Thinking counterfactually supports children's ability to conduct a controlled test of a hypothesis

Angela Nyhout (angela.nyhout@utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6

Alana Iannuzziello (alana.iannuzziello@mail.utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6

Caren M. Walker (carenwalker@ucsd.edu)
Department of Psychology, University of California San Diego,
9500 Gilman Drive, La Jolla, CA 92093 USA

Patricia A. Ganea (patricia.ganea@utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6

Abstract

Children often fail to control variables when conducting tests of hypotheses, yielding confounded evidence. We propose that getting children to think of alternative possibilities through counterfactual prompts may scaffold their ability to control variables, by engaging them in an imagined intervention that is structurally similar to controlled actions in scientific experiments. Findings provide preliminary support for this hypothesis. Seven- to 10-year-olds who were prompted to think counterfactually showed better performance on post-test control of variables tasks than children who were given control prompts. These results inform debates about the contribution of counterfactual reasoning to scientific reasoning, and suggest that counterfactual prompts may be useful in science learning contexts.

Keywords: cognitive development; scientific reasoning; counterfactual reasoning; causal learning; science education

Scientific Reasoning in Development

Equipping children with scientific inquiry skills is a core objective of elementary science education, allowing children to collect evidence and draw inferences about the world around them. However, extensive research has found that children are relatively unequipped to engage in many aspects of scientific inquiry in the absence of direct instruction and frequent scaffolding (Klahr, Fay, & Dunbar, 1993; Klahr & Nigam, 2004; Kuhn & Franklin, 2006; Schauble, 1996). In the present study, drawing from research and theory in cognitive development, science education, and philosophy, we investigate the use of a novel pedagogical tool – counterfactual reasoning prompts – to scaffold children's scientific reasoning skills.

An important sub-skill of scientific inquiry is the ability to control variables. This skill, termed the control-of-variables strategy (CVS) has received a great deal of attention in research on scientific reasoning over the past four decades (for a review, see Zimmerman, 2007). To properly execute this skill, the learner should isolate a single variable at a time, while holding all else constant.

Consider a common task used in studies investigating CVS (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004). Children are presented with a set of ramps that can be varied along a number of dimensions (e.g., ramp height, surface, run length, ball size) and their task is to manipulate the ramps to determine the effect of different variables on where a ball stops after rolling down the ramp. To make warranted inferences about individual variables, the learner should change the values of a single variable (e.g., compare a high ramp to a low ramp), keeping all other variables constant (e.g., smooth surface, same-size balls).

Although children are able to *recognize* a conclusive test of a hypothesis as young as age 6 (Sodian, Zaitchik, & Carey, 1991), they typically fail to *produce* one themselves in the absence of scaffolding through middle childhood (Klahr, Zimmerman, & Jirout, 2011; Zimmerman, 2007). However, with direct instruction, children often show improvement in their ability to design controlled experiments (Chen & Klahr, 1999; Klahr & Nigam, 2004; for a meta-analysis, see Schwichow, Croker, Zimmerman, Hoffler, & Hartig, 2016). For instance, Chen and Klahr (1999) found that 7- to 10-years-old who were given explicit instruction on CVS were better able to transfer this strategy to both similar and dissimilar problems than those who engaged in self-guided inquiry. Younger children frequently failed to design unconfounded tests.

Although past studies have found that children are able to learn the control of variables strategy through direct instruction or demonstrations, science curricula and educational guidelines often recommend teaching scientific inquiry skills through *inquiry-based learning* instead (e.g., US National Research Council, 2000). That is, children's scientific inquiry skills are thought to be best supported by having children explore science concepts based on their own

observations and experiences with phenomena of interest, with little explicit instruction from educators. Thus, there is significant educational value to identifying methods for scaffolding children's hypothesis testing abilities that not only fit within these curricular guidelines, but also harness children's intuitive reasoning skills.

Causal and Counterfactual Reasoning

Whereas the work reviewed above suggests that older children are poor at testing and revising hypotheses, another body of research shows that children are adept at parallel skills when engaging in causal learning tasks.

From a young age, children form, test, and revise hypotheses in building informal theories in various domains (Carey, 1985; Gopnik, Meltzoff, & Bryant, 1997; Keil, 1992;). For instance, toddlers are able to infer higher-order relational causes (Walker & Gopnik, 2014). Preschoolers are able to draw appropriate causal inferences from patterns of dependence, even when evidence conflicts with their prior knowledge (Schulz & Gopnik, 2004), and use evidence from interventions to make inferences about causal structure (Schulz, Glymour, & Gopnik, 2007).

Why do older children (and even adults) fail when applying this skill-set in scientific reasoning contexts? We suggest a few possible explanations for this discrepancy. First, studies of intuitive causal reasoning with toddlers and preschoolers use tasks that are typically decontextualized, placing relatively few demands on children's prior knowledge. Many of these studies rely on a "blicket detector" paradigm, in which children are familiarized with a novel machine, and their task is to determine what makes it switch on (Gopnik & Sobel, 2000). In contrast, scientific reasoning tasks given to older children typically use knowledge-laden tasks that rely heavily on children's existing (and often incorrect) knowledge and theories (e.g., Chen & Klahr, 1999). Second, causal reasoning tasks typically measure children's abilities implicitly, whereas scientific reasoning tasks ask children to explicitly plan and often verbally demonstrate their abilities. Despite these differences, both classes of studies rely on a common set of domain-general inferential skills, including the ability to form and revise hypotheses on the basis of available evidence.

How do we connect the parallel mechanisms children successfully apply in causal reasoning tasks to scientific reasoning contexts? In the current study, we explore the claim that counterfactual reasoning is fundamental to causal and scientific reasoning, and suggest that counterfactual prompts may help to connect these abilities. When we think counterfactually, we compare the way things are to the way things *could have been*. Counterfactual reasoning therefore necessarily involves thinking about causes: As one considers how an event could have turned out differently, one reasons about the causal relationship between an antecedent and outcome. If the event X had not happened, would event Y still have happened? If the answer to this is "no", one can conclude that event X is a cause of event Y (Lewis, 1986).

However, the utility of counterfactual reasoning may not be limited to drawing *specific* causal inferences. Several researchers have drawn theoretical parallels between the *mechanisms* underlying counterfactual reasoning and scientific reasoning (e.g., Buchsbaum, Bridgers, Weisberg, & Gopnik, 2012; Erb & Sobel, 2014; Gopnik & Walker, 2013; Sloman, 2005; Rafetseder & Perner, 2014; Walker & Gopnik, 2013). If a learner believes that X caused Y, they can mentally intervene on X by imagining that it did not occur, follow the causal implications of this change, and then reason about whether it would have led to a change in Y (Gopnik & Walker, 2013; Walker & Gopnik, 2013). We follow an identical process in *scientific* reasoning. We hypothesize that X causes Y, and then make plans to systematically manipulate X in order to investigate its impact on Y. In both counterfactual and scientific reasoning, the learner adjusts a causal system by (mentally or physically) intervening on one event and considering the effects of this change.

Despite the proposed contribution of counterfactual reasoning to science learning, there is relatively little research connecting the two (Engle & Walker, 2018; Frosch, McCormack, Lagnado, & Burns, 2012; Schulz, et al., 2007) and no work linking these capacities to hypothesis testing in children. Only two previous studies to our knowledge have investigated the relationship between counterfactual reasoning and scientific inquiry. Adults primed with counterfactuals were better able to conduct a disconfirming test of a hypothesis than those given neutral primes (Galinsky & Moskowitz, 2000). In another study, counterfactual prompts scaffolded children's ability to detect anomalies to an existing hypothesis in a causal learning task (Engle and Walker, 2018).

Given that counterfactual and scientific reasoning both involve intervening on a single variable to investigate its causal role in an outcome of interest, we propose that engaging children in counterfactual reasoning during a control-of-variables task will scaffold their ability to conduct a controlled test of a hypothesis by activating a parallel underlying cognitive mechanism.

That said, it is worth first considering whether children of the age we tested in the current study (7 to 10 years) are capable of counterfactual reasoning, given the lively debate about its developmental trajectory. Previous research has been mixed, with some findings indicating that children *can* reason counterfactually as young as 3-½ years (Harris, German, & Mills, 1996), and other work suggesting that this ability does not reach maturity until adolescence (e.g., Rafetseder, Schwitalla, & Perner, 2013). However, more recent work suggests that studies showing counterfactual reasoning to be late-developing may have underestimated children's ability by presenting opaque causal structures and by placing large demands on children's memory (McCormack, Ho, Gribben, O'Connor, & Hoerl (2018; Nyhout, Henke, & Ganea, 2019). A recent set of studies demonstrates that children reason counterfactually by age 4 when given a clear and novel causal structure that does not

rely on their background knowledge (Nyhout & Ganea, 2019). Thus, we conclude from these findings that children have the requisite abilities to engage in counterfactual reasoning well before the age of those in the current study.

Current Study

In contrast to previous research (Chen & Klahr, 1999; Klahr & Nigam, 2004), we investigated whether children's ability to control variables could be scaffolded in non-school settings. We also reduced task demands by including a smaller number of variables (2 variables, rather than 4).

Children in the present study were assigned to either a counterfactual or control condition. After watching a video of an actor conducting a controlled test of a hypothesis, children were given either a *counterfactual* prompt, asking them to consider what would happen if the actor had conducted her test differently, or a *control* prompt, in which children were asked to recall what had happened. We predicted that children given counterfactual prompts would be more likely to improve from pre-test to post-test than children given control prompts. We tested a range of ages typically used in CVS research (7 to 10 years), but did not have prior predictions about age-related differences in performance.

Method

Participants

Participants aged 7 to 10 years of age were recruited and tested at a museum in a large urban area. The final sample included 88 children (M = 8.91, SD = 1.13, range = 7.00 to 10.97, 45 girls) whose data are reported below. Participants were placed in two categories, based on their age. The *younger* age category included children between the ages of 7.00 and 8.99 (n = 46, M = 8.00, SD = 0.63) and the *older* age category included children between the ages of 9.00 to 10.99 (n = 42, M = 9.90, SD = 0.59), with categories selected on the basis of similar previous studies (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004). Participants who passed the pre-test phase (n = 24) were excluded as they were determined to be already competent with CVS. Three additional participants were excluded due to experimenter error (n = 2) or language barriers (n = 1).

Materials

For the pre- and post-test phases described below, participants were given two identical ramps with both a down- and up-ramp side. The ramps were ridged on the up-ramp where the ball could stop (Figure 1). Each ridge was painted a different color to allow for unambiguous reference and measurement. There were four binary variables, but participants received only two of the four variables at a time, and the remaining two variables were "fixed". The variables were paired as follows: (1) height (high or low) and ball size (large or small), or (2) starting place (top or middle), and surface (rough or smooth). For instance, at one time-point, participants were given a large and small ball for each ramp, and pieces to adjust the steepness of each ramp ("high" or "low"). At the other time-point, participants were given a rough surface and smooth surface for each ramp, and a piece of cardboard to adjust where the ball started for each.

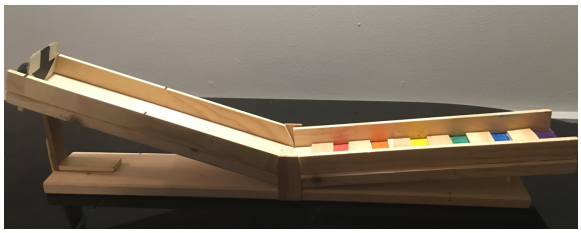

Figure 1: One of two identical ramps used in the study. A ball is launched from the down-ramp (left) and stops on one of the coloured ridges on the up-ramp (right). The apparatus can be adjusted for height, surface type, where the ball starts on the down-ramp, and ball size.

The same set of ramps were used in a video in the scaffolding phase, displayed for participants on a laptop.

Procedure

The study included a warm-up activity (uncertainty training) followed by pre-test, scaffolding, and two post-test phases. Participants were assigned to one of two conditions for the scaffolding phase: counterfactual (n = 45, M age = 8.47, 23 girls) or control (n = 43, M age = 8.44, 22 female girls). The order of all variables and variable sets were counterbalanced between participants.

Uncertainty Training. Given that some of the prompts in the intervention phase required children to acknowledge their uncertainty about an outcome, we included an uncertainty training phase to ensure children were able to recognize and acknowledge their uncertainty. All children, regardless of condition, received the same uncertainty training. Using cards with various colors and suits, the experimenter placed a pair of cards face down, and turned over one card. Before revealing the second card, she asked the participant if they could be "sure or not sure" if the face-down card was the same or different as the face-up card. Regardless of the participant's response, the experimenter instructed children that they cannot be sure if the two cards are the same, and that it is okay to answer the question in this way. The process repeated until the participants answered that they could not be sure three times.

Pre-test. The experimenter placed the two ramps next to each other, directly in front of the participant, and explained that the two ramps were similar and worked the same way. She then showed participants how to operate and adjust the ramps along two of the variables (e.g. height of ramp and size of ball). The other two variables (e.g. ramp surface and run length) were fixed and not introduced until the *post-test transfer phase*. Participants were asked to demonstrate how to manipulate the ramps. If they did not set up the ramp correctly, the experimenter showed them again. All

demonstrations were performed with one ramp, and participants were reminded that the ramps were the same.

To measure children's ability to execute CVS, the experimenter asked them to show how they would find out if one variable plays a role in how far the ball travels down the ramp (e.g., "Can you show me how you would find out if *the size of ball* matters for how far the ball goes down the ramp?"). She told participants they had one chance to set up both ramps at the same time, and then repeated the question a second time. Participants were required to set up both ramps before launching the balls down each ramp one at a time. After each ball was launched, the experimenter labeled the outcome (e.g., "Look! The ball stopped on the yellow line.") but did not compare between the two ramps.

Using the same procedure, the experimenter then asked participants to determine if a second variable mattered for how far the ball would travel down a ramp (e.g., "Can you show me how you would find out if *the height of the ramp* matters for how far the ball goes down the ramp?").

Participants who controlled the correct variable received 1 point for each question for a maximum score of 2. Participants who received a score of 2/2 at pre-test were excluded from the study (and the study was terminated at this point), because they already possessed an understanding of CVS ($n = 24$). Participants who received scores of 0 or 1 went on to the scaffolding phase.

Scaffolding. Participants in this phase watched two videos of an actor exploring the ramps and were told that they would be asked about what they saw after each video. The actor in the videos manipulated the same two variables that participants were asked to isolate during the pre-test, using the same ramps. The video started with the actor stating that she was going to find out if a variable (e.g., height of the ramp) played a role in how far the ball travelled down the ramps. The actor then proceeded to set-up the ramps and labelled the set-up as she went along (e.g., "I'm going to set Ramp 1 to high"). After she set-up both ramps, she launched the balls one at a time and labeled the outcome by stating the color the ball landed on. At the end of the video she stated which ball (on Ramp 1 or Ramp 2) travelled farther. The experimenter then paused the video so that the participants could see the outcome of both ramps at the same time. The videos were identical across conditions; the only difference was in the question prompts children were asked after.

In the *counterfactual condition*, participants were asked to imagine a change to the value of a variable (e.g., "Let's imagine that she set Ramp 1 to low. Would the ball have travelled down the ramp farther on Ramp 1, farther on Ramp 2, or you can't be sure?"). This imagined change would create a confounded (or uncontrolled) test. In the *control condition*, participants were asked to recall what had happened (e.g., "Let's imagine again what happened to the ball on Ramp 1? Did the ball travel farther on Ramp 1, farther on Ramp 2, or you can't be sure?"). Children did not receive feedback on their responses during the scaffolding phase in either condition.

In both conditions, a second video was shown highlighting the other variable (e.g., size of ball). In the counterfactual condition, the experimenter asked the participants to imagine a change to the value of this new variable (e.g., size of ball), creating another confounded test. In the control condition, the experimenter asked the participants the same question as before, but highlighted the other ramp (e.g., Ramp 2).

Post-Test Same. The experimenter removed the laptop and placed the ramps side-by-side in front of the participant. This phase was identical to the pre-test, except participants were not asked to demonstrate how the ramps worked. Responses were coded in the same way, with participants receiving a maximum score of 2.

Post-Test Transfer. The experimenter then told participants that the ramps can work in a different way. The two original variables were fixed (e.g., ramps could only be set to high, and only the big balls could be used) and two new variables were introduced (e.g., surface of the ramp and starting position for the ball). As in the pre-test, the experimenter showed participants how the new variables worked on the ramps and asked participants to demonstrate how to manipulate each new variable.

The procedure was the same as the pre-test and post-test same phases except participants were asked two new questions about each of the new variables (e.g. "Can you show me how you would find out if *the surface of the ramp*/*where the ball starts on the ramp* matters for how far the ball goes down the ramp?"). Again, participants could receive a score up to 2 across the two test questions.

In sum, participants were asked two questions each at pre-test, post-test same, and post-test transfer, and received a score between 0 and 2 for the number of controlled tests they conducted in each phase. In each counterbalancing order, the pre-test and post-test same phases were identical, whereas the post-test transfer phase used two previously unencountered variables. The experimenter live-recorded with paper-and-pencil, and later checked videos for accuracy. A second researcher coded 34% of videos, and inter-rater reliability was excellent (96.6% agreement, Fleiss' $\kappa = 0.93$, $p < .001$).

Results

We first tested whether there were differences between the two conditions at pre-test using a Chi-Square test of independence, and found no significant differences across conditions in pre-test score, $p = .206$. We also found no significant differences between genders ($U = 926$, $p = .687$) or the variable set participants received at pre-test ($U = 902$, $p = .522$), thus we do not consider these variables further.

To investigate the change in children's score (CVS score out of 2) from pre-test to *(1) post-test same* and *(2) post-test transfer*, we conducted two generalized estimating equation (GEE) analyses with multinomial distributions and cumulative logit-log link functions with condition (counterfactual or control) and age group (younger or older) as predictor variables.

For the GEE of pre-test vs. *post-test same* performance, there was a main effect of test, B = -2.56, SE = 0.56, Wald $\chi2(1) = 20.84$, $p < .001$, such that children were 12.82 times more likely to receive a higher score at *post-test same* than at pre-test, Exp(B) = 0.78, 95% CI = [0.03, 0.23]. There was also a main effect of age, B = -1.66, SE = 0.76, Wald $\chi2(1) = 4.78.$, $p = .029$, such that older children were 5.26 times more likely to receive a higher score than younger children, Exp(B) = 0.19 95% CI = [0.04, 0.84]. The main effect of condition was not significant, $p = .436$. The test phase by age category interaction was significant, B = 1.59, SE = 0.72, Wald $\chi2(1) = 4.85.$, $p = .028$, such that older children in the post-test same phase were 4.90 times more likely to receive a higher score than younger children in the post-test same phase, Exp(B) = 4.90, 95% CI = 1.19, 20.13] All other interactions were non-significant.

For the GEE of pre-test vs. *post-test transfer*, there was again a main effect of test, B = -1.59, SE = 0.51, Wald $\chi2(1) = 9.70$, $p = .002$, such that children were 4.90 times more likely to receive a higher score on the *post-test transfer* phase than the pre-test phase, Exp(B) = .204, 95% CI [0.08, 0.55]. There was also a main effect of age, B= -1.35, SE = 0.68, Wald $\chi2(1) = 3.92$, $p = .048$, such that older children were 3.85 times more likely to receive a higher score than younger children, Exp(B) = 0.26, 95% CI = [0.069, 0.987]. The main effect of condition was marginally significant, B = 1.31, SE = 0.67, Wald $\chi2(1) = 3.819$, $p = .051$. Children in the counterfactual condition were 3.71 times more likely to receive a higher score than those in the control condition, Exp(B) = 3.71, 95% CI = [1.00, 13.78]. All interactions were non-significant.

We conducted planned post-hoc comparisons to further investigate performance between groups at each test-phase using Chi-square tests of independence. Performance differed significantly between children in the counterfactual and control conditions at both *post-test same* $\chi^2(2) = 7.28$, $p = .026$ and *post-test transfer* $\chi^2(2) = 6.04$, $p = .049$. Table 1 presents the relevant proportions of children who conducted 0, 1, and 2 controlled tests in each test phase entered into the Chi-square analyses.

Table 1: Proportion of children who conducted 0, 1, or 2 controlled tests in each post-test phase (CVS Score).

		CVS Score (/2)		
Post-test	Condition	*0*	*1*	*2*
Same	Counterfactual	11.1	33.3	55.6
	Control	34.9	20.9	44.2
Transfer	Counterfactual	24.4	22.2	53.3
	Control	41.9	30.2	27.9

Finally, we considered the relation between children's responses to counterfactual prompts in the scaffolding phase and their CVS scores, although we did not make predictions about any such relation. Recall that the correct answer to the counterfactual prompts was "can't be sure", because the counterfactual intervention created a confounded test. Of the 45 children in the counterfactual condition, 13 (29%) answered "can't be sure" to both prompts, 19 (42%) answered "can't be sure" to 1/2, and 13 (29%) did not answer "can't be sure" to either prompt. Children's "can't be sure" responses did not significantly correlate with their performance on any of the CVS tests, Spearman's rho = -.121 to -.231, $p = .127$ to .430.

Discussion

We proposed that prompting children to think counterfactually during a control-of-variables task would scaffold their performance by capitalizing on their underlying causal reasoning skills. The results of this study provide initial support for this proposal. Children given counterfactual prompts showed better performance on the post-test phases than those given control prompts, though these differences were non-significant on post-test same and marginally significant on post-test transfer in the omnibus analyses. Critically, when considering condition differences alone, children in the counterfactual condition performed significantly better than those in the control condition at both post-tests. The largest proportion of control group children scored 0/2 on both post-tests, whereas the largest proportion of counterfactual group children scored 2/2, as displayed in Table 1.

Along with these condition differences, there was also an indication that the video demonstration alone improved children's ability to control variables, given that we found significant main effects of test phase, but no condition by test-phase interaction. The actor did not explicitly comment on the strategies she was using, and the demonstration was devoid of ostensive pedagogical signals (Csibra & Gergely, 2009) that were present in many previous CVS studies (Schwichow et al., 2016). Future work may consider the role of similar demonstrations and counterfactual prompts separately to identify the extent to which they may yield different benefits.

Our findings are surprising in light of previous studies, which found that children required more intensive instruction and scaffolding in order to improve, with some of these interventions even taking place over the course of several sessions (e.g., Schauble, 1996). Even with a subtle manipulation in the form of two counterfactual questions following a demonstration, children showed improvement in their ability to conduct a controlled test of a hypothesis.

Children in the counterfactual condition were able to conduct a controlled test both on the variables they had already encountered and on two new variables, with more than half of children in the counterfactual condition scoring 2/2 on both post-tests. In contrast, children in the control condition showed less evidence of transfer, with a minority of children scoring 2/2 in the post-test transfer phase.

These findings provide preliminary evidence that counterfactual prompts may be a promising pedagogical tool for supporting CVS. However, these results do not allow us to pinpoint the precise mechanism by which counterfactuals may confer this benefit. We have suggested that

counterfactuals may serve as *imagined interventions*, helping learners to connect their intuitive causal reasoning abilities to the current task. This suggestion is in line with previous work emphasizing the relation between causal and counterfactual reasoning (e.g., Gopnik & Schulz, 2007; Sloman 2005; Gopnik & Walker, 2013).

However, other work suggests that counterfactuals may have a *general* effect on reasoning by activating a "mindset" that is open to alternatives. Previous research shows that prompts to consider alternatives in the form of counterfactuals (Galinsky & Moskowitz, 2000) or multiple explanations (Hirt & Markman, 1995) have wide-reaching effects, with individuals showing generally debiased reasoning across a range of settings. Researchers studying these effects have suggested that counterfactuals activate a *mental simulation mindset* that breaks the reasoner free of a singular viewpoint or hypothesis and incites consideration of alternative, and potentially contrasting possibilities. In ongoing research, we are currently investigating whether children prompted with counterfactuals on one task (e.g., ramps) show improvement on a far-transfer task (e.g., pendulums) to better understand the potential mechanisms by which counterfactual prompts may support performance. In the present study, our counterfactual questions were about the experimental design and specifically pertained to the control-of-variables process. It is an open question whether counterfactual questions about a peripheral or irrelevant feature of the task (e.g., the color of the ball) would scaffold performance. An alternative "mindset" account would predict that counterfactuals should be beneficial regardless of their focus.

Our counterfactual prompts not only focused on the control of variables process, but also specifically invited children to imagine a *confounded test*. An alternate explanation for children's success in the counterfactual condition may therefore be that by engaging children in imagining a confounded test, our prompts led them to recognize that such tests were inconclusive and that they should avoid producing such tests themselves. However, the lack of a relation between children's "can't be sure" responses and their ability to control variables suggests that children did not need to explicitly recognize the inconclusiveness of a confounded test in order to benefit from the process of thinking counterfactually. In other words, the effect of the counterfactual prompts appears to be distinct from the specific response they elicit. This finding aligns with research on children's self-explanation showing that the *process* of generating explanations benefits children's causal reasoning, regardless of the specific explanations they produce (e.g., Walker, Lombrozo, Legare, & Gopnik, 2014).

Another possibility is that our counterfactual prompts drew children's attention to both values of the variable that was held constant (e.g., "Let's imagine that she set Ramp 1 to low" when she had set both ramps to high), whereas the control prompts did not (e.g., "Let's imagine again what happened to the ball on Ramp 1"). This may have made children more likely to consider and control the alternate variable. In a follow-up study, we have adapted our control prompt to highlight both levels of the alternate variable to investigate whether this accounts for children's better performance in the counterfactual condition.

Although we are not yet able to identify the precise mechanism by which counterfactuals confer the benefits observed, these findings connect to a wider body of results that suggest that drawing children's attention to alternatives benefits their scientific inquiry (e.g., Sodian et al, 1991; Engle & Walker, 2018). For instance, children in Sodian et al. (1991) were able to recognize a conclusive test of a hypothesis when presented with two contrasting hypotheses, and, as mentioned above, Engle and Walker (2018) found that counterfactual prompts scaffolded children's ability to detect anomalies during causal learning. These results suggest that thinking of counterfactuals and alternatives may benefit a range of scientific inquiry skills.

Conclusion

Children prompted to think counterfactually showed improvements in their ability to conduct controlled tests of a hypothesis – an ability previous studies have suggested requires direct instruction or intensive scaffolding. These results support theoretical proposals about the role of counterfactuals in scientific reasoning, and suggest that counterfactuals may have educational utility. The prompts used in the current study are short and simple, and could easily be implemented in a range of formal and informal learning contexts.

Acknowledgments

This work was supported by funding from the Social Sciences and Humanities Research Council of Canada, including an Insight Grant to P.A. Ganea, a Postdoctoral Fellowship Award to A. Nyhout, and a Canada Graduate Scholarship-Master's to A. Iannuzziello. We are grateful to the children and families who participated in this research, and to the Ontario Science Centre for facilitating data collection. We thank Hanna Lim for assistance with building stimuli, transcription, and coding, Jayun Bae for acting in videos, and Etri Kocaqi for assistance with coding.

References

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development*, *11*, 523-550.

Buchsbaum, D., Bridgers, S., Weisberg, D. S., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*, 2202-2212.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098-1120.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*, 148-153.

Engle, J., & Walker, C.M. (2018). Considering alternatives facilitates anomaly detection in preschoolers. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society. Madison, WI: Cognitive Science Society.

Erb, C.D. & Sobel, D.M. (2014). The development of diagnostic reasoning about uncertain events between ages 4-7. *PLOS One, 9(3):* e92285.

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.

Galinsky, A. D., & Moskowitz, G. B. (2000). Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives. *Journal of Experimental Social Psychology, 36*, 384–409.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Schulz, L. (2007). *Causal learning: Psychology, philosophy, and computation.* New York, NY: Oxford University Press.

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., & Walker, C. M. (2013). Considering counterfactuals: The relationship between causal learning and pretend play. *American Journal of Play*, *6*, 15-28.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*, 233-259

Hirt, E.R., & Markman, K.D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality & Social Psychology*, *69*, 1069-1086.

Keil, F. C. (1992). *Concepts, Kinds, and Cognitive development*. Cambridge, MA: MIT Press.

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.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*, 971-975.

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how)?. In D. Kuhn & R. Siegler (Eds.), *Cognition, perception, and language. Volume 2 of the Handbook of child psychology* (6th ed.). Hoboken, NJ: Wiley.

Lewis, D. K. (1986). *On the plurality of worlds* (Vol. 322). Oxford: Blackwell.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, *45*, 1-9.

National Research Council. (2000). Inquiry and the national science education standards: A guide for teaching and learning. *National Academies Press*.

Nyhout, A., Henke, L., & Ganea, P. A. (2019). Children's counterfactual reasoning about causally overdetermined events. *Child Development*, *9*, 610-622.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition*, *183*, 57-66.

Rafetseder, E., & Perner, J. (2014). Counterfactual reasoning: Sharpening conceptual distinctions in developmental studies. *Child Development Perspectives, 8*, 54–58.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*, 389-404.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*, 102–119.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*, 162-176.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*, 322-332.

Schwichow, M., Croker, S., Zimmerman, C., Hoffler, T., & Hartig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, 39, 37–63.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Cambridge, MA: Oxford University Press.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753-766.

Walker, C.M. & Gopnik, A. (2013). Causality & Imagination. In Marjorie Taylor (Ed.), *The development of imagination*. Oxford University Press: New York.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, *25*, 161-169.

Walker, C. M., Lombrozo, T., Legare, C., & Gopnik, A. (2014). Explanation prompts children to privilege inductively rich properties. *Cognition, 133,* 343-357.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223.