Journal of Experimental Child Psychology 224 (2022) 105509

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

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

Contributions of causal reasoning to early scientific literacy

Margaret Shavlik [a], Özgün Köksal [b], Brian F. French [c], Catherine A. Haden [d], Cristine H. Legare [e], Amy E. Booth [a,*]

[a] Vanderbilt University, Nashville, TN 37235, USA
[b] Ludwig Maximilian University of Munich, 80539 Munich, Germany
[c] Washington State University, Pullman, WA 99164, USA
[d] Loyola University Chicago, Chicago, IL 60660, USA
[e] The University of Texas at Austin, Austin, TX 78712, USA

ARTICLE INFO

Article history:
Received 30 September 2021
Revised 16 June 2022
Available online 16 July 2022

Keywords:
Causal reasoning
Scientific literacy
Preschool children
Confirmatory factor analysis
Cross-lagged panel model
Cognitive development

ABSTRACT

Although early causal reasoning has been studied extensively, inconsistency in the tasks used to assess it has clouded our understanding of its structure, development, and relevance to broader developmental outcomes. The current research attempted to bring clarity to these questions by exploring patterns of performance across several commonly used measures of causal reasoning, and their relation to scientific literacy, in a sample of 3- to 5-year-old children from diverse backgrounds (N = 153). A longitudinal confirmatory factor analysis revealed that some measures of causal reasoning (counterfactual reasoning, causal learning, and causal inference), but not all of them (tracking cause–effect associations and resolving confounded evidence), assess a unidimensional factor and that this resulting factor was relatively stable across time. A cross-lagged panel model analysis revealed associations between causal reasoning and scientific literacy across each age tested. Causal reasoning and scientific literacy related to each other concurrently, and each predicted the other in subsequent years. These relations could not be accounted for by children's broader cognitive skills. Implications for early STEM (science, technology, engineering, and math) engagement and success are discussed.

* Corresponding author.
E-mail address: amy.booth@vanderbilt.edu (A.E. Booth).

https://doi.org/10.1016/j.jecp.2022.105509
0022-0965/

Introduction

Both the educational community (e.g., National Research Council [NRC], 2012) and acadefmic community (Callanan et al., 2020; Gopnik, 2012; Legare, 2014; Sobel, Erb, Tassin, & Weisberg, 2017) argue that causal reasoning is fundamental to scientific literacy. Many of the science-related questions introduced to children in early educational contexts are aimed at understanding causal relations, processes, and transformations (e.g., why ice melts or plants die; Bulunuz, 2013; Mantzicopoulos, Patrick, & Samarapungavan, 2013). Moreover, cause–effect understanding is foundational to experimentation because it depends on manipulating potential causes and observing outcomes to draw inferences about mechanisms of effect (Dunbar & Klahr, 2012). Indeed, "cause and effect" has been identified as a key "cross-cutting concept" fundamental to both the disciplinary core ideas and science and engineering practices, components of the Next Generation Science Standards (NRC, 2012, 2013).

The possibility that causal reasoning is foundational to scientific literacy is consistent with evidence that this capacity emerges long before children enter school and are formally exposed to science (for reviews, see Gopnik & Wellman, 2012; Muentener & Bonawitz, 2017; Sobel & Legare, 2014). For example, young children—and even infants—can learn cause–effect relations by tracking covariation among events (Gopnik, Sobel, Schulz, & Glymour, 2001; Gweon & Schulz, 2011; Schulz & Gopnik, 2004; Schulz, Gopnik, & Glymour, 2007; Sobel & Kirkham, 2006, 2007). Although covariation does not alone indicate causation, young children behave in accordance with this inference. For example, they generate novel interventions to stop an event (Gopnik et al., 2001) and use indirect evidence to infer causality (Sobel, Tenenbaum, & Gopnik, 2004). Although developmental changes are not often reported (e.g., Gopnik & Schulz, 2004; Gopnik et al., 2001; Kushnir & Gopnik, 2005), children's capacity to draw inferences about unknown or indeterminate causal relations in more complex versions of these tasks improves throughout early childhood (Sobel et al., 2004, 2017).

Young children are not just passive observers of event covariations. When the data are insufficient to determine the causal structure of a system, they actively explore in search of disambiguating information (for reviews, see Gopnik & Wellman, 2012; Muentener & Bonawitz, 2017). Gweon and Schulz (2008), for example, found that 4- and 5-year-olds' play behavior was more variable and exploratory after observing confounded evidence regarding the cause of a novel outcome compared with when they observed unconfounded evidence. Further studies have revealed that children's heightened exploration is not wholly random, incorporating informative interventions consistent with a nascent understanding of basic principles of experimentation (Cook, Goodman, & Schulz, 2011; Lapidow & Walker, 2020; Legare, 2012; Legare, Gelman, & Wellman, 2010).

The evidence reviewed so far regarding precocious causal reasoning skills has focused on tasks that require little to no mechanistic knowledge or understanding. Although this measurement approach can be viewed as advantageous in terms of the specificity of response, it might not reflect the full complexity of causal reasoning. Other work shows that conceptual knowledge of causal mechanisms, functions, and viable transformations emerges in infancy and plays a critical role in early causal reasoning (Callanan et al., 2020; Das Gupta & Bryant, 1989; Gelman, Bullock, & Meck, 1980; Legare, 2012; Legare & Lombrozo, 2014). Legare & Lombrozo, 2014, for example, demonstrated that 3- to 5-year-old children are capable of acquiring new mechanistic understanding after brief exposure to novel systems, especially when encouraged to explain verbally how those systems work prior to exploration. Moreover, conceptual knowledge of this nature appears to constrain the hypothesis space and enable children to align their exploratory interventions with the metamechanisms and goals most relevant to specific domains (Kominsky, Zamm, & Keil, 2018; Legare, Wellman, & Gelman, 2009; Waldmann & Hagmayer, 2013).

Children's reliance on conceptual knowledge of causal mechanisms is particularly evident in the context of reasoning about transformations. For example, after seeing pictures of an intact cup (beginning state) and a broken cup (end state), 3-year-old children can correctly identify a hammer from among several alternatives as the most likely cause (Gelman et al., 1980). By 4 years of age, children expand this capability to more complex inferences that require taking into consideration multiple

transformations (e.g., a recently broken cup that then also becomes wet) and/or less canonical state changes (e.g., a broken flowerpot that is fixed to be whole again) (Das Gupta & Bryant, 1989).

Children also bring their conceptual knowledge to bear when reasoning counterfactually about chains of events. Specifically, when reasoning about how events would have unfolded if things had happened differently in the past (i.e., counterfactual reasoning), children must draw on their knowledge of how those events are causally dependent (Gopnik & Schulz, 2007; Harris, German, & Mills, 1996; Lewis, 1973; Mackie, 1974; Pearl, 2000). In their foundational study of children's counterfactual reasoning capabilities, Harris et al. (1996) told 3- to 5-year-olds a story about a character who left her muddy shoes on when entering the kitchen. Even some of the youngest children tested were able to specify that the floor would not have been dirty if the character had taken off her muddy shoes. Although children's competence in reasoning counterfactually about simple scenarios solidifies by 4 years of age (German & Nichols, 2003; Guajardo & Turley-Ames, 2004; Riggs, Peterson, Robinson, & Mitchell, 1998), success in more complex tasks requiring consideration of more than one causally relevant factor is slower to achieve (Rafetseder, Renate, & Perner, 2010; Rafetseder, Schwitalla, & Perner, 2013).

In the current study, we asked how causal reasoning skills like these relate to emergent scientific literacy. Scientific literacy has been generally characterized as encompassing (a) a core understanding of the nature of science, (b) skills and knowledge regarding scientific practice, and (c) conceptual understanding of key scientific ideas (Jenkins, 1994). These components are reflected in the three key dimensions of current U.S. national standards for science education, namely disciplinary core ideas, science and engineering practices, and crosscutting concepts (NRC, 2012, 2013). Disciplinary core ideas include knowledge of specific content domains such as life and physical sciences. Science and engineering practices include the activities of scientific knowledge acquisition such as asking questions, planning and carrying out investigations, and analyzing and interpreting data. Lastly, crosscutting concepts like understanding cause–effect, patterns, and energy and matter are highlighted as broadly essential to both the core ideas and practices components of scientific literacy.

Because cause–effect understanding is a crosscutting concept in and of itself, all the measures of causal reasoning reviewed above might be expected to contribute broadly to scientific literacy. The specific ways in which each instantiation of causal reasoning contributes, however, might be unique. For example, tracking covariations might be particularly important for scientific practices like monitoring variables during observation and experimentation. Causal inference, in contrast, might be key to drawing conclusions from those observed data and to building complex networks of conceptual knowledge central to disciplinary core ideas in science. Counterfactual reasoning might be fundamental to generating alternative hypotheses to test in the scientific process of experimentation. Indeed, counterfactual reasoning has been described as closely mirroring the process of experimentation in that, at its core, it involves mentally manipulating potential causes and considering their effects on imagined outcomes (see Walker & Nyhout, 2020, for a detailed discussion). Children's spontaneous exploration of confounded evidence in ambiguous causal systems can also be viewed as similar to experimentation. In particular, it could offer opportunities for practicing information-seeking strategies related to the systematic forms of hypothesis testing at the heart of scientific inquiry.

In sum, causal reasoning is theoretically foundational to scientific literacy, and a substantial literature is consistent with that possibility. However, almost no evidence directly addresses this potential relation. One challenge to doing so is lack of clarity on how best to characterize and measure early causal reasoning. As reviewed here, many different tasks have been used to assess this ability in young children, varying substantially in their procedural designs and cognitive requirements. It is unclear whether all these tasks reflect a common underlying causal reasoning construct or whether they represent distinct skills that develop independent of each other. To evaluate the relevance of causal reasoning to scientific literacy, it is essential to first establish whether and how these tasks relate to each other developmentally through early childhood.

In pursuing this preliminary goal in the current study, we built on Bauer & Booth, 2019, who found that children's performance on counterfactual reasoning tasks and their performance on causal inference tasks were related to each other, and to scientific literacy, at 3 years of age, but that tracking covariation patterns in a "blicket detector" task was not related to any of these measures. Although this dissociation is suggestive of divergent causal reasoning constructs, the strength and scope of

conclusions that could be reached from this work was limited by the narrow age range studied. Therefore, we adopted a longitudinal lens, following children annually from 3 to 5 years of age, a period during which significant developmental changes take place in the key domains of interest.

After laying this groundwork, we were able to turn to our primary objective of exploring relations between causal reasoning abilities and early scientific literacy. Based on the literature reviewed above, we hypothesized that causal reasoning skills provide a developmental foundation for subsequent scientific literacy. However, we also considered the possibility of more complex reciprocal effects whereby children's scientific knowledge and inquiry skills support performance on causal reasoning tasks. Cognitive skills were also considered as potential contributors to observed relationships between these constructs.

Method

Participants

The data for the current study originate from a larger longitudinal study, first described in Booth, Shavlik, & Haden, 2020. The initial sample included 153 3-year-olds (81 female; M_{age} = 3.41 years, *SD* = 0.26, range = 3.01–3.92). All children were typically developing and understood English "well" or "very well" as reported by parents. The sample was racially, ethnically, and socioeconomically representative of the southwestern U.S. metropolitan recruitment area (see Table 1). At the second wave of data collection, 120 4-year-olds (64 female; M_{age} = 4.59 years, *SD* = 0.26, range = 3.66–5.09) remained in the study, and at the third wave, 112 5-year-olds (61 female; M_{age} = 5.02 years, *SD* = 0.23, range = 5.02–5.92) remained. Throughout the 3 years of assessment, attrition was primarily due to family relocation or our inability to re-establish contact. Missingness was unrelated to key variables of interest; Little's (1988) test was not significant, suggesting that our data were missing completely at random (MCAR), $\chi^2(704) = 758.38$, $p = .076$.

Procedure

Data for this study were collected in a total of nine sessions, three in each age-specific wave. Each session included three to six tasks that were completed by children in a fixed order over the course of 30 to 60 minutes. In the first session of each wave of data collection, parents also completed (or updated) a demographic survey. The average time window in which the three sessions were completed was 4.89 months (*SD* = 2.53) at 3 years, 1.62 months *(SD* = 0.79) at 4 years, and 1.35 months

Table 1
Participant demographics (in %) across 3 years of data collection.

	3 years (n = 153)	4 years (n = 120)	5 years (n = 112)
Race			
White/Caucasian	73.9	74.2	78.4
Black/African American	13.1	12.5	7.2
Asian/Asian American	2.6	1.7	0.9
Mixed/"other"	10.5	11.7	13.5
Ethnicity			
Non-Hispanic/Latino	69.9	72.5	68.5
Hispanic/Latino	30.1	27.5	31.5
Maternal education			
No more than high school	27.5	20.9	18.0
Technical or associate's degree	6.5	6.7	6.3
Bachelor's degree	38.6	44.2	46.8
Master's degree	18.9	20.0	19.8
Advanced degree	8.5	8.3	9.0

(*SD* = 1.15) at 5 years. All but the first session, which was conducted at a local children's museum, took place in the laboratory. Sessions were audiovisually recorded for offline coding of participant attention and responses and for verification of protocol fidelity. Average attention ratings for retained participants fell between 4 and 5 on a 5-point scale for all measures, indicating generally high engagement.

Measures

Causal reasoning

Five causal reasoning tasks were chosen from a wide array of possibilities based on their potential relevance to scientific literacy (Table 2). Additional criteria for task selection at each age included the absence of ceiling or floor effects (as reported in the literature and/or pilot testing) and relatively quick administration time (commensurate with the attention span of young participants). For measures that were used at more than one age, the basic structure of the tasks remained constant, but the specific contents used (i.e., materials and narratives) were revised to minimize potential retest effects. Minor procedural improvements were also sometimes implemented as described in the task descriptions below.

Tracking cause–effect associations. This task, modeled after Gopnik et al. (2001), measures children's ability to draw correct causal inferences based on patterns of dependent and independent probabilities. Four distinct ("blicket detector") boxes were individually presented to children for consideration. After demonstrating a series of events in which one block failed to activate the block, another successfully did so, and then both did so together, children were asked to help make the box stop by removing one of the blocks. To respond correctly, children needed to keep track of the data presented and draw accurate causal inferences about which block would deactivate the box when removed. Performance was scored as the proportion of correct choices (out of 4 trials). Because children's performance on this task approaches ceiling by 4 years of age, it was administered only at the first (3-year) wave of data collection.

Resolving confounded evidence. This task, modeled after Gweon and Schulz (2008), assesses children's ability to generate interventions that are informative in resolving the structure of ambiguous causal systems that are similar to, but more complex than, the blicket detectors used in the tracking cause–effect associations task. The experimenter first demonstrated how two different colored blocks could be used to activate two different colored lights on a novel apparatus. Possible causal factors (block color, panel color, and left/right placement within panels) were confounded in this demonstration, as illustrated in Figure 1. She then said to children, "Now it's your turn! You can play with it however you want!", and let them play freely for 2 minutes (or less if they stated they were done early).

This task was scored in terms of how quickly (in seconds) children produced an intervention that was informative about the causal relation between block placement and light illumination (i.e., places a block in one of the slots in a way that rules out block color, platform color, or left–right placement as causal). For example, placing the yellow block in the leftmost hole on the gray side would have clarified whether left–right positioning was causally relevant to illuminating the green light. Lower scores

Table 2
Causal reasoning and scientific literacy measures used at each time point.

Construct	Skill	3 years	4 years	5 years
Causal reasoning	Tracking cause–effect associations	√		
	Counterfactual reasoning	√	√	√
	Generating causal inferences	√	√	√
	Causal learning		√	√
	Resolving confounded evidence			√
Scientific literacy	Lens on Science	√	√	√
	Science Learning Assessment			√

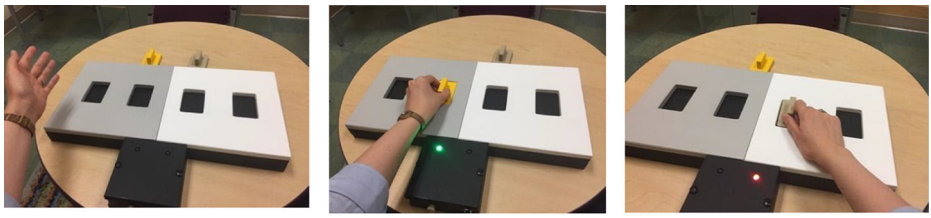

Fig. 1. Confounded evidence task procedure. (A) The starting state. (B,C) The events children observed in the demonstration phase, where the yellow block activated the green light and the brown block activated the red light. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.).

on this task therefore correspond to "better" performance. If children never produced an informative test, they were assigned the maximum possible value of 120 seconds (even if they declined to use the full amount of time available to them). This task had previously been used with children aged 48 to 66 months (Gweon & Schulz, 2008) and therefore could have reasonably been included in our 4-year-old wave of data collection. However, because we did not become aware of the task until after that wave had been completed, it was implemented only with the 5-year-olds.

Counterfactual reasoning. This task, modeled after Guajardo and Turley-Ames (2004), requires children to mentally represent a causal chain of events and how the outcomes would have changed if different events had happened in the past. Children were presented with four vignettes and were asked to reason counterfactually about each situation. For example, one of the four vignettes in the 3-year-old wave described getting mud all over the kitchen floor after playing outside and then returning inside for a drink of water. Children were asked what they could have done differently to avoid getting the floor muddy (e.g., taking shoes off). Each trial of this task was coded dichotomously; if children produced one or more counterfactual scenarios, they received a score of 1 for that trial, and if they produced none, they received a score of 0. Performance was scored as the proportion of trials (out of 4) in which a counterfactual was produced. This task was administered with different vignettes at each wave of data collection.

Generating causal inferences. This task, modeled after Das Gupta and Bryant (1989), measures young children's ability to draw inferences about the causes of event sequences. Children were shown eight pairs of photographs depicting physical transformations (e.g., a broken flowerpot next to a whole flowerpot) and were asked to choose which of four possible instruments (e.g., hammer, light bulb, paintbrush, glue) was responsible. Performance was scored as the proportion of trials (out of 8) answered correctly. Although this task was administered at all waves of data collection, developmental accommodations were made. The 3-year-olds were presented with pictures of relatively familiar items they might have encountered in daily life (e.g., broom, hairbrush), whereas the 4- and 5-year-olds were presented with novel items (e.g., strawberry huller) that put greater demands on their ability to reason broadly about potential mechanisms of effect. To give children a better opportunity to detect these possible mechanisms, real three-dimensional objects were offered to children to select among rather than pictures of the objects.

Causal learning. This task, modeled after Legare & Lombrozo, 2014, measures children's ability to learn about unfamiliar causal mechanisms. In the learning phase of each of 4 trials, children were first introduced to a causal mechanism. For instance, one of the mechanisms was a fan that turned around when a handle, connected through gears, was cranked. The experimenter presented the mechanism and said, "Look, here is a cool machine. I am going to show you how it works, and then you will get a chance to try." The experimenter pointed at the causal parts of the machine and demonstrated how it worked: "Here we have a handle and a fan. Watch how when I turn the handle, the fan goes around." Children were then given a chance to play with the machine for approximately 30 seconds. In the test

phase, the experimenter took the machine away from children's view and removed a causally critical part (e.g., a middle gear). The machine was then re-introduced with the piece missing, and children were asked which of three options would make the machine work again. One option matched the original missing piece in overall appearance but was altered in a way that interfered with its functionality, one differed from the original in appearance but was able to perform the same function, and one differed in both appearance and functionality. If children hesitated to touch the objects, the experimenter encouraged them to point to one (see online supplementary material for a list of the mechanisms presented in each trial together with the corresponding missing pieces). Performance was scored as the total proportion of trials (out of 4) answered correctly. Because piloting revealed near floor performance in 3-year-olds on this task, it was included in only the 4- and 5-year waves of data collection.

Scientific literacy

Few standardized measures of scientific literacy have been developed for the young age range targeted in this study. Among these, even fewer reflect contemporary definitions of scientific literacy that encompass knowledge of both domain-relevant factual information and inquiry practices. However, the Lens on Science (Greenfield, 2015), designed for preschoolers, and the Scientific Learning Assessment (Samarapungavan, Mantzicopoulos, Patrick, & French, 2009; Samarapungavan, Patrick, & Mantzicopoulos, 2011), designed for kindergarteners, are two notable exceptions and therefore were selected for this investigation. More examples of both measures are available at [link: osf.io/z7cgd].

Lens on science

The Lens on Science (Greenfield, 2015) targets all elements of early scientific literacy as currently conceived in the Next Generation Science Standards (NRC, 2012, 2013). Specifically, it goes beyond factual disciplinary knowledge (of earth and space, life, physics and energy, and engineering and technology) and taps into both scientific practices and crosscutting concepts as defined by the NRC's (2012) framework. The test assesses knowledge in the domains of life science (e.g., organisms), physical science (e.g., object properties), engineering/technology (e.g., tools), and earth/space science (e.g., weather) as well as understanding of eight science "process skills" (e.g., observing, questioning, experimenting). The test is administered individually via a touchscreen tablet with instructions and items presented auditorily. Children are adaptively presented with 35 to 40 questions (from a bank of hundreds of possibilities) over the course of approximately 15 minutes. Children select their answers by touching one of the pictured alternatives. For example, one of the life science items shows a picture of a shark and asks children to point to which of three pictured options (a swimming pool, the ocean, and a small fish tank) is where the fish lives. The measure yields a continuous item response theory (IRT) ability score ranging from −3 to 3. A high reliability of .87 and correlation to related measures are reported by Greenfield (2010). The Lens on Science was presented on a 13.9-inch touchscreen Lenovo Yoga laptop.

Science learning Assessment

The Science Learning Assessment (SLA) was developed based on theoretical conceptions regarding science knowledge and learning in kindergarten and is aligned with both state and national standards for early science education (Samarapungavan, Mantzicopoulos, Patrick, & French, 2009; Samarapungavan, Patrick, & Mantzicopoulos, 2011). The test was individually administered and contained 30 weighted questions from two subtests. The Scientific Inquiry Processes subtest taps children's understanding of how science is conducted by asking the children to select among three pictorially and verbally presented answers. For example, children are asked "Which girl asked a science question about a frog?" and must choose between pictures of children asking (a) "What does this frog eat?", (b) "Do you like this frog?", and (c) "Can I call this frog Lilly?" The Life Science Concepts subtest taps children's knowledge of living things and the physical world using multiple-choice and free-response formats. For example, children are asked to point to one of three pictures that "is an insect" and then are asked how they know. The measure was developed for use with children in kindergarten and takes approximately 20 minutes to complete. Internal reliability is reported as .79 by Samarapungavan et al. (2011). SLA-Total scores ranged from 0 (all incorrect) to 38 (all correct).

Cognitive skill

The Early Childhood Cognition Battery of the National Institutes of Health Toolbox (NIH-ECCB), administered on an iPad, includes four adaptive measures of cognitive skill. The Dimensional Change Card Sort Test (Zelazo, 2006) evaluates cognitive flexibility, or the ability to adjust to new tasks and demands. The Flanker Inhibitory Control and Attention Test (Zelazo et al., 2013) assesses children's inhibitory control of visual attention. The Picture Sequence Memory Test assesses children's episodic memory (Dikmen et al., 2014). Finally, the Picture Vocabulary Test (Gershon et al., 2013) measures receptive vocabulary. Based on performance on these four tasks, NIH Toolbox Early Childhood Composite Scores provide a highly reliable assessment of overall general cognitive functioning in young children (Bauer & Zelazo, 2014; Weintraub et al., 2013).

Results

Specifying the structure of early causal reasoning

As a first step in our examination of the early developmental structure of causal reasoning, we calculated descriptive statistics and bivariate correlations for our key measures (see Table 3). For the tasks with multiple measurement time points (i.e., counterfactual reasoning, causal inference, and causal learning), mean scores generally increased with age, consistent with developmental improvement in skill. In addition, with the exception of the tracking associations task, causal reasoning measures correlated with each other at each measurement time point and, in many cases, across waves.

To more rigorously examine the structure of early causal reasoning, we next examined these relations using a longitudinal confirmatory factor analysis (e.g., Schaie, Maitland, Willis, & Intrieri, 1998; Vaillancourt, Brendgen, Boivin, & Tremblay, 2003). This analysis was completed in Mplus using full information maximum likelihood estimation to reduce any biases due to missing data (see Table 4). To accommodate our shift in indicators (see Table 2), we implemented a developmental measurement model (e.g., Bandalos & Raczynski, 2015; Hancock & Buehl, 2008). The common indicators across time points allow for longitudinal analysis of the factor structure, as well as mean levels, of causal reasoning without causing parameter bias (Bandalos & Raczynski, 2015).

Our first model-fitting attempt included all relevant measures of causal reasoning at each of the three waves of data collection (see Table 2). However, this model would not converge. Therefore, we proceeded to test an alternative model that omitted some of the indicators. A theoretical distinction helped to guide this choice; both the blicket detector and confounded evidence tasks are

Table 3
Correlations and descriptive statistics for indicators of causal reasoning across 3, 4, and 5 years.

	1	2	3	4	5	6	7	8	9	10
1. CFR3	–									
2. INFERENCE3	.29*	–								
3. TRACKING3	−.15	−.03	–							
4. CFR4	.47*	.34*	−.08	–						
5. INFERENCE4	.20*	.28*	−.01	.34*	–					
6. LEARN4	.02	.07	.06	.20*	.28*	–				
7. CFR5	.33*	.21*	−.05	.55*	.11	.19	–			
8. INFERENCE5	.18	.15	.04	.24*	.36*	.12	.21*	–		
9. LEARN5	.15	−.08	.10	.27*	.03	−.03	.25*	.07	–	
10. CONFOUND5	−.14	−.10	.06	−.32*	−.22*	.15	−.38*	−.03	−.09	–
M	.34	.56	.64	.64	.69	.28	.73	.68	.52	11.74
SD	.32	.24	.31	.34	.20	.22	.28	.18	.24	18.29

Note. CFR = counterfactual reasoning; INFERENCE = causal inference; TRACKING = tracking.
cause–effect relations (blicket detector); LEARN = causal learning; CONFOUND = resolving confounded evidence. The number following the task abbreviation is the measurement year (3, 4, or 5).
* $p < .05$.

Table 4
Percentages of missing data and reasons for missingness.

Wave	Task	% Missing	Top three reasons for missingness (*n*)[a]
3 years	NIH-ECB		
	PVT	0.65	Attrition (*n* = 33)
	FLANKER	24.84	Attrition (*n* = 40), behavior (*n* = 2), failed training (*n* = 1)
	DCCS	32.68	Attrition (*n* = 42), behavior (*n* = 5), failed training (*n* = 1)
	PICSEQ	39.87	Attrition (*n* = 42), failed training (*n* = 9), behavior (*n* = 3)
	Causal reasoning		
	CFR	24.84	Attrition (*n* = 34), failed training (*n* = 1)
	INF	21.57	Attrition (*n* = 40), behavior (*n* = 2), experimenter error (*n* = 1)
	TRACKING	15.69	Attrition (*n* = 33)
	Scientific literacy		
	LENS	2.61	Attrition (*n* = 33), behavior (*n* = 8)
4 years	NIH-ECB		
	PVT	21.57	Attrition (*n* = 33)
	FLANKER	28.10	Attrition (*n* = 40), behavior (*n* = 2), failed training (*n* = 1)
	DCCS	31.37	Attrition (*n* = 42), behavior (*n* = 5), failed training (*n* = 1)
	PICSEQ	35.29	Attrition (*n* = 42), failed training (*n* = 9), behavior (*n* = 3)
	Causal reasoning		
	CFR	22.88	Attrition (*n* = 34), failed training (*n* = 1)
	INF	28.10	Attrition (*n* = 40), behavior (*n* = 2), experimenter error (*n* = 1)
	LEARN	21.57	Attrition (*n* = 33)
	Scientific literacy		
	LENS	27.45	Attrition (*n* = 33), behavior (*n* = 8)
5 years	NIH-ECB		
	PVT	31.37	Attrition (*n* = 46), behavior (*n* = 1), failed training (*n* = 1)
	FLANKER	32.68	Attrition (*n* = 47), behavior (*n* = 2), failed training (*n* = 1)
	DCCS	35.95	Attrition (*n* = 47), behavior (*n* = 6), failed training (*n* = 1)
	PICSEQ	40.52	Attrition (*n* = 57), behavior (*n* = 2), failed training (*n* = 2)
	Causal reasoning		
	CFR	29.41	Attrition (*n* = 41), technical (*n* = 3), behavior (*n* = 1)
	INF	28.10	Attrition (*n* = 42), behavior (*n* = 1)
	LEARN	40.52	Attrition (*n* = 57), technical (*n* = 4), behavior (*n* = 1)
	CNFND	32.03	Attrition (*n* = 49)
	Scientific literacy		
	LENS	34.64	Attrition (*n* = 41), behavior (*n* = 8), technical (*n* = 4)
	SLA	36.60	Attrition (*n* = 56)

Note. NIH-ECCB = Early Childhood Cognition Battery of National Institutes of Health Toolbox; PVT = Picture Vocabulary Test; FLANKER = Flanker Inhibitory Control and Attention Test; DCCS = Dimensional Change Card Sort Test; PICSEQ = Picture Sequence Memory Test; CFR = counterfactual reasoning; INFERENCE = causal inference; TRACKING = tracking cause–effect relations (blicket detector); LEARN = causal learning; CONFOUND = resolving confounded evidence; LENS = Lens on Science; SLA = Science Learning Assessment.

[a] Total *N* = 153. Attrition within a given wave of data collection was possible given that tasks were administered across multiple sessions.

"knowledge-lean"— that is, unlike the other tasks considered (which rely on real-world cause–effect relations), they employ an arbitrary association between stimulus and response. Their use at just one of the three time points further set them apart from the other causal reasoning variables considered. After omitting these two variables, the remaining model included only counterfactual reasoning, causal inference, and causal learning as indicators of a causal reasoning latent variable.

This model fit the data well, $\chi^2(17) = 24.80$, $p = .09$, root mean square error of approximation (RMSEA) = .06 (90% confidence interval (CI) [.00, .10]), comparative fit index (CFI) = .92, standardized root mean squared residual (SRMR) =.07. At all three time points, the indicators were statistically significant ($p < .05$; see Table 5 for parameter estimates). The test–retest correlations (i.e., stability coefficients) for the causal reasoning construct all were strong and statistically significant ($.84_{ages3,4}$, $.92_{ages4,5}$, and $.75_{ages3,5}$, $p < .05$), with the lowest value not surprisingly being between the most distal measurement time points at 3 and 5 years of age.

Table 5
Parameter estimates for the developmental measurement model across time.

Indicator	PC_{age3}	U_{age3}	PC_{age4}	U_{age4}	PC_{age5}	U_{age5}
Counterfactual reasoning	.625	.610	.892	.204	.656	.570
Causal inference	.450	.798	.389	.848	.321	.897
Causal learning			.214	.954	.302	.909

Note. PC, pattern coefficient; U, uniqueness. Completely standardized solutions.

Table 6
Developmental measurement model fit indices for invariance.

Model	Equality constraints (over time)	χ^2	*df*	*P*	RMSEA	90% CI		CFI	SRMR
						LL	UL		
Configural	None	24.80	17	.09	.06	.00	.10	.92	.07
Metric invariance	Factor loadings	29.03	20	.08	.06	.00	.10	.91	.08
Scalar invariance	Intercepts	69.19	23	<.01	.12	.09	.15	.53	.14
Partial scalar invariance	**CFR and INF intercepts only (LEARN released)**	**30.79**	**22**	**.10**	**.05**	**.00**	**.09**	**.91**	**.08**
Residual invariance	Error (residual) variances	292.95	28	<.01	.26	.23	.28	.00	1.34

Note. CI = confidence interval; LL = lower level; UL = upper level; CFR = counterfactual reasoning; INF = causal inference; LEARN = causal learning. To gauge model fit, the following fit indices were used: chi-square statistic (χ^2), comparative fit index (CFI), root mean square error of approximation (RMSEA) and corresponding CI, and standardized root mean squared residual (SRMR). In model fitting, a *p* value less than .05 for chi-square indicates poor model fit. In addition, a CFI value greater than or equal to .90, an RMSEA value less than .08 (preferably < .05), and an SRMR value at or below .05 indicate good model fit. Bold indicates the best-fitting model.

To examine developmental change in the latent causal reasoning variable, it was necessary to first establish measurement invariance. In other words, before we can test whether the scores on our latent variable differ across time, we need to first establish that we measured the latent variable in the same way across the developmental window of interest. If model invariance is present, invariance of the latent mean structures can be examined (Hancock, 1997; Little, 1997) by first assessing the invariance of the measurement intercepts (i.e., a person's predicted score). If intercepts are invariant, latent mean differences can be compared accurately. Otherwise, in the presence of noninvariant intercepts, expected score differences may reflect differences across time on our variable of interest (causal reasoning), systematic measurement bias, or a combination of both.

Assessing invariance is accomplished by sequentially placing equality constraints on model parameters in a series of model comparisons (see Table 6). Based on the guidelines for strict, strong, and weak forms of factorial invariance presented by Meredith (1993), we achieved strong measurement invariance and therefore proceeded with our mean comparisons. Latent mean estimates (i.e., participants' average "causal reasoning" levels) increased from 3 to 4 years of age ($z = 15.75$, $p < .05$, $d = 1.50$) and from 4 to 5 years of age ($z = 9.52$, $p < .05$, $d = 1.34$). In other words, scores were significantly higher at later time points, with the greatest change seen from 3 to 4 years.

Testing relations between early causal reasoning and scientific literacy

Recall that our primary objective was to investigate the relation between causal reasoning abilities and early scientific literacy developmentally through the preschool years. The ideal approach to doing so would have been to build on this model of the structure of causal reasoning, using latent variable structural equation modeling (SEM) to add an analogous scientific literacy construct at each measurement time point with paths indicating concurrent and predictive relations between both constructs. However, given only partial invariance across time points, and because we had only one indicator of scientific literacy at 3 and 4 years of age (thereby falling short of the recommended three measures per latent construct), we proceeded with the following alternative modeling strategy.

Table 7
Correlations among causal reasoning and scientific literacy composites at each time point.

		Causal reasoning			Scientific literacy[a]		
		3 years	4 years	5 years	3 years	4 years	5 years
Causal reasoning	3 years	–					
	4 years	.44**	–				
	5 years	.31**	.49**	–			
Scientific literacy	3 years	.55**	.39**	.33**	–		
	4 years	.46**	.52**	.47**	.68**	–	
	5 years	.49**	.52**	.55**	.52**	.67**	–
Excluded measures	TRACKING	−.13	−.03	.04	−.05	.11	−.09
	CONFOUND	−.14	−.19	−.22*	−.22*	−.18	−.30**
	M	0.42	0.52	0.63	0.36	1.57	0.00
	SD	0.24	0.19	0.17	1.00	1.00	0.86

Note. TRACKING = tracking cause–effect relations (blicket detector); CONFOUND = resolving confounded evidence. All statistics in this table are the pooled results from multiply imputed datasets.
[a] Scientific literacy is a standardized composite score at 5 years but is a single score at 3 and 4 years.
* p <.05.
** p <.01.

Guided by the results of our confirmatory factor analysis, we calculated composite measures of causal reasoning at each measurement time point based on the average of the scores of available indicators. Although composites were not required for 3- and 4-year scientific literacy (given that only the single Lens on Science measure was available), a standardized average of the SLA and Lens on Science

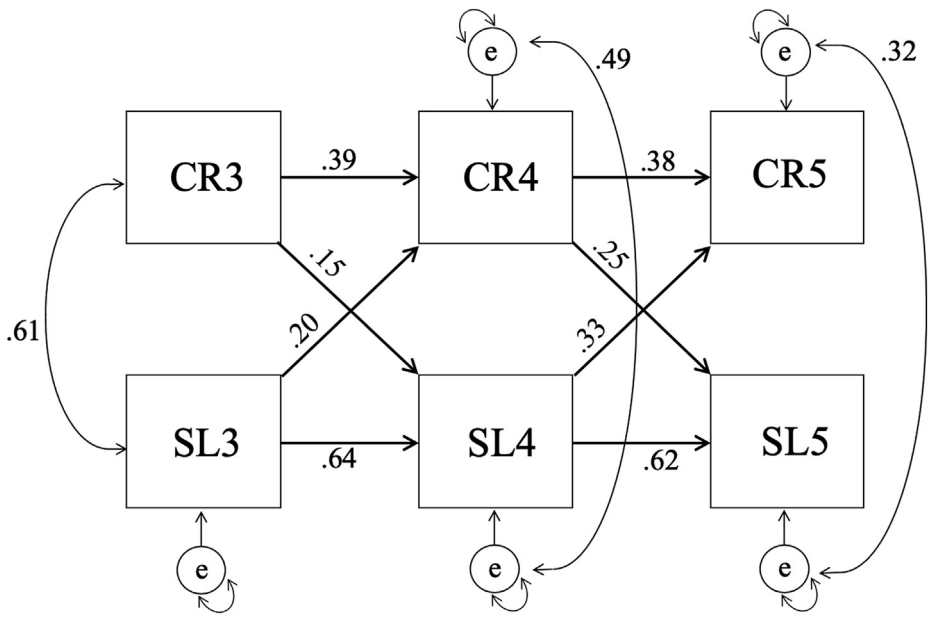

Fig. 2. Cross-lagged panel model for causal reasoning and scientific literacy during the preschool years. CR = causal reasoning composite measure; SL = scientific literacy; e = error. 3 = score at 3 years of age; 4 = score at 4 years of age; 5 = score at 5 years of age. For ease of interpretation, all values are fully standardized. Curved, double-headed arrows indicate correlations, whereas straight arrows indicate regression weights. All *ps* < .05.

Table 8
Follow-up regression analyses examining the role of cognitive skill.

Model	$F(2, 149)$*	Outcome	Predictor	Unstandardized		Standardized			
				B	*SE*	*β*	*t*	*p*	*sr*[a]
1	15.51	CR (4 years)	(Constant)	0.22	0.19		1.17	.244	
			COG (3 years)	0.00	0.00	.16	1.51	.132	.14
			SL (3 years)	0.06	0.02	.31	2.98	.003	.27
2	21.91	CR (5 years)	(Constant)	0.50	0.17		2.97	.003	
			COG (4 years)	0.00	0.02	.00	0.00	.997	.00
			SL (4 years)	0.08	0.02	.47	4.45	<.001	.41
3	37.57	SL (4 years)	(Constant)	−2.50	0.81		−3.10	.002	
			COG (3 years)	0.03	0.01	.38	4.14	<.001	.35
			CR (3 years)	1.31	0.38	.31	4.44	<.001	.29
4	40.87	SL (5 years)	(Constant)	−3.47	0.74		−4.70	<.001	
			COG (4 years)	0.02	0.01	.29	3.17	.002	.28
			CR (4 years)	1.91	0.42	.43	4.54	<.001	.41

Note. CR, causal reasoning composite score; SL, scientific literacy composite score; COG, cognitive skill composite score.

[a] A semipartial correlation (*sr*) indicates the unique relation between a predictor and the outcome variable. In other words, it is the variance in the outcome explained *only* by that predictor (i.e., it does not include variance explained by other predictors, as is the case with zero-order and partial correlations).

* All *p*s <.001.

scores was calculated for the 5-year wave. Several of our participants had data for some, but not all, of the components of these composite variables (see Table 4). Following recommendations by Gottschall, West, and Enders (2012), we therefore created composite scores based on 100 imputations to take full advantage of all available item-level information. Correlations between these composite measures are presented in Table 7.

Using these composite scores, we created a cross-lagged panel model (e.g., Selig & Little, 2012) to test how the variables related to each other over time (see Figure 2) and to specifically test our hypothesis that early causal reasoning would predict subsequent scientific literacy. The resulting model fit the data well, $\chi^2(4) = 8.89$, $p = .064$, RMSEA = .089 (90% CI [.000, .170]), CFI = .989, SRMR = .023. Consistent with our confirmatory factor analysis, causal reasoning at each age predicted causal reasoning scores for the following year; scores at 3 years predicted levels at 4 years ($B = 0.32$, $SE = 0.07$, $p < .001$), and levels at 4 years predicted those at 5 years ($B = 0.32$, $SE = 0.07$, $p < .001$). Similarly, scientific literacy at each age predicted scores for the following year, with scientific literacy at 3 years predicting levels at 4 years ($B = 0.59$, $SE = 0.06$, $p < .001$), which in turn predicted levels at 5 years ($B = 0.51$, $SE = 0.05$, $p < .001$).

Moreover, and of greatest relevance to our core hypothesis, children's causal reasoning skill predicted subsequent scientific literacy. Causal reasoning at 3 years predicted scientific literacy at 4 years ($B = 0.65$, $SE = 0.29$, $p = .027$), and causal reasoning at 4 years predicted scientific literacy at 5 years ($B = 1.07$, $SE = 0.27$, $p < .001$). The reciprocal relation was also evident, with scientific literacy reliably predicting subsequent causal reasoning. Scientific literacy at 3 years predicted causal reasoning at 4 years ($B = 0.04$, $SE = 0.02$, $p = .022$), and scientific literacy at 4 years predicted causal reasoning at 5 years ($B = 0.05$, $SE = 0.01$, $p < .001$). See Fig. 2 for standardized parameter estimates.

Controlling for broad cognitive skills

To rule out the possibility that the observed cross-lagged relationships between causal reasoning and scientific literacy might be spurious due to their common reliance on broad cognitive skills, we ran follow-up regression analyses including the NIH-ECCB as an additional predictor. All four relations between causal reasoning and scientific literacy held with semipartial (part) correlations, revealing that the hypothesized predictor independently accounted for unique variance in the outcome variable of interest over and above cognitive skill (see Table 8).

Discussion

We began this work with the primary goal of furthering our knowledge of how early causal reasoning skills intersect with emergent scientific literacy. As a preliminary step toward this goal, we first examined the structure of causal reasoning across frequently studied tasks in young children. A longitudinal confirmatory factor analysis revealed that some of our measures of causal reasoning (causal inference, counterfactual reasoning, and causal learning) were strong indicators of a common factor (i.e., causal reasoning) at each age tested. Moreover, this structure was stable across the early childhood years. Children with higher causal reasoning ability relative to their peers at 3 years of age were also likely to have relatively high causal reasoning ability at 4 and 5 years.

Two of our measures, however, did not fit well into this final model. The first of these, tracking cause–effect associations (measured at 3 years only), failed to correlate with any other contemporaneous measure (Bauer & Booth, 2019) or subsequent measure of causal reasoning (see Table 3). The second task, resolving confounded evidence, was measured at 5 years only. Although performance on this task correlated with causal inferencing at 4 years and counterfactual reasoning at 4 and 5 years, adding it to the model resulted in worse fit. One notable difference between the measures that coalesced in the model and those that did not is their reliance on understanding meaningful causal mechanisms (see Bauer & Booth, 2019; Tolmie, Ghazali, & Morris, 2016). Specifically, generating causal inferences, counterfactual reasoning, and learning causal mechanisms all require children to understand the precisely sequenced chain of specific actions or events that are responsible for a change in state. In contrast, the causal mechanisms in the resolving confounded evidence and tracking

associations blicket detector tasks are arbitrary and opaque, requiring no understanding of how they work. The "mechanism-free" tasks failed to load onto the same construct as the "mechanism-depen dent" tasks, suggesting that these might represent two distinct forms of causal reasoning (Buchanan & Sobel, 2011). The mechanism-free tracking associations and confounded evidence tasks also failed to correlate with each other. This might be due in part to the disparate ages at which these tasks were administered, but it also might be because the tasks do not tap wholly overlapping skills and therefore do not cohere into a secondary form of causal reasoning. Future research considering a broader range of measures tapping these different potential forms of causal reasoning will be important to confirm whether these are replicable and meaningful dissociations.

Having clarified the structure of causal reasoning in the available data, we proceeded with our primary goal of evaluating whether and how this construct might be related to the development of scientific literacy. Our cross-lagged panel model revealed relations between causal reasoning and scientific literacy at each age tested. More important, temporal precedence consistent with our prediction was established, such that causal reasoning predicted scientific literacy in each *subsequent* year. To our knowledge, this study provides some of the first evidence for the widely held supposition that causal reasoning is developmentally foundational to scientific literacy in early childhood (Gopnik, 2012; Legare, 2014; Sobel, Erb, Tassin, & Weisberg, 2017; Tolmie, Ghazali, & Morris, 2016).

Notably, the reciprocal relations also held in our model, such that scientific literacy predicted subsequent causal reasoning at each measurement wave as well. This finding is consistent with a more complex relation between causal reasoning and scientific literacy in which bidirectional influences are mutually reinforcing. Given the aforementioned dependence of our measures of causal reasoning on metamechanistic and conceptual knowledge, it is perhaps not surprising that scientific literacy (which encompasses this type of knowledge) predicts performance on these tasks. It is also possible that a stronger grasp of scientific inquiry processes facilitates performance on these tasks by helping to structure cognitive comparison of possible mechanisms of effect. This might be especially true for counterfactual reasoning given how similar the requisite processes of mental manipulation of potential causes and evaluation of imagined outcomes are to those involved in experimentation more broadly speaking (Danovitch, Mills, Duncan, Williams, & Girouard, 2021; Walker & Nyhout, 2020).

These relationships could not be attributed solely to shared reliance of the key constructs of broad cognitive capacities like executive function and vocabulary. When scores on the NIH-ECCB were considered as predictors in follow-up regression analyses, all four cross-lagged relations between causal reasoning and scientific literacy held. Of course, this does not mean that cognitive capacities are irrelevant to understanding the development of scientific literacy. Indeed, in three of the four relations probed, NIH-ECCB scores explained a significant amount of unique variance (see Table 8). And although our behavioral observations revealed overall high levels of attention, it is difficult to wholly rule out the possibility that engagement levels contributed to the observed pattern of results. It also remains possible that other cognitive skills or aspects of conceptual knowledge not tested here contribute in important ways to the development of scientific literacy. For example, theory of mind has been shown to predict later experimentation skills (Piekny, Grube, & Maehler, 2013), and metacognition has been highlighted as foundational to scientific thinking (Kuhn, 2011). Further study, perhaps optimally using neural or other physiological measures, will be necessary to further elucidate these complex developmental relations.

Although the current findings add considerably to our understanding of intersections between the development of causal reasoning and scientific literacy, they also highlight several additional directions in need of further investigation. In particular, much more work is needed to developmentally specify the structure of causal reasoning. We were limited in the number of tasks that we could include in this investigation, and replication with a wider range of selections is essential. Extending the ages studies in both directions will also be of value. A number of tests of infant causal perception and reasoning have been developed (Luchkina, Sommerville, & Sobel, 2018; Newman, Choi, Wynn, & Scholl, 2008; Walker & Gopnik, 2014) that might be successfully adapted for use in an individual difference developmental framework like the one used here. Moreover, tests of more sophisticated scientific inquiry skills can be used with older children that might plausibly act as intermediaries in bridging early causal reasoning skills with broad standardized measures of scientific literacy. Toward these goals, we continue to track children enrolled in the current study and have added tasks to

subsequent waves of data collection that more directly assess children's hypothesis generation, explanation evaluation, and testing strategies (e.g., Koerber & Osterhaus, 2019; McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Mills, Danovitch, Rowles, & Campbell, 2017; Piekny & Maehler, 2013; van Schijndel, Jansen, & Raijmakers, 2018). This will offer the added benefit of more explicitly connecting this investigation to research documenting the converging development of causal and scientific reasoning (Köksal, Sodian, & Legare, 2021, Köksal-Tuncer & Sodian, 2018; Sodian, 2018) and the predictors of scientific reasoning in older children (e.g., Mayer, Sodian, Koerber, & Schwippert, 2014; Osterhaus, Koerber, & Sodian, 2017).

In furthering these investigations, it will be important to not lose sight of the tasks that did not fit well into the measurement model specified here. The failure of our tracking associations (blicket detector) task to correlate with other measures of causal reasoning or scientific literacy was unexpected in light of the dominance of this type of task in the causal reasoning literature (but see Bauer & Booth, 2019). Nevertheless, many versions of the blicket detector task have been developed with varying complexity. We included only the simplest of these in the current investigation, and it remains to be seen whether more complex versions will align more closely with other causal reasoning measures and be similarly predictive of emergent scientific literacy.

Our own data speculatively suggest that more sophisticated mechanism-free tasks might be more related to scientific development. Specifically, not only did the confounded evidence task correlate with some of our other measures of causal reasoning at 5 years of age, but it also correlated with scientific literacy at two of three measurement time points (see Table 7). This might be because the confounded evidence task not only requires tracking cause-and-effect relations but also engages key components of the scientific inquiry process, including (a) noticing that a given causal relation is ambiguous and (b) producing an informative disambiguating intervention (Zimmerman, 2007). Moreover, sensitivity to information gaps has been established as a first step in question-asking behavior (Ronfard, Zambrana, Hermansen, & Kelemen, 2018), which is a critical medium for acquiring early science knowledge both in everyday life through parents (e.g., Callanan & Jipson, 2001; Harris & Koenig, 2006) and in kindergarten science activities (e.g., Samarapungavan et al., 2011). That said, the measurement characteristics of the confounded evidence task were unique relative to the other causal reasoning tasks as a result of its reliance on a single continuous record of response time as opposed to cumulative accuracy across a small number of discrete trials. It is possible that this contributed to its observed pattern of relation to other measures independent of the speculative possibilities outlined above. In any case, clarifying these points will require targeted replications with more tasks specifically chosen to represent different potential types of causal reasoning and component skills.

In sum, although relations between early causal reasoning and scientific literacy are clearly complex, the current findings support the view that the former provides a developmental foundation for the latter. In particular, the three measures of causal reasoning included in our final model all required reasoning about causal mechanisms. This may be fundamental to representing more complex scientific phenomena as well as engaging in scientific inquiry to probe various explanations thereof. This is an important finding because many children, especially those from socioeconomically disadvantaged backgrounds, struggle in tasks such as experimentation, interpreting data, and understanding the nature of science (Bustamante, White, & Greenfield, 2016; Curran & Kellogg, 2016; Greenfield et al., 2009). The few studies that have investigated this longitudinally point to stable and potentially widening individual differences that have significant impact on success throughout the elementary and high school years (Byrnes & Miller, 2007; Morgan, Farkas, Hillemeier, & Maczuga, 2016). The current evidence provides new insight into a key precursor to scientific literacy. Ideally informed by additional replications focused more narrowly on at-risk populations, future research might capitalize on this insight in developing early childhood educational curricula and interventions.

Data availability

More information about this project can be found at osf.io/z7cgd.

Acknowledgments

This work was graciously funded by the National Science Foundation (1535102) awarded to Amy E. Booth and Catherine A. Haden. Special thanks go to Kimberly Brenneman, Maureen Callanan, Daryl Greenfield, Robin Gose, and Ala Samarapungavan for their advisory role in this research. We are also grateful to the children and families who generously contributed their time to this project.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2022.105509.

References

Bandalos, D. L., & Raczynski, K. A. (2015). Capturing family–school partnership constructs over time: Creating developmental measurement models. In S. M. Sheridan & E. Moorman Kim (Eds.). *Foundational aspects of family–school partnership research* (Vol. 1, pp. 77–103). New York: Springer.

Bauer, J. R., & Booth, A. E. (2019). Exploring potential cognitive foundations of scientific literacy in preschoolers: Causal reasoning and executive function. *Early Childhood Research Quarterly, 46*(1), 275–284. https://doi.org/10.1016/j.ecresq.2018.09.007.

Bauer, P. J., & Zelazo, P. D. (2014). The National Institutes of Health Toolbox for the assessment of neurological and behavioral function: A tool for developmental science. *Child Development Perspectives, 8*, 119–124.

Booth, A. E., Shavlik, M., & Haden, C. A. (2020). Parents' causal talk: Links to children's causal stance and emerging scientific literacy. *Developmental Psychology, 56*(11), 2055–2064. https://doi.org/10.1037/dev0001108.

Buchanan, D. W., & Sobel, D. M. (2011). Mechanism-based causal reasoning in young children. *Child Development, 82*, 2053–2066.

Bulunuz, M. (2013). Teaching science through play in kindergarten: Does integrated play and science instruction build understanding? *European Early Childhood Education Research Journal, 21*, 226–249.

Bustamante, A. S., White, L. J., & Greenfield, D. B. (2016). Approaches to learning and school readiness in Head Start: Applications to preschool science. *Learning and Individual Differences, 56*, 112–118.

Byrnes, J. P., & Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity-propensity analysis. *Contemporary Educational Psychology, 32*, 599–629.

Callanan, M. A., & Jipson, J. L. (2001). Explanatory conversations and young children's developing scientific literacy. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 21–49). Mahwah, NJ: Lawrence Erlbaum.

Callanan, M. A., Legare, C. H., Sobel, D. M., Jaeger, G. J., Letourneau, S., McHugh, S. R., & Rubio, E. (2020). Exploration, explanation, and parent–child interaction in museums. *Monographs of the Society for Research in Child Development, 85*(1), 7–137. https://doi.org/10.1111/mono.12412.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*, 341–349.

Curran, F. C., & Kellogg, A. T. (2016). Understanding science achievement gaps by race/ethnicity and gender in kindergarten and first grade. *Educational Researcher, 45*, 273–282.

Danovitch, J. H., Mills, C. M., Duncan, R. G., Williams, A. J., & Girouard, L. N. (2021). Developmental changes in children's recognition of the relevance of evidence to causal explanations. *Cognitive Development, 58* 101017.

Das Gupta, P., & Bryant, P. E. (1989). Young children's causal inferences. *Child Development, 60*, 1138–1146.

Dikmen, S. S., Bauer, P. J., Weintraub, S., Mungas, D., Slotkin, J., Beaumont, J. L., ... Heaton, R. K. (2014). Measuring episodic memory across the lifespan: NIH Toolbox picture sequence memory test. *Journal of the International Neuropsychological Society, 20*, 611–619.

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701–718). New York: Cambridge University Press.

Gelman, R., Bullock, M., & Meck, E. (1980). Preschoolers' understanding of simple object transformations. *Child Development, 51*, 691–699.

German, T. P., & Nichols, S. (2003). Children's counterfactual inferences about long and short causal chains. *Developmental Science, 6*, 514–523.

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH Toolbox for assessment of neurological and behavioral function. *Neurology, 80*(Suppl. 3), S2–S6.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science, 337*, 1623–1627.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences, 8*, 371–377.

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

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 theory. *Psychological Bulletin, 138*, 1085–1108.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*, 1–25.
Greenfield, D. B. (2010). *June). Please touch! A computer adaptive approach for assessing early science*. National Harbor, MD: Institute for Education Science.
Greenfield, D. B. (2015). Assessment in early childhood science education. In K. Cabe Trundle & M. Saçkes (Eds.), *Research in early childhood science education* (pp. 353–380). Dordrecht, Netherlands: Springer Netherlands.
Greenfield, D. B., Jirout, J., Dominguez, X., Greenberg, A., Maier, M., & Fuccillo, J. (2009). Science in the preschool classroom: A programmatic research agenda to improve science readiness. *Early Education and Development, 20*, 238–264.
Guajardo, N. R., & Turley-Ames, K. J. (2004). Preschoolers' generation of different types of counterfactual statements and theory of mind understanding. *Cognitive Development, 19*, 53–80.
Gweon, H., & Schulz, L. E. (2008). In *Stretching to learn: Ambiguous evidence and variability in preschoolers' exploratory play* (pp. 570–574). Austin, TX: Cognitive Science Society.
Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science, 332*, 1524.
Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30*(2), 91–105. https://doi.org/10.1080/07481756.1997.12068926.
Hancock, G. R., & Buehl, M. M. (2008). Second-order latent growth models with shifting indicators. *Journal of Modern Applied Statistical Methods, 7*, 39–55.
Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition, 61*, 233–259.
Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development, 77*, 505–524.
Jenkins, E. W. (1994). Scientific literacy. *In T. Husen & T. N. Postlethwaite (Eds.), The International Encyclopedia of Education*, (2nd ed., 9) (pp. 5345–5350). London: Pergamon.
Koerber, S., & Osterhaus, C. (2019). Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *Journal of Cognition and Development, 20*, 510–533.
Köksal, Ö., Sodian, B., & Legare, C. H. (2021). Young children's metacognitive awareness of confounded evidence. *Journal of Experimental Child Psychology, 205* 105080.
Köksal-Tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cognitive Development, 48*, 135–145.
Kominsky, J. F., Zamm, A. P., & Keil, F. C. (2018). Knowing when help is needed: A developing sense of causal complexity. *Cognitive Science, 42*, 491–523.
Kuhn, D. (2011). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 497–523). Hoboken, NJ: Wiley-Blackwell.
Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science, 16*, 678–683.
Lapidow, E., & Walker, C. M. (2020). Informative experimentation in intuitive science: Children select and learn from their own causal interventions. *Cognition, 201* 104315.
Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development, 83*(1), 173–185. https://doi.org/10.1111/j.1467-8624.2011.01691.x.
Legare, C. H. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives, 8*(2), 101–106. https://doi.org/10.1111/cdep.12070.
Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development, 81*(3), 929–944. https://doi.org/10.1111/j.1467-8624.2010.01443.x.
Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology, 126*(0), 198–212. https://doi.org/10.1016/j.jecp.2014.03.001.
Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology, 58*(2), 177–194. https://doi.org/10.1016/j.cogpsych.2008.06.002.
Lewis, D. (1973). Counterfactuals and comparative possibility. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.). *Ifs* (Vol. 15, pp. 57–85). New York: Springer.
Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.
Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53–76. https://doi.org/10.1207/s15327906mbr3201_3.
Luchkina, E., Sommerville, J. A., & Sobel, D. M. (2018). More than just making it go: Toddlers effectively integrate causal efficacy and intentionality in selecting an appropriate causal intervention. *Cognitive Development, 45*, 48–56.
Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, UK: Clarendon.
Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2013). Science literacy in school and home contexts: Kindergarteners' science achievement and motivation. *Cognition and Instruction, 31*, 62–119.
Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43–55.
McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology, 141*, 1–22.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.
Mills, C. M., Danovitch, J. H., Rowles, S. P., & Campbell, I. L. (2017). Children's success at detecting circular explanations and their interest in future learning. *Psychonomic Bulletin & Review, 24*, 1465–1477.
Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educational Researcher, 45*, 18–35.
Muentener, P., & Bonawitz, E. (2017). The development of causal reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 677–698). Oxford, UK: Oxford University Press.
National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

National Research Council (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.
Newman, G. E., Choi, H., Wynn, K., & Scholl, B. J. (2008). The origins of causal perception: Evidence from postdictive processing in infancy. *Cognitive Psychology, 57*, 262–291.
Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*, 450–462.
Pearl, J. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association, 95*, 428–431.
Piekny, J., Grube, D., & Maehler, C. (2013). The relation between preschool children's false-belief understanding and domain-general experimentation skills. *Metacognition and Learning, 8*, 103–119.
Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology, 31*, 153–179.
Rafetseder, E., Renate, C.-V., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child Development, 81*, 376–389.
Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology, 114*, 389–404.
Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development, 13*, 73–90.
Ronfard, S., Zambrana, I. M., Hermansen, T. K., & Kelemen, D. (2018). Question-asking in childhood: A review of the literature and a framework for understanding its development. *Developmental Review, 49*, 101–120.
Samarapungavan, A., Mantzicopoulos, P., Patrick, H., & French, B. (2009). The development and validation of the science learning assessment (SLA): A measure of kindergarten science learning. *Journal of Advanced Academics, 20*(3), 502–535. https://doi.org/10.1177/1932202x0902000306.
Samarapungavan, A., Patrick, H., & Mantzicopoulos, P. (2011). What kindergarten students learn in inquiry-based science classrooms. *Cognition and Instruction, 29*, 416–470.
Schaie, K. W., Maitland, S. B., Willis, S. L., & Intrieri, R. C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. *Psychology and Aging, 13*, 8–20.
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.
Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265–278). New York: Guilford.
Sobel, D. M., Erb, C. D., Tassin, T., & Weisberg, D. S. (2017). The development of diagnostic inference about uncertain causes. *Journal of Cognition and Development, 18*, 556–576.
Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology, 42*, 1103–1115.
Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science, 10*, 286–306.
Sobel, D. M., & Legare, C. H. (2014). Causal learning in children. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(4), 413–427. https://doi.org/10.1002/wcs.1291.
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.
Sodian, B. (2018). The development of scientific thinking in preschool and elementary school age: A conceptual model. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 227–250). New York: Routledge.
Tolmie, A. K., Ghazali, Z., & Morris, S. (2016). Children's science learning: A core skills approach. *British Journal of Educational Psychology, 86*, 481–497.
Vaillancourt, T., Brendgen, M., Boivin, M., & Tremblay, R. E. (2003). A longitudinal confirmatory factor analysis of indirect and physical aggression: Evidence of two factors over time? *Child Development, 74*, 1628–1638.
van Schijndel, T. J., Jansen, B. R., & Raijmakers, M. E. (2018). Do individual differences in children's curiosity relate to their inquiry-based learning? *International Journal of Science Education, 40*, 996–1015.
Waldmann, M. R., & Hagmayer, Y. (2013). Categories and causality: The neglected direction. *Cognitive Psychology, 53*, 27–58.
Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science, 25*, 161–169.
Walker, C. M., & Nyhout, A. (2020). Asking "why?" and "what if?": The influence of questions on children's inferences. In L. Butler, S. Ronfard, & K. Coriveau (Eds.), *The questioning child: Insights from psychology and education* (pp. 252–280). Cambridge, UK: Cambridge University Press.
Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., ... Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology, 80*(Suppl. 3), S54–S64.
Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols, 1*, 297–301.
Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development, 78*(4), 16–33.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.