UC Merced

Title
Clarifying the Causal Logic of a Classic Control of Variables Task

Permalink
https://escholarship.org/uc/item/3bx4z0kp

Journal

Authors
Lapidow, Elizabeth
Walker, Caren M.

Publication Date
2022

Peer reviewed

Clarifying the Causal Logic of a Classic Control of Variables Task

Elizabeth Lapidow (elapidow@ucsd.edu)
Department of Psychology, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093 USA

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

Abstract

Self-directed learners are often described as 'intuitive scientists', yet they also tend to struggle in assessments of their scientific reasoning. We investigate a novel explanation for this apparent gap between formal and informal scientific inquiry. Specifically, we consider whether learners' documented failure to correctly apply the control of variables strategy might stem from a mismatch between their causal intuitions and task presentation. Children (7- and 9-year-olds) and adults were tested on a version of a traditional multivariate reasoning task (Tschirgi, 1980) that we modified to clarify ambiguous elements of the causal logic. A significant majority of participants in all age groups selected informative experiments on this modified task, avoiding confounded actions with positive tangible outcomes. This finding contrasts with the longstanding claim that learners do not correctly employ control of variables without extensive training and suggests that self-directed scientific inquiry may be intuitively suited to support causal learning goals.

Keywords: cognitive development; causal learning; scientific reasoning; control of variables; decision-making; exploration

Introduction

Self-directed learning, in which learners choose what to do to expand upon their existing knowledge, requires two interconnected abilities: *inquiry*, acting to generate informative evidence, and *inference*, drawing rational conclusions from evidence in coordination with prior knowledge. Despite the importance of these abilities, however, there is longstanding disagreement about the competence of self-directed learners' inquiry and inference skills.

Below, we briefly review the claims made by each side of this debate before describing a possible resolution grounded in theories of causal reasoning. We use this novel theory to reexamine a prominent example of the disconnect between formal and informal inquiry: experimentation using the control of variables strategy (CVS). Examined closely, several standard elements of CVS assessments involve ambiguous and counterintuitive causal logic. The current study investigates whether learners' poor performance on a well-known CVS task (Croker & Buchanan, 2011; Tschirgi, 1980) is improved by clarifying this causal logic.

The Gap Between Exploring and Experimenting

The literature documenting inquiry and inference behavior presents two very different accounts of the self-directed learner. On one side, cognitive development research characterizes learners as 'intuitive scientists' who are naturally motivated to seek informative evidence and rationally integrate it with their existing knowledge (Gopnik & Wellman, 2012; Schulz, 2012). Indeed, studies examining *exploration* suggest that children and adults have a spontaneous preference for inquiry that is likely to improve their current knowledge (e.g., Lapidow et al., 2022; Liquin & Lombrozo, 2020; Schulz & Bonawitz, 2007). On the other side, research examining the development of scientific reasoning finds learners' spontaneous and untrained behavior to be highly error-prone (Zimmerman, 2007; Zimmerman & Klahr, 2018). In studies of *experimentation*, children and adults struggle with inquiry, producing confounded and confirmatory actions, and consistently privileging tangible outcomes over information (e.g., Kuhn, 2007; Siler & Klahr, 2012).

There have been considerable efforts on both sides to 'bridge the gap' between these two accounts of self-directed learning (e.g., Kuhn, 2012; Osterhaus et al., 2021; Shtulman & Walker, 2020). However, to date, neither literature has produced an explanation that accounts for the full scope of documented behavior. Scientific reasoning researchers have suggested that cognitive development tasks test an unconscious coordination of theory and evidence that falls short of a genuine understanding of experimental logic (e.g., Kuhn, 2002). Conversely, researchers in cognitive development argue that learners understand this logic, but struggle to meet the extraneous coordination and fluency demands of employing it in scientific reasoning tasks (e.g., Bullock & Ziegler, 1999). Neither of these approaches is wholly satisfying; each side aims to explain away findings that are inconsistent with their own characterization of self-directed learning, rather than to revise this characterization to account for those findings.

Here, we will adopt an alternative approach, in which we abandon the 'gap' narrative and seek to explain how self-directed learners' behavior might be better captured by a single coherent account. This approach is in line with Koslowski's (1996) proposal that learners' apparently unscientific behavior may stem from researchers' incomplete theory of scientific thinking. More recently, Lapidow and Walker (2020, 2021) have proposed that intuitive inquiry and inference behavior may be specifically suited to support the goals of causal learning. In particular, that learners are primarily concerned with acquiring

generalizable causal knowledge that will support future inferences and action. When viewed through this lens, common "unscientific" behaviors may be reconceived as rational attempts to generate and evaluate evidence in pursuit of this goal.

The current project is the first empirical examination of this proposal; we ask whether clarifying the causal logic of a classic scientific reasoning task might account for learners' apparent failure to engage in formal scientific inquiry.

The Control of Variables Strategy

The control of variables strategy (CVS) is a domain-general approach to experimentation: In order to assess the causal relationship between a variable and some outcome of interest, that variable is manipulated while all others are held constant. CVS is considered an essential skill of scientific inquiry, and is included in standard curriculums for science education (Klahr et al., 2011; National Research Council, 2013). It has also been the focus of decades of research, which has overwhelmingly concluded that CVS is challenging for learners – both children and adults typically fail to employ the strategy correctly without extensive training (see Schwichow et al., 2016 for review).

Tschirgi (1980) developed one of the earliest and most influential assessments of CVS in a multivariate problem. In this study, children (2nd-, 4th-, and 6th-graders) and adults observe three variables combine to produce an outcome (e.g., using a sweetener, flour, and fat to bake a cake that comes out well). Learners were then asked to select an experiment to test the hypothesis that one variable (e.g., using honey as the sweetener) caused the outcome (good cake) and that the other two variables (e.g., using wheat flour and butterfat) are non-causal. One of these experiments (the 'VARY' option) changes the suspected causal variable (e.g., replacing honey with sugar) while keeping the other two variables constant. The other experiment (the 'HOLD' option) keeps the suspected cause constant and replaces the other two (see Table 1, top row).

Critically, Tschirgi (1980) treats the VARY option as offering a disconfirming test (following Popper, 1959) of the suspected cause, and thus the only informative choice. The finding that both children and adults select this experiment *only* when the observed outcome is negative (i.e., when the initial combination produces a bad cake), and prefer HOLD when the outcome is positive, is interpreted as evidence that self-directed inquiry does not follow scientific logic. That is, learners select actions based on their *tangible outcomes* (what actually happens) rather than their *information value* (what can be learned).

This study and its conclusions have remained central to research on the development of scientific reasoning. It is considered a standard assessment of CVS, serving as the basis for subsequent empirical research in this area (e.g., Croker & Buchanan, 2011; Varma et al., 2018; Zimmerman & Glaser, 2001). Recent reviews also continue to highlight Tschirgi's study as a central example of learners' failure in

formal experimentation (see Toplak et al., 2013; Zimmerman & Klahr, 2018).

When considered from the perspective of a causal learner however, this task is not a coherent test of experimentation. First, recall that the premise of the assessment is that the VARY option, which applies CVS to the suspected causal variable, is the only informative experiment (Tschirgi, 1980). However, the hypothesis presented to participants makes *two* distinct causal claims: the observed outcome is hypothesized to be (1) causally *dependent* on one variable and (2) causally *independent* of two other variables. Applying CVS to the independence claim would require keeping the suspected cause constant while changing the other two variables, which is the manipulation offered by the HOLD option. Nothing in the task presentation indicates that participants should not evaluate *this* claim, which means that both HOLD and VARY are equally valid responses to the task question. Thus, contrary to standard interpretations (e.g., Schauble et al., 1991; Toplak et al., 2013; Zimmerman & Klahr, 2018), learners' failure to consistently select VARY in this task (as seen in Tschirgi, 1980 but also Croker & Buchanan, 2011; Varma et al., 2018; Zimmerman & Glaser, 2001), is not necessarily evidence of failed scientific inquiry.

Additionally, for either of the options to be informative experiments, participants must make assumptions that are at odds with their real-world experience as causal learners. Specifically, both the VARY and HOLD options exchange the target variable(s) for novel ingredients (e.g., sugar is used to replace the honey in the VARY option), which participants must assume are causally inert. Without this assumption, any effect of removing the variable of interest cannot be distinguished from the possible effect of introducing its replacement. As a result, both experiment options present confounded control of variables, preventing learners from demonstrating their grasp of this inquiry strategy.

The Current Study

If, as proposed above, children and adults' self-directed inquiry behavior is grounded in their intuitions as causal learners, then clarifying the causal logic of this CVS assessment should improve their performance. To test this prediction, we presented participants with a modified version of Tschirgi's (1980) classic task that corrects the ambiguities in its causal logic.

In the current version, children and adults are introduced to a machine that lights up when blocks are placed into each of three differently-shaped slots on top (Fig. 1a). All participants also learn about special blocks called "blickets" that cause the machine to play music when it lights up (Fig. 1b). Next, a set of uncategorized novel blocks (all, some, or none of which could be 'blickets') is placed on the machine, causing it to light up *and* play music (Fig. 1c, left). A character offers a hypothesis about why this positive outcome occurred: that only the novel square block is a

'blicket', and that the novel circle and novel triangle are not (Fig. 1c, right).

Note how this design directly parallels the structure of Tschirgi's (1980) task (see Table 1). In both cases, participants learn that three distinct variables (i.e., types of ingredients, shapes of blocks) can be combined to produce an outcome (i.e., a cake, activating a toy). The tangible value of this outcome (i.e., a good/bad cake, music/no music) depends on whether at least one of these variables is causal (i.e., a causal ingredient, a blicket). At test, participants in both studies observe three novel variables of unknown causal status combine and result in a positive outcome (i.e., good cake, music). A character then offers a hypothesis that just one of the three variables (i.e., the sweetener, the square block) is responsible for this outcome, while the other two are not, and participants are asked to select an experiment from several options in order to test this hypothesis.

At this point, Tschirgi (1980) *intended* to offer a choice between a correct CVS experiment that is less likely to have a positive tangible outcome (e.g., removing the suspected cause of good cakes), and an uninformative experiment that is more likely to have a positive tangible outcome (e.g., keeping the suspected cause of good cakes constant). However, as explained above, these two options present equally informative tests of different parts of the hypothesis and both are subject to novel confounds. In the current task, participants are instead asked to answer *two* questions, each of which offers a choice between one informative and one confounded test of just *one* of the hypothesis' dual claims (Table 1, bottom row). The *Vary-Target Question* options both change the suspected causal variable, targeting the *dependence* claim—that the original outcome was causally dependent on this variable. The *Hold-Target Question* options both change the suspected non-causal variables, targeting the *independence* claim—that the original outcome was causally independent of these two variables. Critically, the two options presented in each question pit information value against an expected tangible outcome: One option replaces the suspected variable(s) with blue 'blickets,' which produces a confounded experiment that is certain to reproduce the positive tangible outcome. The other option replaces the suspected variable(s) with yellow inert blocks, which produces an informative experiment, but does not guarantee the positive tangible outcome. Thus, as intended by the original CVS task (Croker & Buchanan, 2011; Tschirgi, 1980), successful inquiry in this revised version requires identifying controlled experiments and being willing to forgo more desirable tangible outcomes.

Three patterns of behavior are possible. First, participants could consistently choose the option offering an informative experiment across both task questions (*Both Informative*). If a majority of participants follow this pattern, it would provide evidence of successful, scientific inquiry that is consistent with our hypothesis that learners' prior failures were due to causal ambiguities in the standard CVS assessment. Alternatively, performance on the current task

might be consistent with accounts that claim that self-directed learners are motivated by tangible outcomes, rather than information value (e.g., Tschirgi, 1980; Zimmerman, 2000). If so, then we would expect most participants to consistently select uninformative experiments (*Both Uninformative*), since these options are guaranteed to produce the positive tangible outcome (music). However, Tschirgi (1980) also raises the possibility that self-directed inquiry may follow a simple heuristic that avoids changing variables when outcomes are good and preferentially changes variables when outcomes are bad. If so, then we would expect the majority of participants to choose at chance between the two options (*No Preference*), as the options are matched on variable change and initial outcome.

Experiment 1

Like Tschirgi (1980), we present both children and adults with the same stimuli and materials. Adults completed the task entirely asynchronously via the Qualtrics online survey platform, whereas children were tested synchronously in an online video call with an experimenter. Prior to data collection, the design and analysis plan for children was preregistered (see https://aspredicted.org/YKY_ALF).
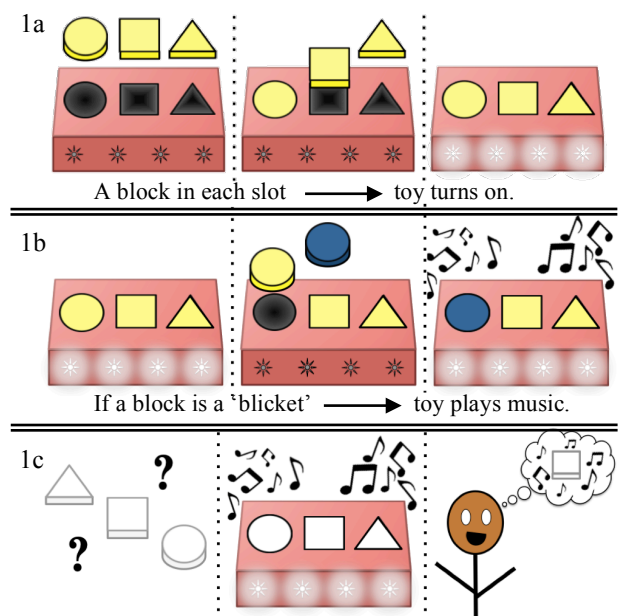

Figure 1: Stimuli introduced in the task: (1a, b) initial demonstration of toy and blocks (1c) the objects (left) and hypothesis (right) presented at test.

Methods

Participants The final sample will include a total of 114 participants, with 48 in each of three age groups tested in Tschirgi (1980): '*second-graders*' (reported mean age: 7 years, 2 months), '*fourth-graders*' (reported mean age: 9 years, 3 months), and '*adults*' (undergraduates). We doubled Tschirgi's sample size (*n* = 24 per age group) to

Table 1. Comparison of the Stimuli Presented in Tschirgi (1980) and the Current Study.

Study	Combination and Outcome Observed	Question	Options Presented	Hypothesis Claim Targeted by Option	Expected Information	Expected Tangible Outcome
Tschirgi (1980)	The combination: [a Flour] [a Sweetener] [a Fat] results in the cake turning out *well.*	Test Question	[same Flour] [*new* Sweetener] [same Fat]	Outcome dependent on Sweetener	Informative if '*new*' is inert	Uncertain (removes suspected cause)
			[*new* Flour] [same Sweetener] [*new* Fat]	Outcome independent of Flour and Fat	Informative if '*new*' are inert	Uncertain (keeps suspected cause)
Current Task	The combination: (circle, square, triangle) results in the toy turning on *with music.*	*Vary-Target* Question	(circle, blue square, triangle)	Outcome dependent on Square	Not Informative	Positive outcome is guaranteed
			(circle, yellow square, triangle)	Outcome dependent on Square	Informative	Positive outcome is uncertain
		Hold-Target Question	(blue circle, square, blue triangle)	Outcome independent of Circle and Triangle	Not Informative	Positive outcome is guaranteed
			(yellow circle, square, yellow triangle)	Outcome independent of Circle and Triangle	Informative	Positive outcome is uncertain

Note. Blue blocks are blickets and yellow blocks are inert. The causal status of white blocks is unknown.

account for collecting two, rather than four, responses per participant. [1]

All 48 adults were recruited from the University of California, San Diego' undergraduate student participant pool and received academic credit. An additional 14 adults were excluded and replaced for failing to answer correctly on one ($n = 11$) or more ($n = 3$) of the comprehension questions.

Seventy-seven children have been tested thus far[2], including 42 *second-graders* ($M = 87.23$ months, $SD = 3.49$ months, range: 78-95 months) and 35 *fourth-graders* ($M = 111.17$ months, $SD = 3.85$ months, range: 107-119 months). All children were recruited from a shared database and received a small gift for participating.

An additional ten children were also tested, but excluded due to experimenter error ($n = 5$), inattention/interference ($n = 3$), or technical issues ($n = 2$).

Stimuli The task was presented using a series of narrated, animated videos, as well as still images, all constructed using PowerPoint. These materials are all available on osf.org at [https://tinyurl.com/osfcake].

Procedure Participants watched several animated videos of a character (Ari) who is "playing with some blocks and a block toy today." They were told that, to turn the toy on, a block must be placed into each of three differently shaped slots (circle, square, and triangle) on top (Fig. 1a, right). Three yellow blocks (one of each shape) were shown next to the toy, and moved, one at a time, into the corresponding slots (Fig. 1a, middle). When the last of the three yellow blocks was in place, a row of lights along the front of the box lit up (Fig. 1a, left).

The yellow blocks were left in the toy and the lights stayed on (Fig.1b, left) while participants were told that "some blocks are a special kind, called 'blickets', and when there is a blicket on the toy, the toy will also play music when it turns on!" The yellow circle block was removed from the toy, causing the lights to turn off (Fig. 1b, middle). A blue circle block was then placed into the empty slot and the lights came back on, accompanied by a musical tone (Fig. 1b, right). This demonstrated that just *one* blicket is sufficient to produce the additional positive outcome. The process of removing the inert yellow block and replacing it with a blue 'blicket' was repeated for the square and triangle. The narration then reminded participants of the underlying causal system: "Remember, you need blocks in all three slots for the toy to turn on, but only one of the three blocks needs to be a blicket for the toy to play music when it turns on."

Next, all three blue blocks were removed from the toy simultaneously, causing the lights to go off and the music to stop. Three white blocks (one of each shape) appeared on the screen (Fig. 1c, left). The narration said, "Here are three new blocks. These blocks have lost their color and Ari doesn't know if any of them are blickets or not. Let's see what happens when we put them on the toy." All three blocks were then placed into the toy simultaneously, the lights turned on, and music played (Fig. 1c, middle).

Figure 2 shows the stimuli as seen by participants in the next step of the task. Initially, only the illustration of Ari and the empty toy were visible. The three sets of blocks (yellow, blue, white) appeared along the top of the screen in turn, and participants were reminded about the causal status of each set as it appeared. Both children and adults heard

---

[1] We did not include '*sixth-graders*' (reported mean age: 11 years, 3 months), since performance was not expected to differ from the other age groups.

[2] Data collection with children is still in progress, but the majority of the target sample for each age group has been tested.

each set described (i.e., "These blocks are blickets. When a blicket is on the toy, it plays music," for blue, "These blocks are not blickets, they don't make the toy play music," for yellow, and "These blocks are new, when we put them on the toy, it played music," for white). To ensure engagement in child participants, the descriptions for the yellow and blue sets were prefaced by the experimenter asking, "Are these ones blickets or not blickets?" Next, a thought bubble with the white square block and music notes appeared above Ari (Fig. 1c, right), and participants were presented the target hypothesis: "Ari thinks that only the new square is a blicket, and that the new circle and the new triangle are not blickets. What should Ari do to find out if this is true?" [3]

Participants then answered the two task questions in counterbalanced order. For each question, two images showing what Ari could do next appeared on the screen (Fig. 2). On the *Hold-Target Question*, the options were to, "try the toy again, still using the new square, but this time using the circle and triangle that are blickets" or "try the toy again, still using the new square, but this time using the circle and the triangle that are not blickets." On the *Vary-Target Question*, the options were to "try the toy again, but using the square that [is/is not] a blicket instead of the new square, and still using the new circle and new triangle." The order of blicket / non-blicket options was counterbalanced across participants. After selecting their answer to the first question, participants proceeded immediately to the second question. After answering the second question, participants saw a video of their second selection being placed on the toy, which lit up and made music.

Comprehension Questions. The procedure also included checks to assess participants' understanding and attention. Adults were given three multiple-choice questions, presented in random order, at the end of the task. Each question showed a set of three blocks of the same color (yellow, blue, or white) and asked participants to indicate "What would happen if Ari put these blocks into the toy?" from a set of four answers: "We know the toy will turn on and not play music" (correct for yellow), "We know the toy will not turn on and play music," "We know the toy will turn on and play music" (correct for blue and white), or "We don't know what the toy will do." Participants who answered any of these questions incorrectly were excluded and replaced.

In order to help maintain attention and engagement, children were asked two comprehension questions *prior* to the final test questions. First, after the three yellow blocks initially caused the toy to light up, the experimenter would pause the video and ask, "What do you have to do to make the toy turn on?" Second, after the first blue block was placed on the toy and the toy turned on, the experimenter

[3] The hypothesis phrasing was taken directly from Tschirgi (1980), but the question wording was altered from "What should s/he do to prove this point?" to "find out if this is true" based on pilot data.
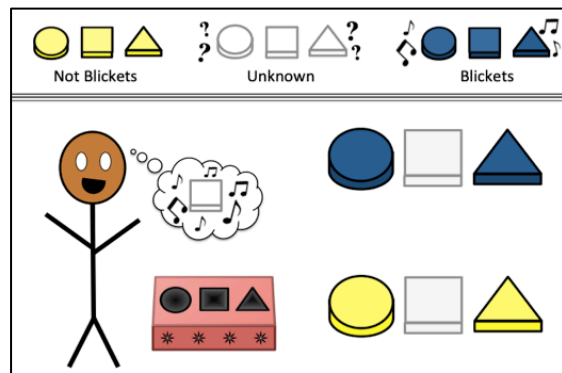
Figure 2: Stimuli at test for the *Hold-Target Question*.

paused and asked, "How many blickets do you need to make the toy play music?" Unlike adults, children were not excluded for answering these questions incorrectly. Instead, the experimenter responded with an explicit correction or confirmation, followed by a reminder of the rule (e.g., "That's right!" or "Not quite!" followed by "The toy will play music even if just one of the three blocks is a blicket and the rest are not!" for the second question).

Table 2. Participant performance by age group.

Age Group (N)	Responses Across Two Task Questions		
	Both Informative	*None Informative*	*No Preference*
Adults (48)	44	1	3
Fourth-graders (35)	24	3	8
Second-graders (42)	26	5	11

Results

Both adults and children were successful on this task (see Table 2). The vast majority of adults (91.67%) followed the *Both Informative* pattern, correctly selecting the informative option on *both* test questions, X^2 (2, $N = 48$) = 73.6, $p < 0.001$. This was also the case for both groups of children tested. A significant majority of *second-graders* (61.9%), X^2 (2, $N = 42$) = 16.71, $p < 0.001$, and *fourth-graders* (68.57%), X^2 (2, $N = 35$) = 20.63, $p < 0.001$, correctly selected the informative option on both test questions. A Fisher's exact test found no relationship between choice pattern (*Both Informative*, *None Informative*, *No Preference*) and age group (two-tailed, $p = 0.83$).

Additionally, there was no evidence that the type of test question impacted performance. Adults choose the correct, informative option on 97.92% of the *Hold-Target Questions*, and on 91.67% of the *Vary-Target Questions*

(Fisher's Exact, $p = 0.36$). Similarly, children (collapsing across age groups) chose correctly on 77.92% of the *Hold-Target Questions*, and 76.62% of the *Vary-Target Questions* (Fisher's Exact, $p = 1$).

Discussion

Contrary to the traditional interpretation of Tschirgi's (1980) results, learners had no difficulty identifying the correct CVS experiment to test a multivariate causal hypothesis. This was true even for the youngest children tested, suggesting that self-directed learners may indeed have an intuitive grasp of scientific inquiry from an early age.

Following Tschirgi's classic CVS task, we presented both children and adults with a multivariate causal system in which three components combine to produce an effect. However, we also corrected critical ambiguities in the original design. First, we addressed the logical flaw created by the two-part causal hypothesis (i.e., that the outcome is causally dependent on one variable and causally independent of the other two) by asking participants to select the appropriate experiments for both the dependence and independence claim in turn. Second, by using the initial demonstration to identify some objects as causal (blue blocks) and some as inert (yellow blocks), the experimental manipulations were not confounded by introducing variables of unknown causal status. Thus, this task presents the choice that Tschirgi (1980) *intended* to offer: between a confounded test that produces a positive tangible outcome, and an informative test that is not guaranteed to produce this outcome. We find that, overwhelmingly, both children and adults correctly selected experiments based on expected information value, rather than tangible outcome.

Importantly, prevailing explanations of the gap between formal and informal self-directed learning cannot account for the difference between our results and past research. Since the format of the final test question is identical to Tschirgi's (1980), the current task requires the same level of explicit coordination of theory and evidence, and places the same demands on fluency and coordination, as the original design. Instead, our results are consistent with the novel suggestion that ambiguous elements of the original task conflicted with participants intuitions as causal learners, leading researchers to underestimate their scientific inquiry skills.

When presented with a causally coherent version of a classic control of variables task, both children and adults preferentially selected correctly controlled and informative experiments. One possible objection to this conclusion is that our task design employs 'knowledge-lean' stimuli. That is, unlike previous studies that relied on familiar, real world content (e.g., Croker & Buchanan, 2011; Tschirgi, 1980), our study employed novel objects. Indeed, scientific reasoning researchers have argued that artificial 'blicket' stimuli simplify tasks in a way that overestimates learners' ability (see Lapidow & Walker, 2021; Weisberg et al., 2020 for discussion).

Two counterpoints to this objection may be raised. First, the existing research does not uniformly support the claim that decontextualized stimuli leads to overestimation of learners' abilities. For example, developmental studies find evidence for early competence in inquiry and inference tasks that challenge participants' real-world prior knowledge (e.g., Bonawitz et al., 2012; Schulz et al., 2007), as well as failure in tasks with novel stimuli (e.g., Kuhn & Phelps, 1982). There is also some evidence to suggest that it is possible to assess principles of formal scientific reasoning using knowledge-lean designs (Köksal-Tuncer & Sodian, 2018; Köksal et al., 2021).

Second, as noted above, the logic of the standard CVS assessment (as presented in Tschirgi, 1980; Croker & Buchanan, 2011) requires that learners assume that any novel variables are inert. This assumption is artificial and unrealistic. Determining the underlying causal structure of real-world multivariate contexts requires reasoning about and controlling for possible confounding effects of the variables used to isolate the targets of interest. Thus, although these "knowledge-rich" paradigms share superficial similarity to realistic contexts about which participants have prior beliefs, they ultimately lack meaningful resemblance to scientific inquiry in the real world.

Planned future work will directly test whether learners are spontaneously sensitive to the specific causal ambiguities we corrected for in the current task. Specifically, we will assess (1) whether learners consider both the dependence and independence claims of the hypothesis presented in the task, and (2) whether they interpret novel variables as potential confounds to control of variables experiments. Although testing these precise claims is beyond the scope of the current study, it *is* central to the proposal that self-directed inquiry is grounded in learners' real-world causal intuitions (Lapidow & Walker, 2021).

To conclude, this study offers a novel explanation for the 'gap' in self-directed learners' performance on informal and formal tests of scientific inquiry. We find evidence to suggest that commonly observed difficulties with applying the control of variables strategy might stem from a mismatch between task presentation and the intuitions of competent causal reasoners. Taken together, our results have the potential both to resolve a long-standing disconnect within the scientific literature, and to help us better understand the character of human inquiry and inference.

References

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. E. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234. https://doi.org/10.1016/j.cogpsych.2011.12.002

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. *Individual Development From*, *3*, 38–54.

Croker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: The effect of prior belief and outcome on children's hypothesis-testing strategies. *British Journal of Developmental Psychology*, *29*(3), 409–424. https://doi.org/10.1348/026151010X496906

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*. https://doi.org/10.1037/a0028044

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*(6045), 971–975.

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. https://doi.org/https://doi.org/10.1016/j.cogdev.2018.06.011

Köksal, Ö., Sodian, B., & Legare, C. H. (2021). Young children's metacognitive awareness of confounded evidence. *Journal of Experimental Child Psychology*, *205*, 105080. https://doi.org/https://doi.org/10.1016/j.jecp.2020.105080

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Mit Press.

Kuhn, D. (2002). What is Scientific Thinking and How Does It Develop? In *Blackwell Handbook of Childhood Cognitive Development* (pp. 371–393). https://doi.org/https://doi.org/10.1002/9780470996652.ch17

Kuhn, D. (2007). Jumping to conclusions. *Scientific American Mind*, *18*(1), 44–51.

Kuhn, D. (2012). The development of causal reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 327–335.

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. *Advances in Child Development and Behavior*, *17*(C), 1–44. https://doi.org/10.1016/S0065-2407(08)60356-0

Lapidow, E., Killeen, I., & Walker, C. M. (2022). Learning to recognize uncertainty vs. recognizing uncertainty to learn: Confidence judgments and exploration decisions in preschoolers. *Developmental Science*, *25*(2), e13178. https://doi.org/https://doi.org/10.1111/desc.13178

Lapidow, E., & Walker, C. M. (2020). The Search for Invariance: Repeated Positive Testing Serves the Goals of Causal Learning. In *Language and Concept Acquisition from Infancy Through Childhood* (pp. 197–219). Springer, Cham. https://doi.org/10.1007/978-3-030-35594-4_10

Lapidow, E., & Walker, C. M. (2021). Rethinking the "gap": Self-directed learning in cognitive development and scientific reasoning. *WIREs Cognitive Science*, e1580. https://doi.org/10.1002/wcs.1580

Liquin, E. G., & Lombrozo, T. (2020). A functional approach to explanation-seeking curiosity. *Cognitive Psychology*, *119*, 101276.

National Research Council. (2013). *Next Generation Science Standards: For States, By States*. The National Academies Press. https://doi.org/10.17226/18290

Osterhaus, C., Brandone, A. C., Vosniadou, S., & Nicolopoulou, A. (2021). Editorial: The Emergence and Development of Scientific Thinking During the Early Years: Basic Processes and Supportive Contexts. *Frontiers in Psychology*, *12*, 629384. https://doi.org/10.3389/fpsyg.2021.629384

Popper, K. R. (1959). The logic of scientific discovery Hutchinson. *Hughes, John,(1987)."La Filosofía de La Investigación Social", Breviarios, Fondo de Cultura Económica, México*.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*(9), 859–882. https://doi.org/10.1002/tea.3660280910

Schulz, L. E. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. In *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2012.06.004

Schulz, L. E., & Bonawitz, E. B. (2007). Serious Fun: Preschoolers Engage in More Exploratory Play When Evidence Is Confounded. *Developmental Psychology*. https://doi.org/10.1037/0012-1649.43.4.1045

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. In *Developmental Psychology* (Vol. 43, Issue 5, pp. 1124–1139). American Psychological Association. https://doi.org/10.1037/0012-1649.43.5.1124

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

Shtulman, A., & Walker, C. M. (2020). Developing an Understanding of Science. *Annual Review of Developmental Psychology*, *2*(1), 111–132. https://doi.org/10.1146/annurev-devpsych-060320-092346

Siler, S. A., & Klahr, D. (2012). Detecting, Classifying, and Remediating: Children's Explicit and Implicit

Misconceptions about Experimental Design. In *Psychology of Science: Implicit and Explicit Processes*. https://doi.org/10.1093/acprof:oso/9780199753628.003.0007

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing the development of rationality. *The Developmental Psychology of Reasoning and Decision-Making*, 7–35.

Tschirgi, J. E. (1980). Sensible Reasoning: A Hypothesis about Hypotheses. *Child Development*, *51*(1), 1–10. https://doi.org/10.2307/1129583

Varma, K., Van Boekel, M., & Varma, S. (2018). Middle School Students' Approaches to Reasoning about Disconfirming Evidence. *Journal of Educational and Developmental Psychology*, *8*(1), 1–28.

Weisberg, D. S., Choi, E., & Sobel, D. M. (2020). Of Blickets, Butterflies, and Baby Dinosaurs: Children's Diagnostic Reasoning Across Domains. *Frontiers in Psychology*, *11*, 2210. https://doi.org/10.3389/fpsyg.2020.02210

Zimmerman, C. (2000). The Development of Scientific Reasoning Skills. *Developmental Review*, *20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

Zimmerman, C., & Glaser, R. (2001). Testing Positive Versus Negative Claims: A Preliminary Investigation of the Role of Cover Story on the Assessment of Experimental Design Skills. In *CSE Technical Report*.

Zimmerman, C., & Klahr, D. (2018). Development of Scientific Thinking. In J. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (4th ed., pp. 1–25). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn407