, 2010; Sanborn, Griffiths, & Navarro, 2010; Shi, Feldman, & Griffiths, 2008) may play a role in addressing this question. Importantly, the developmental improvements found in our study highlight the need for Bayesian models that not only capture idealized learning but also can accommodate, and potentially explain, developmental changes in the quality of children's causal learning. Explaining developmental changes will require additional research that builds on the current findings but also tries to examine in more detail the role of particular types of cognitive processes such as children's strategies.

The second direction for future research stems from our finding that although active learning, in the form of self-selecting interventions, did not seem to benefit children's causal structure learning (beyond simply observing someone else's interventions), our modeling raised the possibility that learning may proceed in a qualitatively different way when children do not have the opportunity to choose interventions themselves. This finding demonstrated the potential benefits of examining not just whether causal learning is affected by whether or not individuals can engage in more active learning but also how it may be differentially related to the sort of information generated in active learning. Further examination of potential qualitative differences looks like an important goal for future empirical and modeling work on causal structure learning.

Acknowledgment

This research was supported by Grant RES-062-23-1799 from the Economic and Social Science Research Council (UK).

References

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*, 523–550.
Baron, J. (2005). *Rationality and intelligence*. Cambridge, UK: Cambridge University Press.
Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology, 64*, 215–234.
Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*, 389–414.
Bramley, N. R., Dayan, P., & Lagnado, D. A. (2015). Staying afloat on Neurath's boat: Heuristics for sequential causal learning. In *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 262–267). Austin, TX: Cognitive Science Society.
Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 708–721.
Bramley, N. R., Nelson, J. D., Speekenbrink, M., Crupi, V., & Lagnado, D. A. (2014). *What should an active causal learner value?* Poster presented at the 2014 annual meeting of the Psychonomics Society, Long Beach, CA.
Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*, 21–32.
Bullock, M., Gellman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time*. New York: Academic Press.
Burns, P., & McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Thinking & Reasoning, 15*, 167–196.
Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*, 1098–1120.
Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology, 79*, 102–133.
Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*, 341–349.
Dean, D., Jr., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education, 91*, 384–397.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 678–693.
Frosch, C. A., McCormack, T., Lagnado, D. A., & Burns, P. (2012). Are causal structure and intervention judgments inextricably linked? A developmental study. *Cognitive Science, 36*, 261–285.
Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.
Good, I. J. (1950). *Probability and the weighing of evidence*. London/New York: Charles Griffin/Hafner.
Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science, 337*, 1623–1627.
Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3–32.
Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*, 1205–1222.
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*, 620–629.
Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory. *Psychological Bulletin, 138*, 1085–1108.
Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7*, 464–481.
Hagmayer, Y., Sloman, S., Lagnado, D., & Waldmann, M. (2007). Causal reasoning through intervention. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford, UK: Oxford University Press.
Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*, 99–107.
Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioural and Brain Sciences, 34*, 169–231.
Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition, 114*, 165–196.
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86.
Klahr, D., & Dunbar, K. (1988). Dual-space search during scientific reasoning. *Cognitive Science, 12*, 1–48.
Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25*, 111–146.
Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*, 661–667.
Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology, 64*, 141–152.
Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*, 674–689.
Kuhn, D. (2007). Is direct instruction an answer to the right question? *Educational Psychologist, 42*, 109–113.
Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. San Diego: Academic Press.
Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction, 26*, 512–559.
Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating the effects of multiple variables: A skill fundamental to scientific thinking. *Journal of Experimental Child Psychology, 103*, 268–284.
Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 44*, 186–196.
Kushnir, T., Gopnik, A., Lucas, C., & Schulz, L. (2010). Inferring hidden causal structure. *Cognitive Science, 34*, 148–160.
Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 856–876.
Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 451–460.
Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General, 143*, 94–122.
Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*, 14–19.
McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 395–416.
Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review, 118*, 120–134.
Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*, 979–999.
Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education, 36*, 334–354.
Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144–1167.
Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31–57.
Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*, 102–119.
Schlesinger, M., & McMurray, B. (2012). The past, present, and future of computational models of cognitive development. *Cognitive Development, 27*, 326–348.
Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences, 16*, 382–389.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*, 1045–1050.
Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, C. A. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition, 109*, 211–223.
Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science, 10*, 322–332.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.
Shi, L., Feldman, N. H., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 745–750). Austin, TX: Cognitive Science Society.
Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47* (1, Serial No. 194).
Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*, 5–39.
Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition, 34*, 411–419.
Sobel, D. M., & Sommerville, J. A. (2009). Rationales in children's causal learning from others' actions. *Cognitive Development, 24*, 70–79.
Sobel, D. M., & Sommerville, J. A. (2010). The importance of discovery in children's causal learning from interventions. *Frontiers in Psychology, 1*. http://dx.doi.org/10.3389/fpsyg.2010.00176.
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*, 303–333.
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453–489.
Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 216–227.
Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*, 776–784.
White, P. A. (2006). How well is causal structure inferred from co-occurrence information? *European Journal of Cognitive Psychology, 18*, 454–480.
Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235–269.
Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99–149.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.