Learners’ Causal Intuitions Explain Behavior in Control of Variables Tasks

Elizabeth Lapidow1 and Caren M. Walker2
1 Department of Psychology, University of Waterloo
2 Department of Psychology, University of California, San Diego

Self-directed learners are described as “intuitive scientists,” yet they often struggle in assessments of their scientific reasoning skills. We investigate a novel explanation for this apparent gap between formal and informal scientific inquiry behavior. Specifically, we consider whether learners’ documented failure to correctly apply the control of variables strategy might stem from a mismatch between task presentation and their intuitions as causal learners. In Experiment 1, children (7- and 9-year-olds) and adults were tested on a version of a traditional multivariate reasoning task (Tschirgi, 1980) that was modified to clarify ambiguous elements of the causal logic in the original design. In all age groups, a significant majority of participants selected informative experiments on this modified task, avoiding confounded actions with positive tangible outcomes. In Experiment 2, we replicate these results with real-world stimuli, and in Experiments 3 and 4, we provide direct evidence that self-directed learners apply specific causal intuitions to experimentation tasks. Together, these findings support a novel alternative interpretation of the apparently paradoxical gap between learners’ success in informal exploration and their error-prone experimentation—both behaviors are consistent with an intuitively causal approach to scientific inquiry.

Public Significance Statement
From early life, we are precocious exploratory learners, and intrinsically motivated to seek informative evidence about the world. However, we also often struggle in contexts that require scientific experimentation. We provide evidence for a novel explanation for this disconnect: the standard control of variables tasks used to assess experimentation skills violates learners’ intuitions about everyday causality. In a modified task, learners overwhelmingly select correct experiments, suggesting a novel continuity between exploration and scientific reasoning behavior.

Keywords: cognitive development, causal learning, scientific reasoning

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 skills, there is longstanding disagreement about learners’ competence in inquiry and inference (Weisberg & Sobel, 2022). Below, we briefly review the claims made by researchers on each side of this debate, before describing a possible resolution grounded in theories of causal reasoning. We then apply this novel theory to reexamine a prominent example of the disconnect between formal and informal self-directed inquiry skills—experimentation using the control of variables strategy (CVS). Specifically, we investigate whether children and adults’ poor performance in standard CVS tasks (Crocker & Buchanan, 2011; Tschirgi, 1980) may be because of everyday causal intuitions about experimentation that run counter to the classic task design.

Elizabeth Lapidow https://orcid.org/0000-0003-1290-6507
This work was supported by a Jacobs Foundation and National Science Foundation Career Award (Social, Behavioral, & Economic Sciences 2047581) to Caren M. Walker and a National Defense Science and Engineering Graduate Fellowship to Elizabeth Lapidow. The authors would like to extend particular thanks to Monica Van and Ashna Singh for their tireless work and invaluable insights on this project, to Dave Sobel for discussion and feedback on the task design, and to the Early Learning and Cognition Lab and the Children Helping Science teams for their support. The authors declare that they have no conflicts of interest. All procedures performed were in accordance with the 1964 Helsinki Declaration and approved by their Institutional Review Board. The design and hypotheses were preregistered: see https://aspredicted.org/db466.pdf for Experiment 1, https://aspredicted.org/a9r29.pdf for Experiment 2, https://aspredicted.org/ud92n.pdf for Experiment 3, and https://aspredicted.org/tx7hd.pdf for Experiment 4. The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/3dpdz/

Elizabeth Lapidow served as lead for conceptualization, formal analysis, investigation, methodology, and writing—original draft and served in a supporting role for funding acquisition and writing—review and editing. Caren M. Walker served as lead for funding acquisition, resources, supervision, and writing—review and editing and served in a supporting role for conceptualization and formal analysis. Elizabeth Lapidow and Caren M. Walker contributed equally to project administration.

Correspondence concerning this article should be addressed to Elizabeth Lapidow, who is now at the Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada. Email: e.lapidow@gmail.com
The Gap Between Exploring and Experimenting

The literature documenting inquiry behavior of self-directed learners presents two different accounts of their competence. Studies examining exploratory learning suggest that children and adults spontaneously prefer to engage in inquiries that are likely to improve their current knowledge (Liquin & Lombrozo, 2020; Schulz & Bonawitz, 2007) and make accurate inferences from the outcomes of these actions (Lagnado & Sloman, 2004; Lapidow & Walker, 2020). Indeed, cognitive development research has characterized learners as “intuitive scientists,” naturally motivated to seek informative evidence and rationally integrate it with existing knowledge (Gopnik & Wellman, 2012; Schulz, 2012). On the other hand, studies of scientific experimentation find that children and adults typically struggle with inquiry, producing confounded and confirmatory actions, and consistently privileging tangible outcomes over information (e.g., Kuhn, 2007; Siler & Klahr, 2012). In light of these findings, research on scientific reasoning abilities characterizes learners’ spontaneous and untrained behavior as biased and error-prone (e.g., Zimmerman, 2007; Zimmermann & Klahr, 2018).

There have been numerous efforts from both sides to “bridge the gap” between these two accounts of self-directed learning (e.g., Kuhn, 2012; Osterhaus et al., 2021; Shulman & Walker, 2020). However, to date, this dominant narrative has largely failed to produce an explanation that accounts for the full scope of learners’ documented behavior (but see Koslowski, 1996; Weisberg & Sobel, 2022 for classic and recent efforts). Scientific reasoning researchers have suggested that cognitive development tasks demonstrate 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 intuitively grasp this logic from an early age but lack the extraneous coordination and verbal fluency to employ it in scientific reasoning tasks (e.g., Bullock & Ziegler, 1999; Lapidow & Walker, 2020).

While each of these approaches captures a truth about early cognition, neither is wholly satisfying. It is true that the coordination of theory and evidence that occurs during early learning precedes the development of metacognition, and that metacognitive thinking is closely connected to scientific reasoning abilities (see Weisberg & Sobel, 2022). It is also true that immature attention, memory, and language capacities limit young learners’ abilities. However, any explanation that appeals to cognitive immaturity to explain the “gap” must contend with the objection that adult learners—with mature language, metacognition, and executive function—still commit many of these same scientific errors (e.g., Kuhn, 2007; Kuhn et al., 1995; Wason, 1960). Indeed, the classic evidence of poor experimentation with control of variables comes from older children (9- and 11-year-olds) and adults (Tschirgi, 1980).

In addition to their inability to account for the persistence of scientific errors over the life span, there is also the conceptual objection that existing accounts are fundamentally incomplete explanations of self-directed learning. Theories from cognitive development and scientific reasoning both tend to demarcate the existing research into “valid” or “invalid” evidence of learners’ scientific competence, albeit on different grounds (i.e., either for insufficient demonstration of explicit understanding or for developmentally inappropriate task demands). In doing so, both camps leave some evidence behind and construct explanations of self-directed learners that ultimately only account for a subset of their actual documented behaviors.

Here, we aim to present an alternative to the “gap” narrative and explain the apparently paradoxical evidence as two pieces of an internally consistent picture of self-directed learning. We propose that self-directed learners’ inquiry and inference behavior may be intuitively and specifically suited to support causal learning (Lapidow & Walker, 2022). That is, the same expectations, considerations, and goals that serve children and adults when learning by exploration in the real world can also cause them to respond in apparently unscientific and uninformative ways. Our position aligns with Koslowski’s (1996) proposal that learners’ unscientific behavior may stem from researchers’ incomplete theory of scientific thinking, and we further suggest that understanding the self-directed learner as intuitively causal provides a more complete theory. The goal of the present research is to investigate the possibility that learners are intuitively causal scientists and that both their “errors” and “successes” are behaviors that are consistent with the considerations of informative experimentation during real-world causal learning.

First, in Experiments 1 and 2, we test untrained children and adults in a novel version of the CVS task that removes the elements that we suggest are in conflict with learners’ causal intuitions. Then, Experiments 3 and 4 isolate these elements of the original task design to directly demonstrate how they conflict with the expectations that learners spontaneously bring to experimentation tasks.

Causal Logic in the CVS

The 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 scientific inquiry skill 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 without extensive training, CVS is challenging for learners of all ages (see Schwichow et al., 2016 for review). Notably, although adults typically outperform children on CVS tasks, use of the strategy remains far from perfect (e.g., Kuhn et al., 1995). Indeed, even adults who have had some formal scientific training and are aware of the importance of the CVS often have significant difficulty adhering to it in both their inquiry and inference (Boudreaux et al., 2008).

Tschirgi (1980) developed one of the earliest and most influential assessments of CVS. Children (second, fourth, and sixth graders) and adults observe a combination of three variables produce a particular outcome (e.g., using a sweetener, flour, and fat in baking a cake that comes out well). Participants are then asked to select an experiment to test the hypothesis that one variable (e.g., using honey as the sweetener) caused this outcome (a 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 Figure 1, left column).

Critically, Tschirgi (1980) treats the VARY option as offering an informative, disconfirming test (following Popper, 1959) of the suspected cause and thus the only correct answer 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 from what happens). This study and its conclusions have remained central to research on the experimentation abilities of self-directed learners. It is considered a standard assessment of CVS, serving as the basis for subsequent empirical research (e.g., Croker & Buchanan, 2011; Moeller et al., 2022; Varma et al., 2018; Zimmerman & Glaser, 2001) and highlighted in recent reviews as the primary example of learners’ failure in formal experimentation (see Toplak et al., 2013; Zimmerman & Klahr, 2018).

However, when considering the nature of informative inquiry during causal learning in the real world, there is no straightforward interpretation of this CVS task. First, the assessment assumes that the VARY option, which applies CVS to the hypothesized causal variable, is the only informative experiment and thus the correct choice (Tschirgi, 1980). However, the hypothesis presented to participants actually consists of two distinct causal claims: the observed outcome is hypothesized to have been (a) causally dependent on one variable and (b) causally independent of two other variables. Although the VARY option applies CVS to test the dependence claim, testing the independence claim requires keeping the suspected cause constant while changing the other two variables—the manipulation offered by the HOLD option.

Critically, both independence and dependence relations are important to real-world causal learners, as both kinds of knowledge are needed to successfully navigate the causal world (e.g., in following a recipe, it is important to know both which ingredients can be substituted and which cannot). In addition, nothing in the task presentation indicates to participants that they should only evaluate the dependence claim. 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 or prioritization of tangible outcomes over information value.

---

**Figure 1**

Comparison of Study Design in Tschirgi (1980) and in Experiment 1 of the Current Study

<table>
<thead>
<tr>
<th>Tschirgi (1980)</th>
<th>Current Study (Vary Trial)</th>
<th>Current Study (Hold Trial)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Initial event observed: margarine honey wheat flour</td>
<td>(1) Initial event observed: unknown unknown unknown</td>
<td>(1) Initial event observed: unknown unknown unknown</td>
</tr>
<tr>
<td>Cake comes out well (desirable outcome)</td>
<td>Toy makes music when it turns on (desirable outcome)</td>
<td>Toy makes music when it turns on (desirable outcome)</td>
</tr>
<tr>
<td>(2) Hypothesis to be tested The honey caused the cake to come out well, the type of flour and fat did not matter</td>
<td>(2) Hypothesis to be tested Only the new square is a blicket, and the new circle and new triangle are not blickets.</td>
<td>(2) Hypothesis to be tested Only the new square is a blicket, and the new circle and new triangle are not blickets.</td>
</tr>
<tr>
<td>(3) Option scored as correct margarine sugar wheat flour</td>
<td>(3) Option scored as correct unknown known known</td>
<td>(3) Option scored as correct known known known</td>
</tr>
<tr>
<td>Considered informative and unlikely to have desirable tangible outcome</td>
<td>Information: informative test Tangible Outcome: uncertain</td>
<td>Information: informative test Tangible Outcome: uncertain</td>
</tr>
<tr>
<td>(4) Option scored as incorrect butter honey white flour</td>
<td>(4) Option scored as incorrect unknown known unknown</td>
<td>(4) Option scored as incorrect known known known</td>
</tr>
<tr>
<td>Considered uninformative and likely to have desirable outcome</td>
<td>Information: confounded test Tangible Outcome: desirable</td>
<td>Information: confounded test Tangible Outcome: desirable</td>
</tr>
</tbody>
</table>

Note. The options on Tschirgi (1980) cannot be classified as informative or confounded tests since they include novel variables that may or may not be confounds. From “Sensible Reasoning: A Hypothesis About Hypotheses,” by J. E. Tschirgi, 1980, Child Development, 51(1), 1–10 (https://doi.org/10.2307/1129583). Copyright 1980 by Wiley. Reprinted with permission. See the online article for the color version of this figure.
The second causal objection to interpreting participants’ performance on the CVS task as evidence of failed scientific inquiry is that neither answer option correctly applies the CVS, since they are both confounded by causally ambiguous variables. Specifically, both the VARY and HOLD manipulations exchange the target variable(s) for novel ingredients (e.g., sugar is used to replace the honey), which participants have no information about. The principle of CVS is to determine the causal influence of one variable on an outcome by observing whether that outcome changes when this variable is removed or replaced. For this manipulation to be informative, this must be the only change. All other potential sources of variation must be controlled for, or the experiment is confounded. However, in the standard CVS assessment, any effect of removing the target variable cannot be distinguished from effects of introducing its replacement, meaning that neither answer option presented to participants is a truly informative experiment. The only way to bypass this confound is to assume that all the novel variables are causally inert (e.g., that introducing sugar has no impact on cake). Critically, however, this assumption is incompatible with competent inquiry in real-world causal learning, in which one variable might easily imitate or inhibit the effects of another. Indeed, learners as young as 5 understand that the ambiguous causal status of untested, novel variables means that they have the potential to be causally efficacious (Fernbach et al., 2012; Sobel et al., 2017). Furthermore, recognizing the potential for new variables to confound experiments is central to a mature understanding of CVS.

To our knowledge, these conflicts between assessments of CVS experimentation and real-world causal considerations have never been previously addressed. Subsequent studies aiming to build upon Tschirgi’s (1980) task have maintained both these elements of the original task design (e.g., Croker & Buchanan, 2011; Moeller et al., 2022). For example, in Croker and Buchanan (2011), the nominally correct experiment to test whether drinking milk is a cause of healthy teeth is to replace milk with cola, introducing a confounding cause of unhealthy teeth. Given the realities of causal learning, the standard method for assessing learners’ ability to employ CVS does not have a correct answer: both are equally valid interpretations of how to apply CVS to test the hypothesis, and both are confounded by the presence of novel variables.

The Current Study

Thus far, we have made two novel proposals: First, that self-directed learners’ approach to experimentation is intuitively suited to the realities of learning about the causal world. Second, that these intuitions are in direct conflict with how learners’ understanding of CVS experimentation has been assessed and interpreted. Importantly, this proposal differs from past explanations of the “gap” that argue that children are intuitive scientists who are hampered by developmentally inappropriate task demands (e.g., Lapidow & Walker, 2020; Weisberg et al., 2020). Instead, this study will test the novel claim that the standard control of variables task conflicts with both children and adults’ causal intuitions about inquiry in the real world. If true, then we would expect self-directed learners to succeed at a CVS task in which these conflicts are corrected. We conducted four experiments to test this prediction. In Experiment 1, we presented participants with a modified version of Tschirgi’s (1980) task that corrects the conflicts outlined above. If the characterization of self-directed learners as competent causally minded scientists is correct, then both children and adults should select experiments based on information value rather than tangible outcomes, as predicted by the standard interpretations (Tschirgi, 1980; Zimmerman, 2000). Next, Experiment 2 rules out the possibility that success on the revised task can be explained by the participants’ lack of prior knowledge about the task stimuli. Finally, Experiments 3 and 4 provide direct evidence of learners’ causal expectations and their sensitivity to the ambiguous elements of the original task.

Experiment 1

In this experiment, we revise the standard version of the CVS assessment (Croker & Buchanan, 2011; Moeller et al., 2022; Tschirgi, 1980, etc.), correcting for the hypothesized conflicts with learners’ causal intuitions. Children and adults are introduced to a machine that lights up when blocks are placed into each of three differently shaped slots on top (square, circle, and triangle; Figure 2A). They learn that, while any three blocks of the correct shapes will cause the machine to light up, there is also a special class of blocks (“blickets”) that cause the machine to also play music when it lights up (an additional, attractive outcome; Figure 2B). Next, a set of uncategorized novel blocks (all, some, or none of which could be “blickets”) is placed into the machine, causing it to light up and play music (Figure 2C, left). A character then offers a hypothesis about the cause of this positive outcome: that only the novel square block is a “blicket” and that the novel circle and novel triangle are not (Figure 2C, right).

Figure 1 illustrates how this design directly parallels the structure of Tschirgi’s (1980) task. 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., baking a cake, lighting up the machine). 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., is an ingredient that causes good cakes, is a blicket). At test, participants in both studies observe three novel variables of unknown causal status combine and result in a desirable outcome (i.e., baking a good cake, lighting up with 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, and the other two are not.

Participants are then asked to select an experiment from several options to test this hypothesis. At this point, Tschirgi (1980) intends to offer a choice between a correct CVS experiment that is less likely to have a desirable tangible outcome (e.g., removing the suspected cause of good cakes) and an uninformative experiment that is more likely to have a desirable 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.1 In the current task, participants are instead asked to answer two questions, each offering a choice between one informative and one confounded test of just one of

---

1 Note that the VARY option does reduce the number of possible states (“blicket” or “nonblicket”) of the three unknown variables to a greater degree than the HOLD option. However, participants’ explicit goal in the task is to select an experiment to determine the truth of a particular hypothesis. Therefore, the information value of the options lies not in the number of possibilities remaining but in whether or not these possibilities confirm or disprove that hypothesis. Thus, both HOLD and VARY have the same information value.
are motivated by tangible outcomes rather than information value (e.g., Tschirgi, 1980; Zimmerman, 2000). If so, then we would expect the majority of participants to consistently select uninformative experiments (both uninformative), since these options are guaranteed to produce the desirable tangible outcome (music). However, Tschirgi (1980) also raises the possibility that self-directed inquiry may follow a simple heuristic of avoiding changes to variables when outcomes are good and preferring to change them 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 both answer options are matched on the value of the initial outcome (desirable) and the direction (change or not) of the manipulation.

Method

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 hypotheses, design, and analysis plan for children were preregistered (see https://aspredicted.org/db466.pdf).

Participants

A total of 96 children, 48 in each age group, were tested in Experiment 1. The age groups and sample size were based on the participants tested in Tschirgi (1980): “second graders” (reported $M_{age} = 7$ years, 2 months), “fourth graders” (reported $M_{age} = 9$ years, 3 months), and “adults” (undergraduates). We doubled Tschirgi’s sample size ($n = 24$ per age group) to account for collecting two, rather than four, responses per participant. Two groups of children were tested: 48 second graders ($M = 86.91$ months, $SD = 3.98$, range: 77.8–95.5 months, 35.4% female) and 48 fourth graders ($M = 112.13$ months, $SD = 4.33$, range: 106.9–125.1 months, 56.2% female). All children were recruited from a shared database maintained by the developmental research labs at the University of California, San Diego. Parent-reported racial identity was 1.7% African American, 9.5% Asian, 1.7% Filipino, 6.9% Hispanic/Latino, 1.7% Indian, 4.3% Mexican American, 0.9% Pacific Islander, 39.7% White, 14.7% multiracial/biracial, and 19% decline to answer. While individual socioeconomic information was not collected, the median household income for the local population from which the sample was drawn was $83,450. Families received a small gift for participating. An additional 12 children were also tested but excluded because of experimenter error ($n = 6$), inattention/interference ($n = 5$), or technical issues ($n = 1$).

Adults ($n = 48$) were recruited from the University of California, San Diego undergraduate student participant pool and received academic credit. During the academic session when the experiment was conducted, this participant pool self-reported race as 21.3% Asian, 21.3% Hispanic/Latino, 18% White, 8.2% multiracial/biracial, and 31.1% decline to answer and gender identity as 47.5% female, 21.3% male, 1.6% nonbinary, and 29.5% decline to answer. While socioeconomic information was not collected, the median family

---

2 We did not include “sixth graders” (reported $M_{age} = 11$ years, 3 months), since performance was not expected to differ from the “fourth grader” age group.
income for the undergraduate population at the University of California, San Diego during the year the study was conducted was $82,000. 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.

**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://osf.io/3dpsr/).

**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 (i.e., to make it light up), a block must be placed into each of three differently shaped slots (circle, square, and triangle) on top (Figure 2A, right). Three yellow blocks (one of each shape) were shown next to the toy, and moved, one at a time, into the corresponding slots (Figure 2A, middle). When the last of the three yellow blocks was in place, a row of lights along the front of the box lit up (Figure 2A, left).

The yellow blocks were left in the toy and the lights stayed on (Figure 2B, 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 (Figure 2B, middle). A blue circle block was then placed into the empty slot and the lights came back on, accompanied by a musical tone and the appearance of music notes (Figure 2B, 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 narrator 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 notes to disappear. Three white blocks (one of each shape) appeared on the screen (Figure 2C, left). The narrator 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 (Figure 2C, middle).

The next step of the task presented a reminder, followed by the test questions (Figure 3 shows the stimuli as it was shown to participants). Initially, only the illustration of Ari and the empty toy were visible. The three sets of three blocks (yellow, blue, and 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 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 preceded by the experimenter asking, “Are these ones blickets or not blickets?” and then either confirming or correcting the child with “yes!” or “no.” Next, a thought bubble with the white square block and music notes appeared above Ari (Figure 2C, 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?”

**Figure 3**

*Stimuli Presented at Test for the Hold-Trial of Experiment 1*

![Stimuli Presented at Test for the Hold-Trial of Experiment 1](image-url)

*Note.* See the online article for the color version of this figure.
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 (Figure 3). On the hold trial, 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 trial, 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/nonblicket 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 (three yellow, three blue, or three 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 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).

**Results**

Both adults and children were successful on this task (see Table 1). The vast majority of adults (91.67%) followed the both informative pattern, correctly selecting the informative option on both test questions, \( \chi^2(2, N = 48) = 73.62, p < .001 \). This was also the case for both groups of children tested. A significant majority of second graders (60.42%), \( \chi^2(2, N = 48) = 17.37, p < .001 \), and fourth graders (66.67%), \( \chi^2(2, N = 48) = 26, p < .001 \), correctly selected the informative option on both test questions. Fisher’s exact test found no relationship between choice pattern (both informative, both uninformative, no preference) and age group (two-tailed, \( p = .75 \)).

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 trial and on 91.67% of the vary trial (Fisher’s exact, \( p = .36 \)). Similarly, children (collapsing across age groups) chose correctly on 77.08% of the hold trial and 76.04% of the vary trial (Fisher’s exact, \( p = 1 \)).

**Discussion**

When presented with a revised version of the classic control of variables task, both children and adults preferentially selected controlled and informative experiments. Contrary to the traditional interpretation of Tschirgi’s (1980) results, learners had no difficulty identifying the correct CVS experiment and were willing to select it even though these options were less likely to bring about a desirable tangible outcome than the confounded experiments. These results contrast with the typical finding that self-directed learners struggle to grasp the scientific logic of multivariate causal hypothesis testing without repeated training. The critical difference, we suggest, between this experiment and previous research is that the current version of the task corrected elements of the CVS assessment that potentially conflict with causal learning in the real world. Our results are consistent with this suggestion and support the view that self-directed learners intuitive grasp of scientific inquiry is tailored to the particular concerns and characteristics of causal learning.

There is, however, an important potential objection to this interpretation: Unlike previous studies that presented CVS problems using familiar, real-world content (e.g., Croker & Buchanan, 2011; Tschirgi, 1980), our task employed novel, “knowledge-lean” stimuli. Researchers in the domain of scientific reasoning have often argued that the use of this type of artificial stimuli may simplify tasks and lead to overestimation of learners’ abilities (e.g., Kuhn, 2002, 2012; but see also Lapidow & Walker, 2022; Weisberg et al., 2020). The existing research does not uniformly support this suggestion: children sometimes succeed on inquiry and inference tasks where the immediate evidence contradicts their prior beliefs about the real world (e.g., Bonawitz et al., 2012; Schulz et al., 2007) and sometimes fail on tasks with entirely novel artificial content (e.g., Kuhn & Phelps, 1982). That said, the ability to coordinate one’s prior beliefs with incoming evidence is a defining skill of mature scientific thinking (e.g., Kuhn, 1993, 2012) and this continues to pose considerable difficulty for self-directed learners well into adulthood (e.g., Greenhout et al., 2004). Therefore, while there is growing evidence that it is possible to assess principles of formal scientific reasoning using knowledge-lean designs (e.g., Köksal et al., 2021; Köksal-Tuncer & Sodian, 2018), Experiment 2 takes this alternative explanation seriously and assesses whether learners’ success replicates when presented in a familiar, real-world context.

**Experiment 2**

Experiment 2 aims to replicate performance in Experiment 1 when the corrected CVS task is presented using familiar, real-world content. We therefore replaced the novel machine with one of the
original domains used in Tschirgi (1980), in which participants hear about a character who is baking cakes by combining three ingredients (a flour, a sweetener, and a fat). Instead of “blickets,” participants learn about special “fancy” ingredients that cause a desirable outcome of cakes coming out “extra-rich and tasty.” At test, the character tries combining three novel ingredients (all, some, or none of which could be “fancy”), and the cake comes out extra-rich. The character then offers a hypothesis about the cause of this desirable outcome: that only the novel sweetener is “fancy” and that the novel flour and fat are not.

Following the corrected logic used in Experiment 1, participants then answer two questions asking them to choose between two possible experiments to test the presented hypothesis. On the vary trial, both options remove the suspected cause (i.e., the sweetener) and replace it with either a fancy sweetener (offering a desirable but uninformative outcome) or a plain sweetener (offering an informative outcome with an uncertain tangible value). On the hold trial, both options keep the suspected cause constant and replace the two suspected inert ingredients (i.e., the flour and fat) with either two fancy or two plain ingredients. The trade-off between information value and tangible outcome is thus made especially salient: any of the three new ingredients may cause desirable cakes, or it may not. Revealing whether the character’s hypothesis is correct requires pairing these ingredients with “plain” ones that are known to be unable to produce that effect on their own. Conversely, creating a desirable outcome requires obscuring the causal status of the new ingredients by pairing them with “fancy” ones that are known to produce the effect.

As in Experiment 1, three patterns of behavior (both informative, both uninformative, or no preference) are possible. If prior knowledge makes it more difficult for learners to employ correct CVS, then we would not expect the same strong preference for informative experiments (both informative) that we saw in Experiment 1. Instead, participants may be more likely to choose at chance between the two options (no preference), reflecting uncertainty about which option is the better choice, or they may fall back on the simple heuristic of changing behavior in response to bad outcomes. Another possibility is that the familiar, realistic stimuli will reveal a preference for tangible outcomes over information value that Experiment 1 was unable to capture. That is, self-directed learners might, in fact, prioritize actions that will bring about desirable events, but the music outcome was simply not high-value enough to elicit that behavior. If so, presenting learners with the same desirable tangible outcome (a good cake) that was at stake in Tschirgi’s experiment should lead participants to consistently choose the uninformative experiment that is guaranteed to produce this event (both uninformative).

Given that children and adults’ performance followed the same pattern, both in Experiment 1 and in the classic control of variables task (Tschirgi, 1980), only adults were included in Experiment 2. Critically, this age group also presents a more challenging test of the effect of real-world stimuli, since adults are more likely to have prior beliefs and domain knowledge about baking than young children.

Method

The design, predictions, and analysis plan for Experiment 2 were preregistered prior to the start of data collection (see https://osf.io/3dpsr/).
The Figure 4 illustrates the next set of stimuli seen by participants. As in Experiment 1, each set of variables appeared one at a time, along the top of the screen, as the narrator reminded participants of their causal status (i.e., “These ingredients are not fancy. They make plain cakes,” “These ingredients are fancy. When one ingredient in a cake is fancy, it comes out extra-rich,” and “These ingredients are new. When they were used in a cake, it came out extra-rich.”). A thought bubble with the new sweetener appeared above Ari, and participants heard the target hypothesis: “Ari thinks that only the new sweetener, honey, is a fancy ingredient and that neither the new rice flour nor the new fat margarine are fancy. What should Ari do to find out if this is true?”

The two task questions were presented in counterbalanced order. On the hold trial, the options were, “Ari can bake another cake, still using the new sweetener, but this time using the plain white flour and the plain butter fat” or “Ari can bake another cake, still using the new sweetener, but this time using the fancy wheat flour and the fancy milk fat.” On the vary trial, the options were to “bake another cake, but using the [plain powdered sugar/fancy brown sugar] sweetener instead of the new sweetener, and still using the new rice flour and new fat margarine.”

Comprehension Questions. The understanding and attention check in Experiment 2 was structurally identical to Experiment 1, with content matching the new stimuli. Each of the three multiple-choice questions showed one of the three sets of ingredients seen during the task (plain, fancy, or new) and asked participants to indicate “What would happen if Ari put these ingredients into the oven?” from a set of four answers: “We know they will bake a plain cake” (correct for plain), “We know they will not bake a cake,” “We know they will bake an extra-rich cake” (correct for fancy and new), or “We don’t know what they will bake.”

Results and Discussion

When asked to select experiments to test a multivariate causal hypothesis embedded in real-world stimuli, untrained participants continued to show an understanding of correct CVS. As in Experiment 1, a significant majority of participants (83.67%) followed the both informative pattern, correctly selecting the informative experiment on both test questions, $\chi^2(2, N = 49) = 56, p < .001$. Again, there was no difference between the two types of test questions: The informative option was selected on 89.8% of hold trials and 83.67% of vary trials (Fisher’s exact, $p = .55$).

Despite the use of more detailed stimuli, about which adults are very likely to have real-world prior knowledge, participants in Experiment 2 behaved almost identically to those in Experiment 1. Fisher’s exact test showed no association between the cover story (knowledge-lean in Experiment 1 or knowledge-rich in Experiment 2) and choice pattern (both informative, both uninformative, no preference), two-tailed, $p = .37$. Indeed, the only variation in the results is the slight increase of participants in the both uninformative category ($n = 1$ in Experiment 1, $n = 5$ in Experiment 2). This suggests that, as intended, the more realistic stimuli likely offered a higher value tangible outcome than the music in Experiment 1. However, this temptation did not disrupt learners’ willingness to conduct informative experiments.

Experiment 2 provides compelling evidence that learners’ successful use of CVS in Experiment 1 was not because of their lack of prior knowledge or the use of simplified stimuli. This result rules out a potential objection to our claim that the difference in performance between the current study and previous research stems from relative alignment with causal intuitions. With this established, Experiments 3 and 4 aim to generate direct evidence for our claim
that causal intuitions influence learners’ behavior on assessments of experimentation by demonstrating that learners spontaneously apply these intuitions when reasoning about control of variables. First, Experiment 3 assesses whether learners recognize the difference between the dependence and independence claims embedded within the target hypothesis. Then, Experiment 4 tests whether learners spontaneously treat the introduction of casually ambiguous variables as a potential confound for control of variables experiments. Together, these experiments will not only shed light on the precise cause of the errors on the classic task, but also provide direct evidence for our proposal that self-directed learners approach to inquiry is grounded in real-world causal intuitions.

Given the lack of stimuli effects, the remaining experiments use only the novel, knowledge-lean stimuli from Experiment 1. Additionally, we opted to only include adult participants in these two follow-up studies. This decision was both pragmatically and theoretically motivated. Isolating the elements of the task design and evaluating learners’ specific causal intuitions about them required increasing the cognitive and verbal demands of the task. In particular, participants had to follow an instruction to try to identify which of the options was “most correct,” given the particular situation presented. We expected that children might perform differently on these tasks for reasons unrelated to their understanding of experimentation. Critically, as noted above, both Experiment 1 and the original Tschirgi (1980) study show the same pattern of performance for both adults and children, suggesting there is very little difference in how learners of different ages approach the task. We therefore feel confident that learners’ intuitions about conducting informative experiments and the way this informs their subsequent behaviors are the same for both grade schoolers and adults.

**Experiment 3**

Recall that participants in the standard CVS task are presented with an ambiguous choice—selecting which of two equally valid applications of CVS is the better test the target hypothesis. This is because the hypothesis itself is composed of two distinct causal claims that (a) one of the variables is responsible for the outcome (e.g., the quality of the cake is causally dependent on the sweetener used) and (b) the other two variables are not (e.g., the quality of the cake is causally independent of the flour and fat used). Manipulating the suspected causal variable while keeping the other two unchanged, typically labeled the “VARY” option, applies CVS to test the dependence claim. However, the alternative “HOLD” option, keeping the suspected cause unchanged and manipulating the other two, is an equally valid application of CVS to test the independence claim.

The standard interpretation assumes that the dependence claim is the only one that participants consider when making their choice. However, in the absence of any direct instruction, there is no reason why learners should have more interest in determining which variable is responsible for the outcome than in determining whether any variable improves quality. The results of Experiments 1 and 2 show that learners are equally competent at reasoning about CVS in the context of either independence or dependence claims. However, these studies cannot reveal whether learners who are confronted with the original (ambiguous) presentation of the CVS test question recognize the independence and dependence claims as distinct, or whether they are inclined to prioritize either one during inquiry.

Here, we test whether learners’ choice of experiments reflects an understanding that the HOLD and VARY options are differently conclusive applications of CVS, depending on which particular claim they intend to test. We present the original two-part hypothesis and a choice between “HOLD” and “VARY” options. However, we explicitly instruct participants to focus on just one of the two claims when selecting their experiment. In the dependence condition, participants are asked to find out whether the claim that “the square block is a blicket” is correct, and in the independence condition, they are asked to find out whether the claim that “the circle and triangle are not blickets” is correct. Participants in both conditions are given a choice between two options: replace-causal replaces the suspected causal variable (square) with an inert yellow block and leaves the suspected noncausal variables (circle and triangle) unchanged. This is analogous to the “VARY” option in the classic task and offers a conclusive test of the independence claim (it also has some potential to be informative about the dependence claim but only under certain outcomes). Conversely, replace-noncausal keeps the suspected causal variable (square) the same and replaces suspected noncausal variables (circle and triangle) with inert blocks. This is analogous to the classic “HOLD” option and offers a conclusive test of the dependence claim (again, it is also potentially informative about the independence claim, depending on the outcome).

As in the original CVS task design, both options presented to participants are potentially informative experiments. However, if learners understand the differences between causal independence and dependence claims, then they should recognize that one option is the “better” choice in each condition and select it (see Table 2 for an overview of the logic of this design). Each option has two possible outcomes (the machine either will or will not make music) and whether that outcome is conclusive depends on which claim is being tested. Specifically, if the suspected cause is removed (replace-causal), then either outcome will reveal whether the remaining variables are causal. However, this manipulation is only informative about the suspected cause if the machine does not make music. Similarly, either outcome of keeping the suspected cause (replace-noncausal) reveals whether or not it is causal, but is only informative about the two suspected noncauses if the machine does not make music. If learners recognize the dual nature of the target hypothesis and are equally ready to accept either claim as the target of their inquiry, we should see a preference for the most informative test of each claim: a preference for replace-causal in the independence condition and for replace-noncausal in the dependence condition.

In contrast, if learners do not distinguish these claims, then we would expect participants in both conditions to choose at chance between the two options. Alternatively, participants might behave more similarly to what is expected in the standard CVS assessment and latch on to one portion of the hypothesis as the more important or “real” claim. However, it need not be the case that the dependence claim, which is presumed to be the focus in the standard interpretation, will be more appealing. In this case, we would expect all participants to show a preference for the same option, either replace-causal or replace-noncausal, regardless of condition.
Table 2
Task Design for Experiment 3

<table>
<thead>
<tr>
<th>Answer option</th>
<th>Outcome</th>
<th>Indicates</th>
<th>Dependence condition: (square is blicket)</th>
<th>Independence condition: (circle and triangle are not blickets)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Replace-noncausal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Circle: yellow (inert) Square: white (new) Triangle: yellow (inert)</td>
<td>Machine plays music</td>
<td>1. New square is a blicket. 2. New circle and new triangle are unknown.</td>
<td>Outcome is informative (claim is true)</td>
<td>Outcome is uninformative</td>
</tr>
<tr>
<td></td>
<td>Machine does not play music</td>
<td>1. New square is not a blicket. 2. New circle and/or new triangle are blickets.</td>
<td>Outcome is informative (claim is false)</td>
<td>Outcome is informative (claim is false)</td>
</tr>
<tr>
<td><strong>Replace-causal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Circle: white (new) Square: yellow (inert) Triangle: white (new)</td>
<td>Machine plays music</td>
<td>1. New square is unknown. 2. New circle and/or new triangle are blickets.</td>
<td>Outcome is uninformative (claim is true)</td>
<td>Outcome is informative (claim is true)</td>
</tr>
<tr>
<td></td>
<td>Machine does not play music</td>
<td>1. New square is a blicket. 2. New circle and new triangle are not blickets.</td>
<td>Outcome is informative (claim is true)</td>
<td>Outcome is informative (claim is true)</td>
</tr>
</tbody>
</table>

Method

The design, predictions, and analysis plan for Experiment 3 were preregistered prior to the start of data collection (see https://aspredicted.org/ud92n.pdf).

Participants

Participants for Experiment 3 (n = 51) were recruited from the University of California, San Diego undergraduate student participant pool and received academic credit. This experiment was conducted in the same academic session as Experiment 2, so participants had the same overall demographics. Participants were randomly split into two conditions: the dependence condition (n = 27) and the independence condition (n = 24). Six additional participants were excluded and replaced answering one (n = 2) or more (n = 4) of the comprehension questions incorrectly.

Materials

The task was presented entirely asynchronously via the Qualtrics online survey platform 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://osf.io/3dpsri/).

Procedure

The procedure was identical to that used with adults in Experiment 1 until the point at which the character offered their hypothesis. In Experiment 1, the hypothesis “Ari thinks that only the new square is a blicket, and that the new circle and the new triangle are not blickets” was visualized by a thought bubble showing the white square block surrounded by music notes appearing above Ari. In Experiment 3, this was accompanied by two additional thought bubbles, each showing one of the two other new blocks (the white circle and the white triangle) surrounded by x-marks. Following this, the target claim was indicated. The narrator said, “Let’s try to find out if Ari is right about the [square/circle and triangle]!” and the thought bubbles for the nontarget blocks were grayed out. The test options were presented on the following screen, with only the thought bubble(s) showing the target claim visible (e.g., in the independence condition, the circle and triangle were shown, paired with the test question, “What should Ari do to find out if it’s true that the new circle and triangle are not blickets?”).

Participants were given a forced choice between two options: “Ari can try the toy again, still using the new square, but this time using the circle and triangle that are not blickets” (replace-noncausal) and “Ari can try the toy again, but using the square that is not a blicket instead of the new square, and still use the new circle and new triangle” (replace-causal). The order of these options was counterbalanced across participants. To avoid the possibility of participants treating a choice between two potentially informative experiments as a trick question, written instructions told participants, “Look carefully! Both options will tell Ari something about the blocks, but which one is the best one to make sure they find out whether the [Square is a blicket/Circle and Triangle are blickets] or not?”

Comprehension Questions. In order to ensure that our comprehension questions were not inadvertently screening for above-average participants, we reduced and simplified the answer options on the multiple-choice questions from Experiment 1. As before, each question asked participants to indicate, “What would happen if Ari put these blocks into the toy?” for one set of three blocks (yellow, blue, or white). However, there were only three answer options: “We know the toy will not play music” (correct for yellow), “We know the toy will not play music” (correct for blue and white), or “We don’t know what the toy will do.” As before, participants who answered any of these questions incorrectly were excluded and replaced.

Results and Discussion

Collapsing across conditions, participants in Experiment 3 chose the “better” experiment (i.e., the option that was certain to produce conclusive evidence about the target claim) 82.35% of the time, significantly above chance, p < .001, 95% confidence interval (CI) [69.13, 91.6], two-tailed binomial. This preference was also significant within each condition: 88.89% of participants in the dependence condition selected replace-noncausal (p < .001, 95% CI [70.84, 97.65], two-tailed binomial) and 75% of participants in the independence condition selected replace-causal (p = .02, 95% CI [53.29, 90.23], two-tailed binomial). There was also no significant difference in the proportion of participants choosing the more conclusive option between the two conditions (p = .12, 95% CI [70.84, 97.65], two-tailed binomial).
As predicted, the majority of participants selected the answer option that offered the “better” experimental manipulation of the portion of the two-part hypothesis they were asked to investigate. This is particularly impressive since both options presented were unconfounded experiments (correctly replacing the target variables with variables known to be causally inert) and had at least some potential to generate informative evidence. Identifying the better experiment required not only recognizing that each manipulation could have multiple outcomes but also evaluating the potential conclusiveness of each one for the specific causal claim they had been directed to test. The results of this experiment suggest that learners readily recognize the distinct causal independence and dependence claims embedded within the classic CVS task design and understand the implications of this distinction for conducting informative experiments.

**Experiment 4**

In Experiment 4, we examine whether learners are spontaneously sensitive to the second counterintuitive element of the standard CVS assessment. Recall that in many prior investigations of CVS ability (e.g., Croker & Buchanan, 2011; Moeller et al., 2022; Tschirgi, 1980), the experiment options replace one or more of the original three variables with novel variables. Critically, participants have not previously observed how these variables behave, nor are they given any information about them by the experimenters. This task design therefore depends on participants assuming that these variables are causally inert and that their introduction has no influence on the final outcome. Without this assumption, all of the experiment options offered to participants violate a basic principle of controlled, informative experimentation and any effect of removing a variable could in truth be caused by introducing its replacement.

As noted above, the assumption that novel variables are causally inert contrasts with learners’ real-life experience of causal relationships. The world is a complicated place. Even a simple event is likely subject to the influence of a vast number of variables, many of which will be unknown to the learner. In order to conduct informative experiments in the world, learners—like scientists—must be acutely aware of potential confounds posed by novel variables. Indeed, even 5- to 7-year-olds recognize the potential for novel variables to be causes (Fernbach et al., 2012; Sobel et al., 2017). If participants apply these same intuitions to formal assessments of experimentation, they are unlikely to assume the novel variables introduced in each manipulation are inert.

In Experiment 3, we assess how learners interpret novel variables that are introduced in the context of the control of variables task. Do they assume these variables are inert controls as prior work as assumed? Or, do they spontaneously recognize their uncertain causal status? To test this, we created a version of the CVS task that asks participants to choose between experiments using either novel or known variables. As in previous experiments, participants were introduced to a character with a block toy and several blocks of different colors and learn the causal status of two these colors (blue “blickets” and yellow “nonblickets,” or “known-blocks”) during the introduction of the task. However, the pool of blocks was increased to six colors (white, yellow, blue, red, purple, and green), several of which (red, purple, and green “unknown-blocks”) were never observed on the machine and about which participants were not given any information.

Participants were tested on two forced-choice trials that asked them to choose between an experiment where the target variables were replaced with unknown-blocks and one where they were replaced with known-blocks. Since the unknown-blocks might be either blickets or nonblickets, the unknown-block option might produce either confounded or unconfounded evidence. That is, if this option results in the machine making music, the possibility that the unknown-blocks are blickets renders the experiment inconclusive. If it does not result in music, then the learner has conclusive evidence that none of the blocks used are blickets, which either proves the independence claim or refutes the dependence claim. Whether the known-block option was confounded differed between the two trials (see Table 3). In the known-informative trial, the known-block option replaced the target(s) variables with causally inert yellow blocks. In the known-confound trial, this option used blue blocks, which participants know cause the machine to play music.

If participants are sensitive to the uncertain causal status of novel variables, then they should treat the unknown-block option differently on each of these trials. Since it may produce confounded, inconclusive evidence, this option is a worse experiment than one that is definitely informative. However, it also has some chance of producing informative evidence, making it a better choice than a definitely confounded experiment. We predict that self-directed learners will approach CVS experiments in a way that is sensitive to this difference: avoiding the unknown-block option on the known-informative trial and selecting it on the known-confound trial. This pattern of choice behavior would suggest that learners spontaneously recognize and are sensitive to the ambiguity of novel variables, consistent with the claim that they approach the task from the perspective of a causal learner.

There are also several alternative possibilities that would not be consistent with the current proposal. First, participants may recognize this ambiguity but approach the task in a way that prioritizes tangible outcomes over information value. If this were the case (which seems unlikely given the results of Experiments 1 and 2), we would expect the opposite pattern of behavior: choosing the unknown-block option on the known-informative trial, and the known-block option on the known-confound trial. If, contrary to our proposal, learners do not spontaneously recognize or treat novel variables as potential confounds, then we would expect one of two patterns. A preference for the unknown-block option on the known-confound trial and no preference on the known-informative trial would support the classic assumption that participants treat the novel variables in CVS tasks as causally inert. This would suggest that they are interpreting the known-confound trial as a choice between a confounded and unconfounded test and the known-informative trial as a choice between two unconfounded tests. Finally, learners might make the opposite assumption and treat the unknown-block is causally efficacious. In this case, we would expect a preference for the known-block option on the known-informative trial and no preference on the known-confound trial.

**Method**

The design, predictions, and analysis plan for Experiment 4 were preregistered prior to the start of data collection (see https://aspredicted.org/xx7hd.pdf).

---

3 The introduction to the task clarified that these were the only two categories, so participants need not consider the possibility of novel variables being inhibitory causes.
Table 3
Task Design for Experiment 4

<table>
<thead>
<tr>
<th>Trial</th>
<th>Target block(s) replaced with</th>
<th>Possible tangible outcome(s)</th>
<th>Information value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known-confound</td>
<td>Red (unknown-block option)</td>
<td>Music</td>
<td>Not informative</td>
</tr>
<tr>
<td></td>
<td>Blue (known-block option)</td>
<td>No music</td>
<td>Informative</td>
</tr>
<tr>
<td>Known-informative</td>
<td>Red (unknown-block option)</td>
<td>Music</td>
<td>Not informative</td>
</tr>
<tr>
<td></td>
<td>Yellow (known-block option)</td>
<td>No music</td>
<td>Informative</td>
</tr>
</tbody>
</table>

Participants

Participants for Experiment 4 (n = 52) were recruited from the University of California, San Diego undergraduate student participant pool and received academic credit. This experiment was conducted in the same academic session as Experiment 2, and participants had the same overall demographics. A sample size of 48 was preregistered and increased because of the nature of timing online data collection. Four additional participants were excluded and replaced answering incorrectly on one (n = 2) or more (n = 2) comprehension questions.

Materials

The task was presented entirely asynchronously via the Qualtrics online survey platform 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://osf.io/3dpsr/).

Procedure

The only difference from the procedure in Experiment 1 was the additional novel blocks. On the first screen, five sets of blocks (one circle, square, and triangle, in sets of blue, yellow, purple, red, and green) were shown next to Ari and the toy. At test, the blue and yellow sets moved from this collection to positions at the top of the screen, while the red, green, and purple remained in their initial position, while the experiment options were presented.

As in all previous experiments, participants heard the character offer a hypothesis and selected an experiment to test it. To avoid confusion with the novel variables, the narration in this experiment referred to the white blocks as “colorless” (e.g., “Ari thinks only the colorless square is a blicket”) rather than “new.” The unknown-block was one of the three unobserved colors and referred to by its color name (e.g., “Ari can try the toy again, still using the colorless square, but this time using the red circle and red triangle”).

Each participant completed two test trials with a forced-choice between a known-block and an unknown-block experiment. Since the critical dependent variable was the pattern of choice behavior across the known-informative and known-uninformative trials, both trials had to be either both hold-target manipulations or both vary-target manipulations. Given that there was no difference between hold- and vary-target questions on any of the previous trials, this was counterbalanced across participants and collapsed in analysis. Participants also completed the same set of comprehension questions used in Experiment 3.

Choice Behavior Coding. For analysis, we coded participants into categories (see Table 4) based on their choice behavior: choosing the unknown-block on the known-confound trial and the known-block on the known-informative trial (Category A), choosing the unknown-block on both trials (Category B), choosing the unknown-block on both trials (Category C), and choosing the known-block on the known-confound trial and the unknown-block on the known-informative trial (Category D). Note that only two of these categories correspond directly to possible patterns of behavior described above: If learners are sensitive to the uncertainty of novel variables in conducting CVS experiments, then we would expect the majority of participants to be in either Category A (if they prioritize information gain) or Category D (if they prioritize tangible outcomes). If, contrary to our predictions, learners treat the novel variables as though their causal status and potential to produce confounded experiments were known, then participants will be split across multiple categories. Participants assuming that novel variables are causally inert should result in a split between Category A and C, whereas assuming that novel variables are definitely confounded should result in a split between Category A and B.

Results and Discussion

The results of Experiment 4 were overwhelmingly consistent with our predictions for learners’ choice behavior, with the majority of participants (67.31%) falling into Category A, as compared to 15.38% in Category B, 5.77% in Category C, and 11.54% in Category D. In other words, a significant majority of participants selected the unknown-block option over a definitely confounded experiment but not over a definitely informative experiment, \( \chi^2(3, N = 52) = 50.61, p < .001 \). We find the same results when using Fisher’s exact test to account for low cell count (two-tailed, \( p < .001 \)). This pattern suggests that participants recognized the ambiguous causal status of the novel variable and its potential to confound an experiment and spontaneously considered this in choosing between CVS experiments.

We also find that the proportion of participants choosing the more informative experiment was significantly above chance in both the

Table 4
Response Categories for Experiment 4

<table>
<thead>
<tr>
<th>Response category</th>
<th>Choice—known-confound trial</th>
<th>Choice—known-informative trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Unknown-block</td>
<td>Known-block</td>
</tr>
<tr>
<td>B</td>
<td>Known-block</td>
<td>Known-block</td>
</tr>
<tr>
<td>C</td>
<td>Unknown-block</td>
<td>Unknown-block</td>
</tr>
<tr>
<td>D</td>
<td>Known-block</td>
<td>Unknown-block</td>
</tr>
</tbody>
</table>
known-confound and the known-informative trials. That is, a significant majority of participants selected a definitely informative experiment over the potentially confounded experiment with a novel variable (82.69% of choices, \( p < .001, 95\% \text{ CI } [69.67, 91.77] \), two-tailed binomial), but chose that same potentially confounded experiment over a definitely confounded, uninformative experiment (73.08% of choices, \( p = .001, 95\% \text{ CI } [58.98, 84.43] \), two-tailed binomial). Learners identified and selected the experiment most likely to be informative regardless of whether the ambiguous potential confound was pitted against an informative or uninformative test. However, they did show a stronger preference for the definitely informative test over the unknown-block (82.69%), than for the unknown-block over the definitely confounded test (73.08%). The difference is only marginally significant (\( p = .09 \), two-tailed binomial) but is consistent with the proposal that learners are approaching the task as an instance of real-world causal learning: They are reluctant to select an CVS experiment that introduces a novel variable (and potential confound), even when it is the only option with a chance of providing informative evidence.

**General Discussion**

The claim that self-directed learners struggle with control of variables tasks despite their precocious abilities in exploratory learning has remained largely unchallenged for over three decades. The current work is a significant departure from this dominant narrative. We argue that documented difficulties with CVS do not stem from a true “gap” between learners’ competence in informal and formal experimentation but from a conflict between the standard assessments and learners’ real-world intuitions across the life span. Over four experiments, we not only demonstrate children and adults’ untrained competence on version of the CVS task that does not conflict with their intuitions as causal learners but also provide direct evidence for how elements of the classic task conflict with their expectations.

First, in Experiment 1, we replicate the classic CVS task—presenting both children and adults with a choice between experiments to test a hypothesis about a multivariate causal system—while also correcting for the two conflicts between the original design and real-world causal learning. We address the issue of the two-part causal hypothesis, which claims both that an outcome is causally dependent on one variable and causally independent of the other two variables, by asking participants to choose between two experiments testing just one of these claims at a time. Critically, however, one of these two experiments had an uninformative, desirable outcome, while we also had the usual demonstration of the causal system to identify some classes of objects as causal and some as inert, allowing us to present CVS manipulations that did not introduce causally ambiguous novel variables. Thus, Experiment 1 presented the choice that Tschirgi (1980) intends to offer: between correct control of variables that is less likely to produce a desirable tangible outcome, and an alternative confounded test. We find that, overwhelmingly, both children and adults correctly selected the appropriate, informative CVS experiments in this task. This suggests that when an assessment of experimentation ability does not contradict their expectations as causal learners, untrained children and adults spontaneously select experiments based on the expected information value, rather than on the tangible value or by applying a general heuristic.

Importantly, prevailing explanations of the gap between formal and informal learning cannot account for learners’ success in our task. The majority of these prior explanations have appealed either to an immature understanding of scientific inquiry (e.g., Kuhn, 2002) or to an inability to execute that understanding (e.g., Lapidow & Walker, 2020) to account for errors in scientific reasoning. However, the structure and format of the task scenario and test questions used in the current study is identical to Tschirgi’s (1980). Presumably, therefore, it requires the same level of explicit coordination of theory and evidence and has the same execution demands as the original design. Thus, children and adults’ success in the current experiments cannot be explained away as the result of our use of a less rigorous assessment of scientific thinking. Instead, these findings are consistent with the claim that the original CVS task conflicts with participants’ causal intuitions, leading to an inaccurate estimation their scientific inquiry (Kosowski, 1996; Lapidow & Walker, 2022).

In Experiment 2, we address a potential criticism to the interpretation of our results as evidence of scientific reasoning due to the “knowledge-lean” nature of the stimuli. To do so, we replicate our initial results in a task using a domain and narrative drawn directly from Tschirgi (1980). The design retained the changes from Experiment 1 but presented in a story about a character combining three types of ingredients to bake a cake that could turn out more or less well. The increases in exclusions based on comprehension check failures and participants’ increased preference for tangible outcomes suggest that the more detailed, content-rich, and familiar stimuli did present a more demanding test of learners’ abilities. Despite this, the pattern of behavior was nearly identical to Experiment 1: The vast majority of learners selected the informative test, even though it was less likely to produce a desirable tangible outcome than the uninformative alternative option.

We find further support for our interpretation in Experiments 3 and 4 by providing direct evidence that learners spontaneously draw on their real-world causal intuitions when asked to execute the CVS in formal assessments. Each experiment addressed one of the two elements obscuring the interpretation of the standard CVS assessment: (a) the simultaneous presentation of two distinct causal claims and (b) the inclusion of novel variables. Experiment 3 shows that adult participants recognize the distinction between the independence and dependence claims embedded in the target hypothesis and understand that the potential informativeness of both “VARY” and “HOLD” options depends on which of the two claims one intends to test. Finally, in Experiment 4, we demonstrate that adult participants are also sensitive to the ambiguity of novel variables, selectively avoiding them when a truly unconfounded test is available and reluctantly selecting them when the alternative is a definitely confounded test.

Taken together, these findings respond to fundamental questions about the underlying character of self-directed learning. Over the past 20 years, there has been increasing evidence of even young learners’ competence in many aspects of formal scientific thinking, including hypothesis testing (e.g., Chen & Klahr, 1999; Köksal-Tuncer & Sodian, 2018), hypothesis-evidence relations (Koerber et al., 2005), reasoning about the informativeness of evidence (e.g., Köksal et al., 2021; Masnick & Morris, 2008), and even explicit understanding of the nature of science (Sodian et al., 2002). The current study not only extends this literature but also provides a novel explanation of the apparent inconsistency between this
evidence for early competence and the classic evidence of self-directed learners’ errors. We suggest that, across studies, participants are approaching inquiry and inference tasks with expectations and goals attuned to learning in the causal world. Whether this approach leads them to perform well or poorly, depends on whether or not these intuitions are appropriate for the task. For example, the experiments presented in many CVS tasks are only informative if the novel variables are presumed to be inert (e.g., Moeller et al., 2022), but learners have an expectation that untested variables could in fact be causal (Fernbach et al., 2012; Sobel et al., 2017). Here, we find that this expectation leads them to spontaneously avoid untested variables as potential confounds when selecting experiments. Future work is needed to establish whether this explanation also is consistent with apparent gaps observed in other scientific reasoning tasks.

Past researchers have sought to explain why self-directed learners sometimes behave in ways that are inconsistent with their characterization either as intuitive scientists or as poor scientists. Here, we depart from this well-trod narrative and instead seek a characterization that is consistent with all of self-directed learners’ documented behaviors. Given that we characterize learners as competent, our novel account is more closely aligned with the cognitive development side of the ‘gap.’ That is, our reexamination of the control of variables task suggests that at least some of learners’ apparent difficulties with scientific reasoning may result from the particular methods used to assess those skills. Critically, however, we do not aim to repeat prior claims that learners are intuitive scientists or as poor scientists thinking. In particular, our results highlight the importance of checking that the assumptions of empirical and educational assessments align with learners’ lived experience. By understanding, accommodating, and leveraging learners’ intuitive strengths, we can design more effective educational materials and more accurate tests for assessing scientific thinking over the course of development.

References


Received July 28, 2023
Revision received December 7, 2023
Accepted January 20, 2024