Journal of Educational Psychology

What Skills Related to the Control-of-Variables Strategy Need to Be Taught, and Who Gains Most? Differential Effects of a Training Intervention

Sonja Peteranderl, Peter Adriaan Edelsbrunner, Anne Deiglmayr, Ralph Schumacher, and Elsbeth Stern

Online First Publication, May 25, 2023. https://dx.doi.org/10.1037/edu0000799

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

ISSN: 0022-0663

https://doi.org/10.1037/edu0000799

What Skills Related to the Control-of-Variables Strategy Need to Be Taught, and Who Gains Most? Differential Effects of a Training Intervention

Sonja Peteranderl[1], Peter Adriaan Edelsbrunner[1], Anne Deiglmayr[2], Ralph Schumacher[1], and Elsbeth Stern[1]

[1] D-GESS, ETH Zurich

[2] Empirische Schul- und Unterrichtsforschung, University of Leipzig

Building on rich training literature, we examined which skills constituting the control-of-variables strategy (CVS) benefit from a comprehensive training, and which develop similarly during content-focused inquiry at ages 10–12. In addition, we examined whether prior knowledge, reasoning abilities, and reading comprehension explain variation in intervention effects. In a within-classroom, controlled field-experiment, half of $N = 618$ children from schools located in the German-speaking part of Switzerland were randomly assigned to a training on the CVS, and the other half to an active control group engaging in content-focused inquiry. Mixed-effects models revealed that the CVS training improved children's skills in planning controlled experiments and understanding the indeterminacy of confounded experiments, whereas it did not show specific effects on children's skills in identifying and interpreting controlled experiments. Children with better reasoning abilities and reading comprehension showed the strongest intervention effects on the more difficult skills. The general and differential effects of training remained mostly stable after a period of 6 months. More basic CVS skills seem to develop without targeted training, whereas more advanced ones benefit most from training that meets learners' preconditions.

Educational Impact and Implications Statement

Children's understanding of more difficult aspects of controlled experimentation benefits from training, particularly for those with better reasoning abilities and reading comprehension. These findings encourage teaching more advanced aspects of experimentation already in elementary school, which can have visible developmental impact on children's understanding. Classroom-based trainings should consider heterogeneity in children's cognitive abilities and be designed such that all children's understanding benefits, for example, through providing sufficient guidance and support through structured activities that induce cognitive activation.

Keywords: control-of-variables strategy, scientific reasoning, experimental study, elementary school, differential effects

Supplemental materials: https://doi.org/10.1037/edu0000799.supp

The abilities to design controlled experiments and draw valid conclusions from experimental outcomes are core competences, along with further experimentation skills such as formulating research questions and reasoning the conclusions. These competences are considered as a major goal in STEM (science, technology, engineering, and mathematics) education and are listed in the educational standards of many countries (e.g., D-EDK, 2016; National Research Council, 2012). A central strategy for conducting controlled experiments is described as the control-of-variables strategy (CVS; Chen & Klahr, 1999): In order to conduct a controlled experiment, the experimenter needs to keep all variables constant, "controlling" for their influence, except for one focal variable whose effects are being tested.

Sonja Peteranderl https://orcid.org/0000-0003-4116-3984

Sonja Peteranderl and Peter Adriaan Edelsbrunner contributed equally. We thank the physicist Dr. Brigitte Hänger-Surer for developing the training for the control group. We also thank all students, parents, and teachers for their participation and support.

Sonja Peteranderl served as lead for data curation, formal analysis, and investigation. Peter Adriaan Edelsbrunner contributed equally to supervision. Anne Deiglmayr served in a supporting role for supervision, writing–original draft, and writing–review and editing. Ralph Schumacher served in a supporting role for writing–review and editing. Sonja Peteranderl and Peter Adriaan Edelsbrunner contributed equally to methodology and visualization. Sonja Peteranderl, Peter Adriaan Edelsbrunner, and Elsbeth Stern contributed equally to writing–original draft and writing–review and editing.

The data are available at https://osf.io/4t7cf/.

The experiment materials are available at https://osf.io/4t7cf/.

Correspondence concerning this article should be addressed to Sonja Peteranderl, ETH Zurich, D-GESS, RZ H22, Clausiusstrasse 59, 8092 Zurich, Switzerland. Email: sonja.peteranderl@ifv.gess.ethz.ch

The CVS is the core strategy behind more advanced scientific reasoning skills such as complex problem-solving (Greiff et al., 2015), it is applicable both in scientific and everyday contexts (Song & Black, 1992), and it predicts science achievement beyond more general reasoning skills (Bryant et al., 2015). Consequently, the question arises how an understanding of the rationale behind the CVS, and the competence to put it into practice when designing and interpreting experiments, develop, and how these skills can be fostered by targeted training. This has been a major research topic in research on scientific reasoning and broader cognitive development that has provoked ongoing exchange among science educators and psychologists (Edelsbrunner et al., 2018; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016; Zimmerman, 2007).

Rich literature on CVS training interventions shows that core understanding of the CVS can be successfully improved by targeted educational interventions (Chen & Klahr, 1999; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016). A fact that has been widely overlooked in the CVS-training literature is that the CVS encompasses multiple subskills that differ in their cognitive affordances and developmental trajectories (Bullock et al., 2009; Schwichow et al., 2020). In addition, CVS is correlated with general and more specific cognitive abilities, such as logical reasoning and reading comprehension, which might explain variations in individuals' developmental trajectories and in the effects of trainings (Edelsbrunner et al., 2018; Wagensveld et al., 2015). This raises the question how CVS trainings building on the available literature benefit different CVS subskills, and how their benefits depend on learners' preconditions for learning such as their cognitive abilities.

In the present study, we systematically consider differential effects of a CVS training, considering both the differential difficulties of different CVS subskills to be learned, and the individual cognitive preconditions of the learners. Specifically, at ages 10–12, we examine which CVS subskills are affected by explicit training, and which skills develop during an active control intervention that is also inquiry-based, but does not involve explicit instruction of the CVS. In addition, we examine how learners' cognitive preconditions (prior knowledge, general reasoning, reading comprehension) affect their learning gains. The overall aim of this study is to examine which skills gain from a targeted training, which in the present study's context is when children are 10–12 years old, and to what extent the benefit of the training depends on learners' individual cognitive preconditions.

The Emergence of CVS From Preschool to Adolescence

There is a long tradition in studying the acquisition of the CVS in cognitive-developmental and educational psychology (Inhelder & Piaget, 1958; Koslowski, 1996; Kuhn & Phelps, 1982; Siegler & Liebert, 1975; for a review, see Zimmerman, 2007). Against the early assumption that children under the age of 12 are cognitively unable to develop CVS skills (Inhelder & Piaget, 1958), numerous studies have demonstrated younger children's capabilities in recognizing controlled experiments and interpreting experimental outcomes (Kuhn, 2002). Sodian et al. (1991) found that, although first and second graders showed problems in spontaneously generating causal hypotheses and planning controlled experiments, they nonetheless were able to identify conclusive tests and distinguish them from nonconclusive ones. Bullock and Ziegler (1999) delivered consistent findings on a task requiring the identification and planning of adequate tests of hypotheses and the relations between variables in third- to sixth-graders. Children had to choose between provided experimental designs that were either confounded or controlled for testing a given hypothesis. On average, by the age of approximately 8 years, children preferred unconfounded comparisons over confounded ones. Progress in mastering CVS in the course of elementary school was also found by Osterhaus et al. (2017), particularly for less complex tasks that encompass a limited number of easily identifiable variables.

Such findings indicate that, at least to some degree, understanding of the CVS is already in place prior to secondary education (Bullock & Ziegler, 1999; Koerber et al., 2015; Koerber & Osterhaus, 2019). They also show the importance of distinguishing between different subskills of CVS when examining its development. Despite the sometimes considerable skills in young children, their ability to apply the CVS is still fragile, for instance when results are not in accordance with their beliefs (Croker & Buchanan, 2011; Gopnik & Schulz, 2004; Sodian et al., 1991). Many children are able to generate hypotheses to confirm their existing beliefs, but they do not see the necessity to also seek counterevidence (Croker & Buchanan, 2011; Kuhn et al., 1995; Schauble, 1996). Moreover, although children's average CVS competence is constantly increasing during elementary school (Koerber et al., 2015), a considerable number of children do not develop CVS skills on their own (Zimmerman, 2000, 2007, for reviews). Mastery of CVS remains difficult even for some adults (Bullock et al., 2009). In addition, not all skills constituting the CVS develop concurrently. In the study by Bullock and Ziegler (1999), for example, already 8-year-olds preferred unconfounded comparisons over confounded ones, which Bullock and Ziegler referred to as choice tasks. In contrast, when children were asked to design their own experiments, which Bullock and Ziegler referred to as production tasks, even in sixth grade only about 40% planned controlled comparisons. Osterhaus et al. (2020) found that in accordance with Bullock and Ziegler (1999), third graders mostly succeeded in selecting a controlled experiment among multiple alternatives. When asked to interpret the effects of confounded comparisons, in accordance with findings of Kuhn et al. (1988), they still struggled with understanding the inherent indeterminacy of such designs. Overall, these results show that whereas some CVS skills and broader scientific reasoning skills develop during regular schooling, driven through informal and casual learning opportunities, other skills are less driven by these factors. More advanced skills such as planning controlled comparisons and understanding the indeterminacy of confounded comparisons appear to require more targeted learning opportunities.

In accordance with these findings, recent studies discuss that the CVS consists of multiple distinct subskills that are rarely systematically differentiated (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016). These subskills differ in their cognitive affordances and developmental patterns. Building on a commonly used definition of the CVS by Chen and Klahr (1999), Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016) provides a categorization of four different subskills of the CVS. Table 1 provides an overview of the four skills and how these have been labeled in related research.

The first skill is the *identification of* a controlled experiment from among multiple alternative designs. This skill has been shown to develop in many children at about ages 8–10 (e.g., Bullock & Ziegler, 1999). The second skill, *interpretation*, denotes the ability to draw the correct inference from a controlled design. This skill

Table 1
Overview of CVS Subskills, Their Assessment, and Results Regarding Their Age of Development in Recent Research

Subskill	Description	Comparable tasks for testing subskill	Age of development
Interpretation (IN)	Ability to draw the correct inference from a controlled design	CVS (IN) tasks (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016)	Development is relatively early (kindergarten), but not in all children (Koerber & Osterhaus, 2019; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow et al., 2020; Schwichow, Zimmerman, et al., 2016)
Identification (ID)	Ability to identify a controlled design among multiple alternative designs	Choice-tasks (Bullock & Ziegler, 1999); ramp task (Chen & Klahr, 1999); CVS (ID) tasks (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016)	Development at 8–10 years (Bullock & Ziegler, 1999; Sodian et al., 1991)
Planning	Ability to plan controlled experiments	Production-tasks (Bullock & Ziegler, 1999)	Development from early to middle adolescence, but not for all (Schwichow et al., 2020)
Understanding (UN)	Understanding of the interdeterminacy of confounded experiments	Data-interpretation task (Kuhn, 1988); understanding experiments UNEX (Osterhaus et al., 2015); CVS (UN) tasks (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016)	Development, if any, in adolescence (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow et al., 2020; Schwichow, Zimmerman, et al., 2016)

Note. CVS = control-of-variables strategy.

also develops relatively early in many, although not in all children (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow et al., 2020; Schwichow, Zimmerman, et al., 2016). The third skill, *planning* a controlled experiment, requires more active involvement from individuals in designing correctly controlled comparisons. Schwichow et al. (2020) found that many but not all individuals develop this skill during early- to mid-adolescence. Finally, *understanding* the indeterminacy of confounded experiments and thus the impossibility of drawing reliable inferences from such designs, appears to be the most demanding skill (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016). This skill is not developed in elementary school children who otherwise show some CVS proficiency (Osterhaus et al., 2020), and some individuals do not develop a thorough understanding of this aspect throughout adolescence (Schwichow et al., 2020). The first and fourth of the listed skills, identification of a controlled experiment and understanding the indeterminacy of confounded designs, not only relate to the CVS principle, but they can also be seen as part of the skills required to examine interactions.

Overall, psychometric and developmental investigations indicate that planning controlled experiments and understanding the indeterminacy of confounded designs are the two most difficult among these four skills, and also the last skills to develop (Osterhaus, et al., 2020; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow et al., 2020; Schwichow, Zimmerman, et al., 2016). This raises the question how to best train all subskills of the CVS in one comprehensive training, particularly those that appear to require educational support before children enter secondary school.

How to Train the CVS

Numerous intervention studies have investigated whether and how mastery of the CVS can be promoted through systematic training (e.g., Chen & Klahr, 1999; Dean & Kuhn, 2007; Klahr, 2005; Klahr & Nigam, 2004; Kuhn & Dean, 2005; Strand-Cary & Klahr, 2008). Some studies found that elementary school children can be successfully instructed in applying the CVS, particularly by direct instruction and by guided inquiry (Chen & Klahr, 1999, 2008; Klahr & Nigam, 2004; Matlen & Klahr, 2013; Schalk et al., 2019; Strand-Cary & Klahr, 2008). Direct instruction and inquiry activities can achieve sustainable learning gains and lead to transfer across time and tasks (Chase & Klahr, 2017; Lorch et al., 2010, 2014). This relates to meta-analytic evidence showing that guidance is pivotal in inquiry-based learning (Lazonder & Harmsen, 2016). Guidance at varying levels of specificity can help learners in obtaining information from experiments, and it can positively affect inquiry skills.

Although by far not all intervention studies succeed in training the CVS (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016), many successful approaches have been implemented within the last five decades. Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016) summarized the findings from 72 intervention studies and reported a mean overall effect size of $g = 0.61$. Already early studies found evidence for the assumption that students' understanding of the CVS is trainable (Case & Fry, 1973; Siegler et al., 1973). Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016) found that instructional interventions involving (a) demonstrations of good experimental designs and (b) cognitive conflict particularly benefit students' acquisition of CVS skills.

Although numerous studies have shown the trainability of the CVS, it appears that two, in our view important, aspects have been mostly neglected so far. First, although training studies typically encompass more than one CVS subskill (e.g., Chen & Klahr, 1999; Lorch et al., 2010), we are not aware of any studies evaluating training effects on all four subskills listed by Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016). We are also not aware of any literature indicating differential relations of these skills to school achievement and other outcome variables. However, as discussed, there is variation in the difficulty, as well as in the developmental trajectories of

these skills. These differences indicate that some CVS subskills might be better trainable in children than others. For example, whereas the skill of identifying controlled experiments might be relatively easy to master even for young children (Bullock & Ziegler, 1999, 2009), a training aimed at the more advanced understanding of the indeterminacy of confounded designs showed limited effectiveness with 6- to 7-year-olds (Case, 1974). In addition, even after thorough training, by far not all individuals succeed on the subskill of planning controlled comparisons (e.g., Chen & Klahr, 1999). These findings indicate that it is important to evaluate CVS training effects systematically with regard to different CVS subskills. In addition, there are usually substantial individual differences in learning gains, and in children's skills after training (Chen & Klahr, 1999; Lorch et al., 2014; Wagensveld et al., 2015). Such variation points to the relevance of individual preconditions that explain variation in training effects (Chen & Klahr, 1999; Wagensveld et al., 2015). For example, CVS is correlated with general cognitive abilities (Bullock & Ziegler, 2009; Edelsbrunner et al., 2018; Kuensting et al., 2013; Mayer, 2012; Mayer et al., 2014; van Schijndel et al., 2015; Veenman et al., 2014). In addition, correlations have been found between CVS and children's verbal skills (Siler et al., 2010; van der Graaf et al., 2018).

First efforts have been made to examine which individual preconditions might contribute to variation in the effects of CVS trainings. In a training study with sixth graders, Wagensveld et al. (2015) found that children's prior CVS knowledge, verbal reasoning, vocabulary, and reading comprehension predicted learning gains in a CVS training based on discovery learning, but not in a second group that received more direct instruction on the CVS. Possibly due to sample size limitations, Wagensveld et al. (2015) did not examine differential effects of the interventions, that is, they did not test whether effects of the condition differed for children with varying preconditions. In a study by Schalk et al. (2019), prior knowledge regarding the CVS predicted negatively how much third graders' CVS skills improved during another study that examined differential effects focused on variation across assessment instruments, rather than differential effects across learners (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016). Thus, although there are substantial individual differences in the development of the CVS as well as in learning gains and posttest achievement in trainings, we know little about the sources of these individual differences.

Research Questions and Design of the Present Study

Our overview of the literature showed that the emergence of scientific reasoning is a protracted process that is intertwined with general cognitive development and stimulated by learning opportunities (see also Edelsbrunner et al., 2022). At the end of elementary school, around the age of 10–12 years, some children master the CVS without having undergone any formal training, while others do not improve despite having received such training (Wagensveld et al., 2015; Zimmerman, 2007).

Building on this state of research, the goal of the present study is to examine the differential effects of a comprehensive CVS training that targets all four CVS skills described by Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016). We trained these skills at ages 10–12, which is a developmental period during which children undergo substantial transitions in their understanding of CVS (Bullock & Ziegler, 1999; Schwichow et al., 2020). This allows testing which subskills of the CVS benefit from targeted training, and which ones develop without it. In addition, we evaluated which cognitive preconditions determine how much children gain in the different subskills. There are various kinds of preconditions for learning that can affect the overall efficacy of educational interventions, such as cognitive, metacognitive, and affective/motivational factors (Tetzlaff et al., 2021). In this study, we focus on factors that are typically described as cognitive preconditions for learning (e.g., Grimm et al., 2023). Under this term, we examine students' prior knowledge about CVS, their general reasoning abilities, and reading comprehension. These three cognitive preconditions have been related to children's acquisition of the CVS in prior research (Edelsbrunner et al., 2022).

In a randomized experimental field trial employing a within-classroom design, we compared children's CVS skills after half of them had undergone a comprehensive CVS training (intervention group) while the other half had undergone inquiry-based science lessons that did not explicitly address the CVS (active control group). The strong control condition was meant to provide a learning environment in which students could engage in inquiry. A prior study showed that such inquiry can benefit students' understanding of the CVS (Schalk et al., 2019). The students in the control group did not receive explicit instruction on the CVS. This choice of conditions (intervention group undergoing the CVS training, and the active control group) allowed examining whether explicit instruction of the CVS benefits its development beyond more content-focused inquiry during which CVS could in principle also develop. The within-classroom randomization of participants prevented a confounding of potential training effects with effects of classroom-specific experience. As we were not aware of any pen-and-paper instruments for children that capture all four mentioned subskills of the CVS, we developed such a test for the purpose of this study. This newly developed CVS test, which allowed assessing all four CVS skills (*identification, interpretation, planning, understanding*), was applied as pre-, post-, and follow-up measures. This instrument allowed building an overall CVS score (cf. Schwichow et al., 2020) but also four scores for the four individual CVS subskills. In order to examine the sustainability of training effects, we assessed students not only shortly after the training, but also half a year later.

> *Research Question 1:* Does a comprehensive CVS training have a positive effect on students' overall CVS score, compared to an active control training?

We developed a comprehensive CVS training intervention that encompassed three lessons and built on the principles identified by Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016) as leading to the largest training effects. We, therefore, expected that this CVS training would induce large learning gains in comparison to the active control training, in which students did not receive explicit instruction on the CVS. In addition, sustainability of trainings often falls short of expectations (Bailey et al., 2017). At the same time, Bailey et al. (2017) found sustained intervention effects for naturally developing skills that can be further enhanced through targeted training. As CVS falls into this category, posttest effects,

once achieved, can be expected to remain rather stable. Out of these considerations, we derived the following hypothesis regarding the overall effects of the CVS training at posttest and follow-up.

Hypothesis 1: The intervention group will outperform the control group in the overall CVS score at posttest, as well as at follow-up 6 months after training.

Research Question 2: Are all four CVS skills (*identifying* controlled comparisons, *interpreting* controlled comparisons, *planning* controlled comparisons, and *understanding* the indeterminacy of confounded designs; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016) equally affected by the training?

The training aims to advance the full spectrum of skills, but starting points may differ depending on a child's prior knowledge of CVS. Some children may already enter the training with mastery of the easier subskills *interpretation* and *identification*, but struggle with *understanding* and *planning*, therefore particularly gaining on the latter skills. For other children, there might still be room for improvement in *interpretation* and *identification*, while *planning* and *understanding* are still out of reach, resulting in substantial effects for the first two skills only. These assumptions point toward differential effects, which we discuss next. Since the training explicitly targets all four skills, we assume that it will improve all of these to some extent in comparison to the control group.

Hypothesis 2: The intervention group will outperform the control group in each of the four CVS subskills in the posttest as well as in the follow-up test.

Research Question 3: How do students' individual cognitive preconditions affect the extent to which they gain from the CVS training?

In line with theoretical assumptions and empirical findings of previous studies (Chen & Klahr, 1999; Cruz Neri et al., 2021; Wagensveld et al., 2015), we expect children's (a) prior knowledge (i.e., their CVS skills measured at pretest), (b) reasoning abilities, and (c) reading comprehension to explain variation in differences between the intervention and the control group. Building on the large sample size of our study (over 600 learners), we can examine whether these variables interact with the effects of the two conditions for the overall CVS score, as well as for the four more specific CVS subskills.

Regarding prior knowledge, studies on the expertise-reversal effect have shown that learners with higher prior knowledge tend to benefit more from conditions with less guidance (e.g., Kalyuga, 2009). In this study, prior knowledge is referred to as CVS skills at pretest. For learners with lower prior knowledge of a specific CVS skill, instructional guidance might be helpful in pointing out the principles underlying the skill. For these learners, a strongly guided intervention condition should thus evoke stronger learning gains. As many elements of our CVS training were highly scaffolded, it might serve low prior-knowledge learners to a higher degree than high prior-knowledge learners. For children who have already developed some understanding of a specific CVS skill at pretest, the guidance they receive within the intervention condition might be unnecessary. These children might be better able to refine their already better-developed schemata of the relevant principles with less guidance (Kalyuga, 2009) while engaging in inquiry more freely within the control condition.

Hypothesis 3a: Benefits of the CVS training are more pronounced for learners with lower prior knowledge.

The second individual cognitive precondition, general reasoning ability, is understood as the ability to learn (Gottfredson & Lapan, 1997). Numerous studies have shown that this ability determines the degree to which an individual makes productive use of learning opportunities (e.g., Vaci et al., 2019; Ziegler et al., 2021). We expect that students with higher general reasoning abilities will better exploit the learning opportunities provided by the CVS training. These students might be better able to generalize the CVS principles from the multiple instances in which these are applied and explained within the training (Lotz et al., 2022).

Hypothesis 3b: Benefits of the CVS training are more pronounced for learners with better reasoning abilities.

Regarding reading comprehension, we undertook careful pilot testing with our CVS test, including cognitive interviews, to ensure that students' achievement did not depend on their reading comprehension to an unduly extent. Nevertheless, reading comprehension is substantially related to more general verbal abilities (Cain et al., 2004; Schroeder, 2011) which have been shown to be predictive of children's development of skills related to the CVS (van der Graaf et al., 2016) as well as to science learning (Cruz Neri et al., 2021). The better the children's verbal abilities, the better they might be able to follow the instructors' explanations, and engage in more effective verbalization as well as verbal interactions with their peers during collaborative inquiry (Van Boxtel et al., 2000). Better comprehension also helps integrating sentences into coherent mental representations, in turn fostering students' science literacy (Cruz Neri et al., 2021; Hall & Miro, 2016).

Hypothesis 3c: Benefits of the CVS training are more pronounced for learners with higher reading comprehension.

Method

Sample

Based on a power analysis simulating a multilevel model and differential effects of the intervention for children with varying preconditions, we aimed at a sample of about 40 school classes with about 700–800 children in order to reach a power of .85. We eventually recruited a sample of 38 classrooms ($n = 758$ children) that participated in the present study. All classrooms came from schools located in the German-speaking part of Switzerland. Twenty-nine of the 38 classrooms were fifth grade and nine were sixth grade. The reason for this grade distribution was that we started implementing the study in the second half of one school year and continued in the beginning of the next school year. The parents or guardians of each child gave informed consent. Ethical approval was provided by the first author's institution ethical review board. Half of the children in each class were randomly assigned to the intervention group, the other half to the control group. From the initial sample, we had to exclude 140 children due to missing parental consent (48), being absent during the CVS training (62), or being absent during the

whole experiment (11). The remaining sample for analysis consisted of $n = 618$ children ($M_{age} = 11.67$, $SD_{age} = 0.65$, 309 male, 309 female). Of these, $n = 318$ children received the CVS training, and $n = 300$ the active control training.

Procedure

We employed a test on experimentation skills (CVS test) and the tests on general reasoning ability and reading comprehension as group tests in the classrooms administered by trained research assistants. The CVS test was presented at three measurement points: 1–2 weeks before the training started (pretest), 1 week after the training was finished (posttest), and approximately 6 months after the training (follow-up test). The tests on reasoning abilities and reading comprehension were presented together with the pretest, and participants were asked for their age and gender.

Trainings

After the children had been randomly assigned to the intervention (CVS training) and the control group (active control training) within their classes, the respective training was implemented within 2 weeks. Both the intervention and the control groups received parallel trainings, each lasting three 45-min-lessons (the typical duration of a Swiss school lesson). In both trainings, different physical contents and materials were used (comparable to Chen & Klahr, 1999). For each class, the two trainings took place at the same time, in two separate rooms of the school. In all classes, the intervention group was taught by the first author, who is a trained elementary school teacher. The control trainings were implemented by children's regular class teachers. This allowed teaching students from both conditions at the same time in different rooms. Analysis provided in Appendix A and Table A1 indicate that students' learning gains within the control condition did not depend on the teachers. The classroom teachers were briefed in detail about the scientific contents and learning objectives of the active control training, and they received all necessary materials as well as a detailed lesson script (available from the online supplemental materials).

The CVS training applied in the intervention group was developed by the authors of this article under direction of the first author, based on the following rationales for training elements. We decided to split the three sessions of the training into 2 days for logistic reasons and to prevent exhaustion. We used probe questions after every task, based on the seminal training study by Chen and Klahr (1999). Additionally, as instructional elements, we implemented demonstration experiments and cognitive conflict in the first training session, two elements that have been found to support conceptual understanding (Lee & Byun, 2012) and to be particularly effective in CVS trainings (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016). Cognitive conflict can act as a cognitively activating instructional element that increases learners' engagement (e.g., Limón, 2001). Demonstration experiments can function as worked examples that allow reducing learners' cognitive load (Bichler et al., 2020) and focusing on explaining, discussing, and modeling (Grimm et al., 2023) the principles behind controlled and confounded experiments. In the second training session, we used worksheets containing confounded experimental designs, which has been shown to be effective in former CVS trainings (e.g., Lorch et al., 2014). In the last training session, we used hands-on tasks with marble runs including four variables with two levels, similar to the ramps-task which has been used in the effective training by Chen and Klahr (1999).

In the first session (45 min), experiments demonstrating confounded and controlled experiments were used with materials similar to those developed by Chen and Klahr (1999). Cognitive conflict was induced in the first demonstration experiment by showing the result of an experiment that was not consistent with the initially posed hypothesis. The unexpected outcome was discussed within the group before the next experiment was demonstrated. The terms "research question," "variable," and "influence" were introduced at the first session and their correct use demonstrated. These terms were introduced as part of supporting students' science literacy, as they are also part of Swiss Science curricula (D-EDK, 2016). During all training sessions, a controlled experiment was referred to as a "fair experiment" and a confounded experiment as an "unfair experiment."

The second session (45 min) was administered on the same day after a short break. The CVS was repeated by working with worksheets including confounded experiments (Lorch et al., 2014) in pairs (*identification* and interpretation of controlled experiments) and in whole-class discussions about the results and the probe question "Can you be sure?" (Chen & Klahr, 1999). In a second task, the children had to judge both controlled and confounded experiments and decide whether these experiments could be interpreted or not (promotion of the *understanding*-subskill).

In the third session (45 min), approximately one week later, the researcher brought marble runs, which allowed for a setup broadly similar to the classic ramp task introduced by Chen and Klahr (1999). This session started with a short repetition of the CVS principles and of relevant terms, and continued with a hands-on activity in small groups. Every group was equipped with a marble run, marbles, and a worksheet for documenting their experiments and findings regarding which variables (e.g., the steepness of the ramp, or the material of the marbles) affected how far the marbles would run. The worksheets were scaffolded, guiding children through the experimentation process (predict-observe-explain) as an effective way of guidance in inquiry-based learning (Lazonder & Harmsen, 2016). Afterward, all group-designed experiments were discussed before they were conducted (subskill *planning*) and their outcomes evaluated (subskill *interpretation*).

In all three sessions, principles of guided discovery learning (Hardy et al., 2006) were applied, aiming for an appropriate basis to enable children to connect inquiry-based content knowledge to explicitly instructed domain-general experimentation skills. Across all three training sessions, all four CVS subskills were addressed: In worksheet tasks (session 2) as well as hands-on experimentation (session 3), the students were guided in *planning* and *interpreting* controlled experiments and practiced the *identification* of confounded and controlled experiments. In the demonstration experiments as well as when discussing students' own experimentation plans, the rationale underlying the CVS was explicated, and it was shown and discussed how no certain conclusions could be drawn from confounded experiments (*understanding*). Overall, all four subskills explicated by Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016) and assessed by our CVS test were part of the CVS training and the training was aimed at providing sufficient space for developing all of these skills. The training did not include tasks from the CVS test.

The Active Control Training was developed by a Physics educator for the purpose of this study. The control group underwent a curriculum on electrical circuits, which focused on content knowledge instead of scientific reasoning. The topic of electrical circuits was chosen to provide a control training in which students can obtain content knowledge, including declarative and procedural knowledge, about a relevant topic for science education. The topic further lends itself to systematic inquiry in the course of experimenting with electrical circuits. In contrast to the intervention condition, in which the CVS was exemplified across various topics (e.g., marble runs, funnels) to help students in generalizing the underlying principle, within this more content-focused condition the topical context remained the same throughout the three lessons. In the first session, the material and useful terms (e.g., "switch," "crocodile clamps") were introduced. The children learned about when an electrical circle is closed and why this is important. In the second session, the children learned about the use and application fields of "parallel circuits" and "serial circuits." They worked in pairs with worksheets by rebuilding them with given electrical material. Similar as in the intervention group (CVS training condition), the third session also involved hands-on problem-solving tasks that allowed systematic inquiry and experimentation (building circuits using physical elements such as wires, batteries, switches, and bulbs). The control training did not include guided experimentation, guided hypothesis testing via controlled experiments, or explicit instruction on the CVS.

Measures

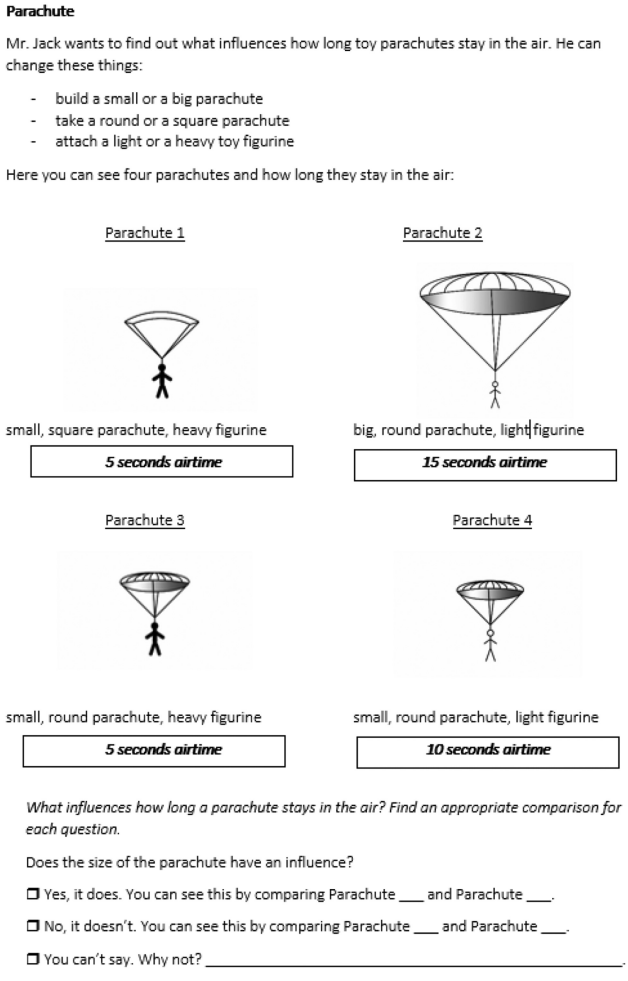
Control-of-Variables Strategy (CVS Test)

The newly developed CVS test encompassed 15 items, some of which were based on scenarios from existing tests (e.g., the airplane-task from Bullock & Ziegler, 1999 and the ramp-task from Chen & Klahr, 1999). Everyday concepts and topics from elementary school science education were used to build cover stories for the items. However, no specific scientific knowledge was needed in order to solve the test items correctly. This was ensured in cognitive interviews with children in pilot studies, and by involving experts in item construction who were science teachers with degrees and practicing educational research. All items have a multiple choice-format, and some items in addition ask for open answers. In pilot studies, construct validity was examined in cognitive interviews and by asking students for additional open answers (Peteranderl, 2019). Based on the information from these interviews and quantitative data collection in pilot classes, the test underwent multiple rounds of revision. The 15 items encompass the four skills *identification*, *interpretation*, and *planning* of controlled comparisons, as well as *understanding* the indeterminacy of confounded comparisons.

The skills *interpretation* and *identification* were assessed within the same 10 cover stories. An example item (translated into English by the study authors) is depicted in Figure 1. The children receive descriptions of three input variables (with two levels each) that could have an influence on one outcome variable. Four pictures show the setup and results of four tests, each realizing a different combination of levels of the input variables. The child's task is to identify two tests resulting in a controlled comparison (*identification*) and to find out whether and in which direction the input variables affect the outcome variable of the chosen comparison (*interpretation*). To solve the *identification* task, the children need to select a controlled comparison between two tests, that is, a comparison in which only the focal variable given in the research question differs between the two setups (*identification*). The children receive credit for selecting an unconfounded comparison involving the correct focal variable. In addition, the children need to interpret this comparison with regard to the research question by deciding whether the respective variable has an influence, or not (*interpretation*). The children receive credit for selecting the correct answer. Because of this procedure, *identification* and *interpretation* could be scored separately. For each of the skills, the maximum score was 10. Children's scores on *identification* showed estimated internal consistencies of McDonald's $\omega = .92$ at all measurement points, and their scores on *interpretation* $\omega = .91$, .91, and .85 at pre-, post-, and follow-up tests, respectively.

Figure 1

Example Item Assessing the Identification and Interpretation Skills



Note. In this item, the focal variable is the time to drop a parachute, independent variables are the weight of the toy figure, the form of the para chute, and the size of the parachute. The children must, given the provided evidence, decide for each individual independent variable whether it has an influence or not (interpretation) and select the correct comparison supporting their interpretation (identification).

The *planning* skill was assessed with four items. An example item is depicted in Figure 2. The items presented a research question and

Figure 2

Example Item Assessing the Skill to Plan a Controlled Experiment

Spaghetti

Sara and Matteo test how long you have to cook Spaghetti. They believe that it might make a difference

- whether they do or don't add salt to the water
- whether they cook thin or thick Spaghetti
- whether they use an electric cooker or a gas cooker
- how much Spaghetti they cook.

Sara and Matteo take the same pot, fresh water and fresh Spaghetti for every new test. They stop the time the Spaghetti need to cook.

First, they want to find out when it makes a difference if they *add salt* to the water.

How should their tests look like?

Test 1			Test 2		
water	❒with salt	❒without	water	❒with salt	❒without
thickness	❒thick	❒thin	thickness	❒thick	❒thin
cooker	❒electric	❒gas	cooker	❒electric	❒gas
amount	❒200g	❒500g	amount	❒200g	❒500g

Note. Children are asked to pick the levels of variables in order to produce a controlled comparison for testing the protagonist's hypothesis.

four independent variables (with two levels each) embedded in a cover story that involved protagonists wanting to test a causal hypothesis (e.g., that brooms with a longer broomstick are more effective for sweeping). The child was asked to design an experiment that was suitable for testing the protagonist's hypothesis. The child had to configure two test events (e.g., two broomsticks) by ticking checkboxes in a table that indicated the levels of the input variables they wanted to select for each test. To receive credit, the child had to produce a controlled comparison, that is, two tests (e.g., two brooms) that differed only on the focal variable (e.g., length of the broomstick). For this skill, the maximum score was 4. Children's *planning* scores showed estimated internal consistencies of $\omega = .92$, .95, and .95 at pre-, post-, and follow-up tests, respectively.

The tasks assessing *understanding* (five items) also presented a protagonist wanting to test a causal hypothesis. Figure 3 shows an example of an *understanding* item (translated into English by the authors). The setups presented in these items involved three input variables (with two levels each) and one outcome variable. In these stories, the protagonist had already conducted a comparison of two tests, which were presented to the child. The child's task was to select the answer (in a multiple choice format) that best described how the comparison could be interpreted. The protagonist's comparisons always had a confound (i.e., did not implement the CVS), thus, the correct answer always was that the protagonist could not be sure of what had caused the observed effect. For this skill, the maximum score was 5. Children's scores on *understanding* showed estimated internal consistencies of $\omega = .71$, .95, and .95 at pre-, post-, and follow-up tests, respectively.

As on overall indicator of children's CVS skills, we also computed an overall mean. Prior research has shown that in addition to scores on individual skills, an overall score of CVS represents a valid way to capture students' overall CVS skills (Schwichow et al., 2020). This was done by first transforming each of the scores for the four different skills into proportion correct-scores ranging from 0 to 1. We then computed a grand mean of these four correctness-scores to yield an overall indicator of CVS in which all four skills are represented equally, ranging again from 0 to 1. Children's overall CVS scores showed estimated internal consistencies of $\omega = .96$, .98, and .97 at pre-, post-, and follow-up tests, respectively.

Student's Cognitive Preconditions

Prior Knowledge

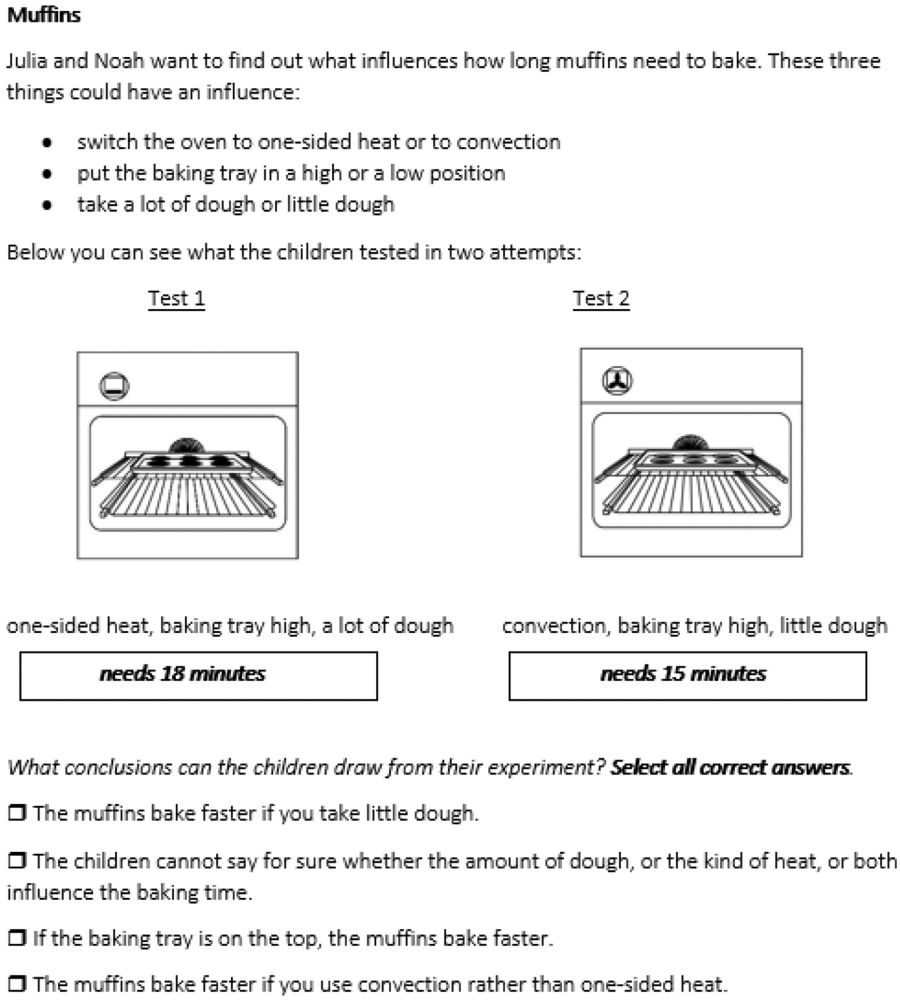
Students' prior knowledge on each of the four CVS subskills was assessed by their scores on the respective subskills scores at pretest. For example, in predicting overall CVS scores at posttest or follow-up, prior knowledge was represented in overall CVS scores at pretest, and in predicting scores on the understanding-subskill at posttest or follow-up, scores on this subskill at pretest were considered as prior knowledge.

Reasoning Abilities

Reasoning abilities were measured by the number series and figural analogies-scales of the Germany-wide established cognitive abilities test for primary school children (Kognitiver Fähigkeitstest [KFT]; Heller & Perleth, 2000). In the number series task (20 items), the children had to identify mathematical rules behind a number sequence and decide from among five alternative which number comes next according to the rule. In the figural analogies task

Figure 3

Example Item Assessing Children's Skill to Understand the Indeterminacy of Confounded Comparisons



(25 items), the children had to decide based on a base analogy (e.g., a larger square and a smaller square) which of five further objects yields the same relation together with a third presented object (e.g., a smaller circle yielding the same relation compared to a larger circle). Each student received a sum score built from all correctly answered items on both scales, which encompassed an overall of 45 items. The overall score showed an estimated internal consistency of $\omega = .92$ for the present sample.

Reading Comprehension

Students' reading comprehension was assessed using the latest version of a test appropriate for this age group assessing reading speed and reading comprehension (Lesegeschwindigkeits- und -verständnistest [LGVT, 2. Auflage]; Schneider et al., 2007). Children's task is to read a text and fill in 47 blank spots with appropriate words by selecting from among three options for each spot within a maximum time of 6 min. The maximum score on this test was 52 and it showed an estimated of internal consistency of $\omega = .80$ for the present sample.

Statistical Analyses

To examine our three research questions, we investigated descriptive statistics and set up multilevel models. For students' overall CVS scores, as well as for each of the four CVS skills, one model was set up with posttest scores (transformed into solution rates ranging from zero to one) as a dependent variable, and a second model with follow-up test scores as a dependent variable. The first predictor variable in each of the models was children's score on the respective skill at pretest. This variable was included to control for children's prior knowledge. In order to improve the interpretability of model results, pretest scores were *z*-standardized in all models. The second predictor variable was condition (experimental vs. control group). Inclusion of this variable in the model predicting children's overall CVS scores covered Research Question 1 (training effects on overall CVS skills), and including it to predict either of the four individual CVS skills covered Research Question 2 (training effects on four individual skills). Students in the control group were indicated by a 0 and those in the intervention group by a 1, implying that all reported model

intercepts and main effects represent estimates for the control group. The third and fourth predictor variables were students' *z*-standardized scores on reasoning abilities and reading comprehension. Through *z*-standardization, model intercepts represent estimates when all continuous predictor variables are at their mean. Effects of *z*-standardized predictor variables can be interpreted as change in the dependent variable when the predictor variable changes by one *SD*. To cover Research Question 3 (differential effects for students with varying individual characteristics), we included main effects and interactions with condition for prior knowledge, reasoning abilities, and reading comprehension. Since condition was coded with 0 for the control group and 1 for the intervention group, intervention effects indicate how much the main effect of the respective predictor variable (e.g., prior knowledge) changes when a child is in the intervention condition instead of the control condition. To cover the hierarchical data structure, we included a random intercept in all models to control for the multilevel structure caused by unexplained systematic variation between school classes. We used restricted maximum likelihood-estimation in the R package *lme4* to estimate these models (Bates et al., 2014).

For statistical inferences regarding the main effects of individual variables, we used a significance level of .05. For inferences regarding interaction effects (Research Question 3), we used a significance level of .10. It has been known since the seminal work by Cronbach and Snow (1977) that reliably finding interactions requires very large sample sizes. Since as described above, we had not fully reached the sample size that our power analysis had proposed, we increased the alpha error-level to optimize the beta-error and thus emphasize our aim not to overlook interaction effects. To visualize interaction effects, we produced Johnson–Neyman plots that show effects of the condition across levels of the moderator variable (Preacher & Sterba, 2019), as well as regions of statistical significance, via the R package *interactions* (Long, 2019). We report 90% confidence intervals for all model parameters.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All data, analysis code, and research materials are available at https://osf.io/4t7cf/. Data were analyzed using R, Version 4.1.1 (R Core Team, 2021) and the packages described above. This study's design and its analysis were not preregistered.

Results

Descriptive statistics of the overall CVS score and of students' individual cognitive preconditions are presented in Table 2. Descriptive statistics and intercorrelations of all further variables are presented in Tables A2–A4. Student age and gender had almost no associations with any other study variables (Table A4). We first checked the similarity of the intervention and control groups regarding prior knowledge, reasoning abilities, and reading comprehension. As visible from Table 3, independent sample *t*-tests did not indicate significant differences between the groups on any of these variables (all p's $> .5$), with Cohen's *d*s and their confidence intervals indicating high similarities. The mean estimates at pre- and posttest indicated that both groups increased in their overall CVS scores. Within-group comparisons between pretest and posttest served as manipulation check to find out whether the intervention affected CVS. The intervention group improved strongly from pre- to posttest by $d = 0.85$ ($p < .001$), but the control group also improved by $d = 0.41$ ($p < .001$). The considerable improvement of the control group has to be taken into consideration when interpreting the results. The results in Table 3 also show improvement from posttest to follow-up test in the intervention group ($d = 0.14$, $p < .02$) and in the control group ($d = 0.47$, $p < .001$).

Research Question 1: Does a Comprehensive CVS Training Have a Positive Effect on Students' Overall CVS Score, Compared to an Active Control Training?

The regression models (Table 3) indicate a main effect of condition after controlling for prior knowledge (i.e., students' pretest scores) and the other covariates, confirming that the CVS training indeed led to higher scores at posttest than the active control training. The standardized mean difference between the two conditions at posttest appeared moderate ($d = 0.39$, independent samples *t*-test: $p < .001$). At the follow-up, the effect of condition was also significant (Table 3), but with a smaller standardized group difference than at posttest ($d = 0.23$, independent samples *t*-test: $p < .001$). These findings confirm Hypothesis 1, which predicted higher achievement in the intervention group compared to the control group at posttest and follow-up.

Research Question 2: Are All Four CVS Skills Equally Affected by the Training?

Figure 4 depicts the mean achievement rates for each skill across time for both conditions. A manipulation check revealed that performance in all four skills was better after training than before (identification: $d = 0.51$, interpretation: $d = 0.33$, planning: $d = 0.87$, understanding: $d = 0.93$, all $p < .001$). This was also

Table 2
Descriptive Statistics for Overall CVS Scores and Individual Characteristics Across Conditions

	Control group		Intervention group		Group difference	
Variable	*M*	*SD*	*M*	*SD*	*d*	90% CI
Overall CVS pretest	0.44	0.23	0.43	0.24	−0.06	[−0.19; −0.08]
Overall CVS posttest	0.55	0.26	0.65	0.28	0.39	[0.26; 0.52]
Overall CVS follow-up	0.61	0.26	0.67	0.27	0.23	[0.10; 0.36]
Reading comprehension	17.45	10.36	18.00	10.59	0.05	[−0.08; 0.19]
Reasoning abilities	32.31	9.42	31.33	9.63	0.10	[−0.03; 0.24]

Note. CVS = control-of-variables strategy; CI = confidence interval.

Table 3
Linear Mixed-Effects Models Predicting Students' Overall CVS Scores at Posttest and Follow-Up

Parameter	Posttest			Follow-up		
	B	*t*	*p*	*B*	*t*	*p*
Intercept	0.54	58.72	<.001	0.61	61.42	<.001
Condition	0.11	9.23	<.001	0.07	5.26	<.001
Prior knowledge	0.19	15.78	<.001	0.17	13.24	<.001
Reasoning abilities	0.04	3.56	<.001	0.05	4.41	<.001
Reading comprehension	0.02	1.6	.111	0.02	1.93	.055
Condition × Prior Knowledge	−0.02	−1.51	.132	−0.05	−2.61	.010
Condition × Reasoning Abilities	0.02	2.02	.044	0.03	1.90	.058
Condition × Reading Comprehension	0.02	1.21	.229	0.00	0.17	.869

Note. Dependent variable was the mean solution rate across four skills ranging from 0 to 1. Condition: 0 = control, 1 = intervention, such that intercept stands for control. Prior knowledge (overall CVS scores at pretest), reasoning abilities, and reading comprehension were *z*-standardized. CVS = control-of-variables strategy.

true for the control group (identification $d = 0.44$, interpretation: $d = 0.34$, planning: $d = 0.23$, understanding: $d = 0.35$, all $p < .001$). As visible from Figure 4, in both groups additional small to moderate (Cohen, 1988) achievement gains appeared between posttest and follow-up, except for the planning skill in the intervention group.

To test whether the CVS training benefitted all four skills, we estimated individual mixed-effects regression models for each of the four skills (posttest results: Tables 4 and 5; follow-up results in Tables A5 and A6). There were no significant group differences for *identification* and *interpretation*. In both groups, performance increased from pretest to posttest, and from posttest to follow-up (Figure 4, Tables A5 and A6). The model estimates (Table 4), including narrow confidence intervals of the condition-effect around zero (Tables A7 and A8), indicated that *interpretation* and *identification* were not further improved by the CVS training. On the other hand, for the more advanced skills *understanding* and *planning*, statistically significant group differences emerged (Table 5), with substantially stronger gains in the intervention than in the control group. Overall, Hypothesis 2 was confirmed only for the two skills of *planning* and *understanding*.

Research Question 3: How Do Students' Individual Cognitive Preconditions Affect the Extent to Which They Gain From the CVS Training?

The impact of the individual cognitive preconditions was assessed by including the variables prior knowledge, reasoning abilities, and reading comprehension in linear mixed-effects Models. Originally, Research Question 3 and the respective hypotheses were limited to the overall CVS score. As the results of Research Question 2 showed different training effects for the subskills, separate analyses were included. For the overall posttest and follow-up score, the results of the linear Mixed-Effects Models are presented in Table 3. The posttest results for the subskills are presented in Table 4 (interpreting and interpretation) and Table 5 (understanding and planning). Results for follow-up scores are depicted in Tables A5 and A6.

Regarding simple main effects, overall, each of these individual preconditions had an impact on students' CVS scores at posttest and follow-up. Prior knowledge appeared to have the largest general estimated impact on the overall score and the subskills. At the same time, there was a significant independent contribution of reasoning abilities for all outcome measures. For reading comprehension, a

Figure 4
Solution Rates (±90% Confidence Intervals) on Overall Control-of-Variables Strategy (CVS) Score and the Four Different CVS Skills at the Three Measurement Points for the Intervention Group and the Control Group

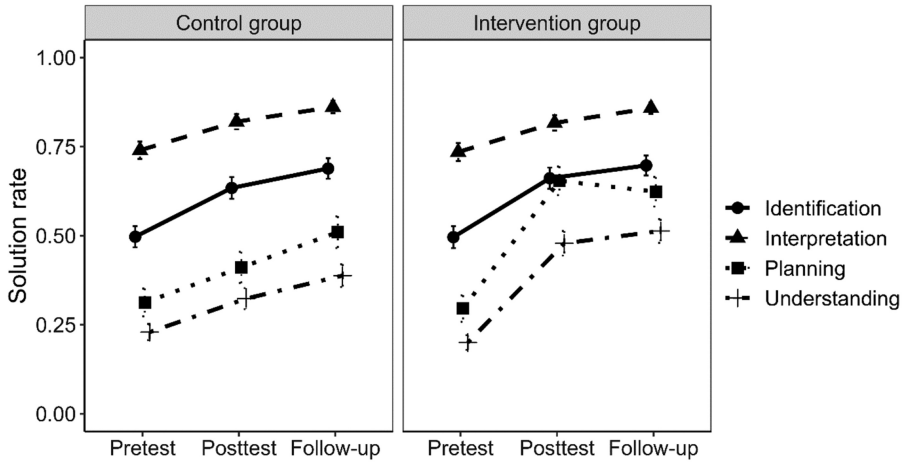


Table 4
Linear Mixed-Effects Models Predicting Students' Identification and Interpretation Scores at Posttest

Parameter	Identification			Interpretation		
	B	*t*	*p*	*B*	*t*	*p*
Intercept	0.64	53.01	<.001	0.82	77.68	<.001
Condition	0.02	1.28	.200	0	−0.12	.904
Prior knowledge	0.21	13.44	<.001	0.09	6.65	<.001
Reasoning abilities	0.05	3.75	<.001	0.04	3.66	<.001
Reading comprehension	0.01	0.83	.407	0.02	1.85	.064
Condition × Prior Knowledge	−0.03	−1.5	.135	0.04	2.05	.041
Condition × Reasoning Abilities	0.01	0.36	.720	−0.01	−0.49	.623
Condition × Reading Comprehension	0.03	1.4	.161	−0.01	−0.28	.780

Note. Dependent variable was solution rate ranging from 0 to 1. Condition: 0 = control, 1 = intervention, such that intercept stands for control. Prior knowledge, reasoning abilities, and reading comprehension were *z*-standardized.

statistically significant independent contribution was only found for the sub-skills of planning and understanding. Since the groups were dummy-coded, these estimates represent main effects in the control group.

The focus of Research Question 3 was to find out whether effects of the individual preconditions would differ between the two groups. This is visible from the estimated interactions between condition and the three individual variables presented in Tables 3–5 (follow-up results for the four individual skills presented in Tables A5 and A6). A positive interaction means that the impact of an individual precondition is more pronounced in the intervention group, and a negative interaction means that it is more pronounced in the control group. The exact interpretations of the interactions were figured out by modeling the slopes of the condition at different sections of the individual characteristics. For all significant interactions at our selected level of $p < .10$, the slopes modeled thereby are visualized in Figure 5 for the posttest results, and in Figure A1 for the follow-up results. In these figures, the slope of condition indicates the predicted difference between control and intervention group on the respective outcome variable, controlling for all other variables in the model. It is depicted on the y-axis. This condition effect is always shown across levels of the respective individual characteristic, presented on the x-axis.

For students' *overall CVS score*, we only found a positive interaction with reasoning abilities (Table 3). As visible from Figures 5A and A1, this interaction was positive, meaning that the effect of condition (the benefit of CVS training) was stronger for learners with better reasoning abilities (more to the right of the a-axis in Figure 5A). Note that for learners with lower reasoning abilities (more to the left of the x-axis), the condition effect was still positive (i.e., the 90% confidence band does not include 0), but much weaker. For the other individual characteristics, there were no significant interactions, although all parameter estimates were in the expected directions (i.e., negative for prior knowledge, positive for reading comprehension). Confidence intervals in Tables A7 and A8 further indicate that meaningful effect sizes for these nonsignificant interactions cannot be excluded based on the present data.

For students' skill in *identifying* controlled experiments, we did not find any significant interactions (Tables 3 and A5). The parameter estimates for the interactions of all three variables were in the expected directions (i.e., negative for prior knowledge, positive for reasoning abilities, and reading comprehension). The confidence intervals for reasoning abilities were close to 0 (Tables A7 and A8), whereas those for the other interactions did not exclude larger effects.

For students' skill in *interpreting* controlled experiments, we found a positive interaction with prior knowledge (Table 4, Figure 5B). For students with low prior knowledge (i.e., pretest scores more to the left of the x-axis), the condition showed a slightly negative effect. In this group, learners in the control group showed higher predicted interpretation-scores at posttest than those in the intervention group. This effect was not visible for students with prior knowledge-scores within about $-1SD$ and $+1SD$ from the mean. For these students, as indicated by the darker-shaded confidence band, there were no condition differences and students in

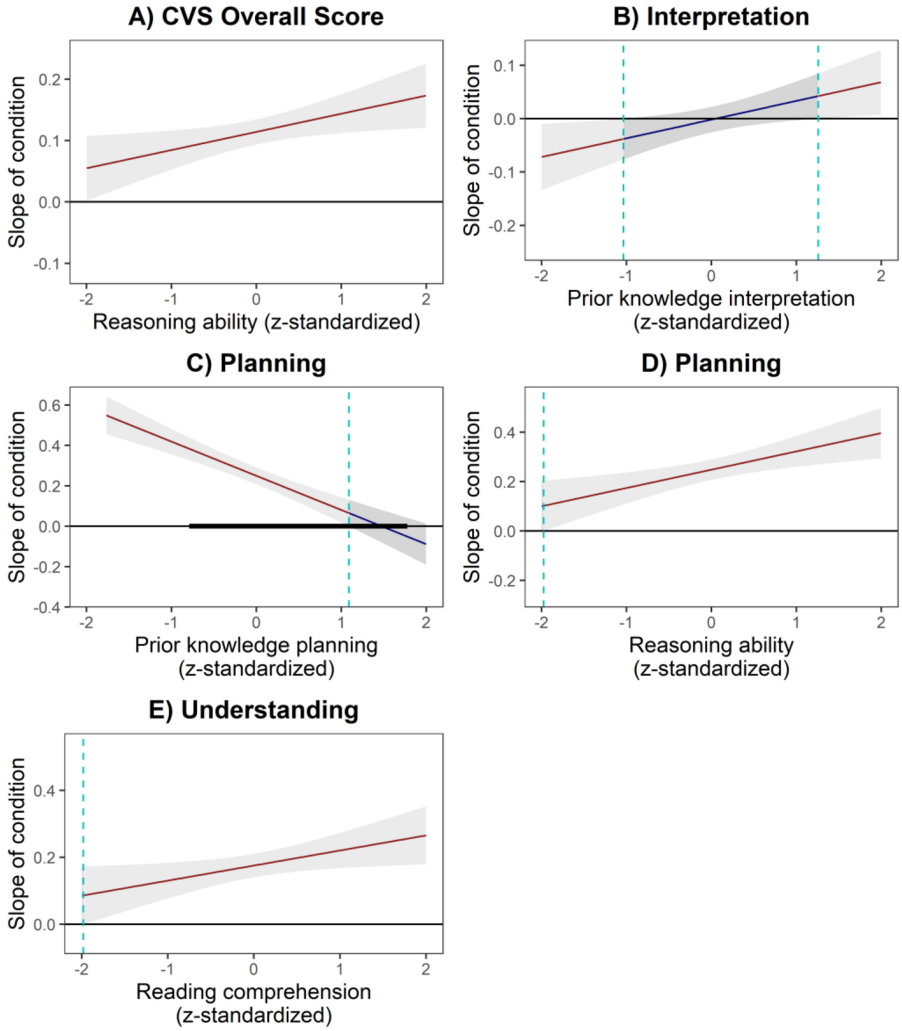
Table 5
Linear Mixed-Effects Models Predicting Students' Planning and Understanding Scores at Posttest

Parameter	Planning			Understanding		
	B	*t*	*p*	*B*	*t*	*p*
Intercept	0.41	22.51	<.001	0.314	20.24	<.001
Condition	0.25	9.91	<.001	0.176	8.28	<.001
Prior knowledge	0.29	14.3	<.001	0.144	8.81	<.001
Reasoning abilities	0.08	3.56	<.001	0.058	3.32	.001
Reading comprehension	0.05	2.41	.016	0.048	2.8	.006
Condition × Prior Knowledge	−0.17	−5.95	<.001	−0.004	−0.2	.842
Condition × Reasoning Abilities	0.08	2.6	.010	0.024	1.01	.315
Condition × Reading Comprehension	0.04	1.5	.134	0.046	1.87	.062

Note. Dependent variable was solution rate ranging from 0 to 1. Condition: 0 = control, 1 = intervention, such that intercept stands for control. Prior knowledge, reasoning abilities, and reading comprehension were *z*-standardized.

Figure 5

Visualizations of Interaction Effects and Their Regions of Statistical Significance



Note. Lines indicate regression weight (i.e., slope) of condition (intervention vs. control group) across levels of standardized moderator variables. Shaded area indicates 90% confidence bands. Dashed lines indicate points of moderator at which effect of condition becomes nonsignificant (areas with darker confidence band) or significant (areas with brighter confidence bands). CVS = control-of-variables strategy. See the online article for the color version of this figure.

both groups showed similar achievement at posttest. For learners with higher prior knowledge (more than 1*SD* above average, on the right of the a-axis), condition showed a positive effect, meaning that learners in the intervention group achieved higher learning outcomes than those in the control group. This is contrary to our Hypothesis 3a, in which we predicted lower condition-effects for learners with higher prior knowledge. This result inverted at follow-up (Table A5 and Figure A1). Specifically, at follow-up, there was a negative interaction, with students with lower prior knowledge showing positive effects of the condition, those with about average prior knowledge no significant effect, and those with high prior knowledge a negative effect of condition. This finding is in accordance with Hypothesis 3a.

For students' skill in *planning* controlled experiments, we found a negative interaction with prior knowledge, and a positive interaction with reasoning abilities. The negative interaction with prior knowledge (Table 5 and Figure 5C) indicated that whereas students with low prior knowledge benefitted substantially from the training, those with average prior knowledge showed a more moderate benefit, and those with high prior knowledge (more than +1 *SD* above mean) none. The positive interaction with reasoning abilities appeared similar to that of the overall CVS score. While even students with low reasoning abilities gained from the training, those with higher reasoning abilities had a much stronger benefit. The parameter estimate for reading comprehension was in the expected positive direction, although not significant. All estimates also were in the expected directions at follow-up, but only that of prior knowledge remained significant. For students' *understanding* of the indeterminacy of confounded designs, there was a positive interaction with reading comprehension that looks similar to the interactions

of training with reasoning abilities for the overall CVS and planning scores (Table 5 and Figure 5). While even students with low reading comprehension showed a positive effect of the training, those with better reading comprehension showed much larger benefits. The interactions of prior knowledge and reasoning abilities were also in the expected directions but not significant. At follow-up, the picture of reading comprehension and reasoning abilities reversed, with reasoning abilities showing a stronger positive interaction, and reading comprehension having a slightly lower and nonsignificant interaction estimate than at posttest. Prior knowledge showed an interaction very close to zero.

Overall, Hypothesis 3a, predicting a negative effect of prior knowledge on training benefit, was confirmed for the planning skill at the posttest as well as follow-up. For interpretation, we found a hypothesis-inconsistent positive interaction at posttest, but a hypothesis-consistent negative interaction at follow-up. For the identification and understanding-skills, we could not find interactions with prior knowledge, although their parameter estimates were negative both at posttest and at follow-up, with varying magnitudes. Hypothesis 3b, predicting a positive effect of prior reasoning ability on training benefit, was confirmed for the overall CVS score at posttest and follow-up, as well as for *planning* at posttest, and for *understanding* at follow-up. Hypothesis 3c, predicting a positive effect of reading comprehension on training benefit, could only be confirmed for *understanding* at posttest.

Discussion

The goal of the study was to further examine and explain the wide variation in CVS mastery during late childhood and adolescence. Developing CVS is a lengthy process stimulated by casual and informal learning opportunities that can be boosted by targeted trainings already at an early age (e.g., Chen & Klahr, 1999; Klahr et al., 2008; Klahr & Nigam, 2004; Kuhn, 2005; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016; Strand-Cary & Klahr, 2008). At the same time, difficulties in understanding experimental designs and drawing appropriate conclusions still arise in adolescence (Schwichow et al., 2020). In a randomized controlled design with an intervention group and an active control group, we investigated to what extent children at the end of elementary school (fifth and sixth grades) could gain from a targeted training directly after the training as well as 6 months later. The test we developed allowed to measure transferability of the CVS, as the situations described in the items were not referred to in the training. The test allowed for the analysis of overall performance in CVS as well as the four subskills that we focused on separately. By also including the individual cognitive preconditions of prior knowledge, general reasoning ability, and reading comprehension, this design allowed us to investigate differential training effects in more detail than previous studies had done.

Our first hypothesis, according to which the intervention group was expected to outperform the control group on the overall score of the CVS test in the posttest as well as in the follow-up test after 6 months, was confirmed. This is in line with prior studies showing positive effects for most CVS trainings (Chen & Klahr, 1999; Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016; Strand-Cary & Klahr, 2008). With an effect size of $d = 0.39$ between intervention and control group at posttest, our training triggered clear yet moderate effects (Cohen, 1988), which remained significant but decreased to $d = 0.23$ in the follow-up test. This moderate magnitude of the effect might appear surprising, given that in our training we kept to best-practices by implementing measures of cognitive conflict and demonstration experiments (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016), introduction of useful scientific terms (van der Graaf et al., 2016; Wagensveld et al., 2015), scaffolded worksheets (Lazonder & Harmsen, 2016), and hands-on activities (Chen & Klahr, 1999).

The moderate effect size can be explained by looking into the results regarding the second hypothesis, according to which we expected an advantage of the intervention group on all fours CVS skills. However, we found such a difference on only two of the four skills which means that Hypothesis 2 did not fully apply, as there were no specific benefits of the training for the skills of identification and interpretation. The overall effect size thus constitutes itself of two null effects on the easier skills, and two large effect sizes on the more difficult skills. Former studies showed that a considerable part of children younger than 10 years already mastered the subskills of *identification* (Bullock & Ziegler, 1999) and *interpretation* (Koerber & Osterhaus, 2019; Schwichow et al., 2020). We nonetheless expected that part of the children aged 12 would still struggle with these skills and therefore would gain from a training. This was not the case, as the intervention group and the control group did not differ visibly at any time point. A ceiling effect can account for the results in *interpretation*, while there was still room for improvement for *identification*. Both groups improved similarly and reached the same level on these skills. It appears that within the control group, the engagement in inquiry and the stimulation by the pretest were sufficient for triggering effects.

On the other hand, *planning* and *understanding* clearly benefitted from the CVS training, as the achievement in the posttest and the follow-up test was more pronounced in the intervention group than in the control group. It appears that both, *planning* a conclusive experiment and *understanding* the indeterminacy of confounded experiments, which are the more difficult CVS skills (Schwichow et al., 2020), need guided instruction to develop even for children in fifth and sixth grades. De Van and Csapó (2021) and Schwichow et al. (2020) claimed that many students struggle to understand the indeterminacy of confounded experiments before entering higher secondary school. In our study, we could show that already 12-year-old students can learn to neglect drawing conclusions from confounded designs in items that are not related to the training contents. Although based on the information from our multiple choice-items we cannot be sure how far students' understanding in this regard goes, this is a notable transfer effect was preserved even half a year later.

Our results demonstrate the malleability of all CVS subskills in the age group around 12 years: To improve in the short term, the more difficult skills need comprehensive training, while the easier ones may need only occasional stimulation. The latter argument was confirmed by the remarkable improvement of the control group. Although the participants did not receive a targeted training on CVS, their performance increased from the pre- to the posttest in the overall score as well as in the CVS skills within 2–3 weeks. This change most likely goes beyond age-related development, an interpretation also confirmed by the fact that the increase from pre- to posttest was stronger than from posttest to follow-up despite the

greater time gap (6 months) of the latter. Children from the active control group may have been stimulated to think about experimental evidence and causal conclusions in new ways when working on the pretest (i.e., re-test effects). Earlier studies have found similar testing effects in control groups (e.g., Bohrmann, 2017). Moreover, the training on building and trouble-shooting electric circuits may have indirectly affected CVS skills. Although the control training did not focus on CVS, building electric circuits requires a systematic approach, which may have indirectly stimulated some CVS skills. This would be in line with interventions finding effects on children's CVS skills in more self-guided and implicit interventions (Schalk et al., 2019; Strand-Cary & Klahr, 2008).

Hypotheses 3a–c concerned the impact of individual cognitive preconditions on the acquisition of CVS with and without a targeted training. Overall performance in the posttest as well as in the follow-up test was best predicted by prior knowledge in CVS, but there was also a unique though smaller contribution of reasoning abilities, which is in line with findings from Wagensveld et al. (2015). Unexpectedly, reading comprehension only had an effect on the most difficult subskills of understanding and planning, a finding that will be discussed later.

Considering interactions between condition and individual characteristics allowed to shed light on differential benefits of the training. We predicted that benefits of the CVS training would be more pronounced for learners with lower prior knowledge (Hypothesis 3a), resulting in a negative interaction between group and prior knowledge. Children with higher prior knowledge were expected to improve without a targeted training. For the overall score, this could only be confirmed for the follow-up test, suggesting that without a targeted training the development of CVS is delayed. For the skill of *interpretation of controlled experiments*, there appeared to be interaction effects, but this skill showed ceiling effects particularly for those high on prior knowledge, undermining reilable conclusions. For the difficult subskill of *planning*, however, it could be confirmed that those with high prior knowledge could improve without training.

Overall, we would have expected stronger negative interaction effects of prior knowledge and training, given the wide-reaching evidence that learners with higher prior knowledge benefit less from strongly scaffolded interventions (e.g., Kalyuga, 2009). Whereas we could not find evidence for such an effect in our study, prior knowledge still showed the strongest main effect for all skills. This is in line with research showing the importance of prior knowledge for learning (Simonsmeier et al., 2022), although, as Simonsmeier et al. emphasize, a strong standardized regression weight does not yet indicate a positive relation between prior knowledge and learning gains. Follow-up analysis to our study might examine in more detail how prior knowledge and learning relate in CVS interventions, for example, by employing models allowing nonlinear interactions to bolster against ceiling effects (cf. Vaci et al., 2019; Ziegler et al., 2021). For the

For reasoning abilities, the predicted positive interaction (Hypothesis 3b) could be confirmed for both time-points, indicating that benefits of the CVS training were more pronounced for learners with higher reasoning abilities in the overall score. Considering the subskills, we found an ordinal interaction for planning in the posttest and for understanding in the follow-up test. This finding goes beyond the main effects shown by Wagensveld et al. (2015), showing that the effect of reasoning abilities can be harnessed to further improve learning outcomes on the more difficult CVS skills. Also students scoring at the lower end of the reasoning scale showed benefits from the training at posttest, even though this effect faded out 6 months later. Overall, these findings regarding reasoning abilities indicate that a training implementing cognitive conflict, demonstration experiments, and a phase of collaborative inquiry, benefits all, and in particular those with better reasoning abilities.

For reading comprehension, the predicted positive interaction (Hypothesis 3c) did not reach significance for the overall score, while there was a significant interaction effect only for children's understanding of the indeterminacy of confounded comparisons. This is probably the most difficult skill among the skills that we encompassed (Schwichow et al., 2020). The interaction indicates the important role of language in comprehending the role and logic of the CVS principle. This result goes beyond earlier findings by Wagensveld et al. (2015) who showed a main effect of reading comprehension for children's acquisition of the CVS, and it is in line with Siler et al. (2010), who found that verbal/deductive reasoning is a key variable in children's acquisition of the logic of CVS and in enabling far transfer. The interaction pattern that we observed indicates that even learners with weaker reading comprehension gained from the training, but those with stronger reading comprehension benefitted even more.

Limitations

In our study, we have encompassed all four CVS subskills that Schwichow, Christoph, et al. (2016), Schwichow, Croker, et al. (2016), and Schwichow, Zimmerman, et al. (2016) identified as making up CVS mastery. Although this goes beyond prior studies that we are aware of in its comprehensiveness, it would be informative to broaden the breadth and detail of CVS skills further when undertaking similar future studies. For example, instead of CVS in its most basic sense in which one tries to find out about the effects of one variable, transfer to more complex problem-solving in which the causal status of multiple variables is scrutinized one after another might by examined (Greiff et al., 2015). In addition, the interplay of CVS skills with further aspects of epistemic cognition (a recent label for scientific thinking and reasoning; Greene et al., 2016) such as children's epistemic beliefs, aims, values, and ideas should be examined in the course of interventions to see which broader role CVS skills take in the development of elaborate and comprehensive scientific reasoning (Chinn & Rinehart, 2016).

Our randomized experimental field trial employed as a within-classroom design allowed to study training effects under real learning conditions not confounded with effects of classroom-specific experience prior to our study. In such designs, the respective class teacher can only be appointed under one condition. We chose the control condition, while the intervention group was taught by the same teacher, whom the students did not know before. This design made it possible to compare the effects of the CVS training with realistic content-related science classes. This also meant that we know little in detail about what the children in the control group really learned. Instances of treatment diffusion may have happened because children might have talked to each other about what they had learned while divided into groups. Most importantly, the teacher used in the intervention group was a researcher with a teacher diploma. It remains to be seen, whether the training effects achieved by her would have been achieved to the same extent by in-service teachers.

Another limitation concerns the use of multiple choice tests, which was the downside of our large sample size. Letting children express their ideas in their own words might have provided more detailed insights into their understanding. Finally, although our interaction analyses were based on hypotheses and a relatively large sample, these should be replicated in future research to ensure that the derivations of implications for designs of trainings remain appropriate beyond the present sample.

Implications for CVS Trainings and Future Research

While former studies delivered evidence for considerable improvement of CVS at the latest from the beginning of elementary school, our study emphasizes the need for further support at the end of elementary school, when children have reached the age of 12. With the exception of the interpretation skill, which already came close to a ceiling effect at pretest, all other CVS subskills showed substantial variance at all measurement points, indicating room for improvement. This was also the case for the relatively easy CVS subskill of *identification*, which was not affected by our training in comparison to the control group, although the average solution rate did not exceed .70 at any measurement point. Further research is required on how to support students who do not master this skill around the age of 12.

First indications regarding how to support the easier skills arise from our findings within the control group. The spontaneous improvements in the active control group showed that children gain from undergoing a comprehensive training as well as from getting casual opportunities to think about systematic experimentation. This finding is in accordance with prior studies finding that longer-term engagement with the CVS benefits its application even without targeted training (Schalk et al., 2019), and that the CVS can develop triggered by repeated exposure to test situations (Bohrmann, 2017). In order to optimally support children of all ability levels in all CVS subskills, both learning opportunities should be offered regularly in elementary school. To achieve some development on the easier subskills, triggering their application in repeated assessment or inquiry-situations appears to be sufficient. At the same time, a targeted training that directly addresses principles and fallacies of experimentation further boosts particularly the more difficult subskills *understanding* and *planning*. These prior and current findings corroborate the view that, on the one hand, certain aspects of the CVS develop spontaneously during engagement in inquiry, whereas this is less the case for more advanced aspects, which benefit substantially from targeted training.

An important insight of our study is that generalized statements about the trainability of CVS need to be put in perspective regarding aspects such as the trained age group, the exact CVS skills trained, and the learners' individual cognitive preconditions. Thus, it is of utmost relevance to meet children in their diversity of cognitive preconditions and create appropriate learning opportunities already in elementary school. Our study showed that also children with weaker general reasoning abilities benefitted from the intervention condition in comparison to the control condition. At the same time, the training achieved even stronger effects for those with better general reasoning abilities. While our training managed quite well to support children with varying cognitive preconditions, there is a risk that the achievement gap will further increase. Combined efforts in further disentangling differential effects of trainings and finding instructional means to meet the affordances of all learners are required to improve instruction of scientific reasoning. This could include the implementation of scaffolding that helps children in acquiring the stepwise process involved in designing and interpreting controlled designs (Grimm et al., 2023). Furthermore, additional individual preconditions such as inhibition ability might be considered in exploring further factors that affect individual differences in the success of trainings (Grimm et al., 2023; Osterhaus et al., 2019; van der Graaf et al., 2016, 2018). We assume that in accordance with our results and prior literature, training elements such as demonstration experiments (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016), inducing cognitive conflict (Schwichow, Christoph, et al., 2016; Schwichow, Croker, et al., 2016; Schwichow, Zimmerman, et al., 2016), modeling design and reasoning-processes (Grimm et al., 2023), and providing verbal support (Studhalter et al., 2021; van der Graaf et al., 2019) will also benefit the acquisition of other scientific reasoning skills for students with varying preconditions.

Differential learning effects also concern, for example, the role of language during instruction. The interaction between reading comprehension and achievement in the subskill of understanding raises the question of which particular aspects of our training drew so much on children's language. We expect that following the logic during induction of cognitive conflict, during which the teacher shows a confounded experiment and emphasizes the reason for the inability to draw a conclusion from such situations, requires strong language-related skills. Research has shown that the verbal behavior of the teacher in such situations and during inquiry more generally affects children's learning gains (Mercer, 2013; Studhalter et al., 2021). Mercer (2013) found that children who are supported in their academic language use do strengthen their own scientific thinking. Strong linguistic abilities were also found to be supportive for better performance in CVS tasks (van der Graaf et al., 2016), and in transferring the CVS to new domains (Wagensveld et al., 2015). For future interventions, we suggest considering in detail how the teacher verbally structures the guided parts of the training, to find out how such trainings draw on children's verbal skills and to optimize trainings accordingly to improve learning gains for all children. This suggestion is in accordance with the finding of van der Graaf et al. (2019) that the combination of direct instruction of scientific reasoning with verbal support had a positive effect on the effectiveness of an inquiry-based lesson.

The findings of this study further suggest specific approaches to the implementation of experimentation in elementary school curricula. Specifically, in lower grades teachers could start with easier explanations about identifying controlled experiments and interpreting the results of controlled experiments. The focus in later classes should be on promoting more demanding CVS skills. This might lead to better domain-general experimentation skills overall, which are important in the context of STEM in later school curricula (e.g., secondary schools). In accordance with prior findings indicating effects of the CVS on STEM achievement even beyond general reasoning (Bryant et al., 2015), trainings might evoke long-term achievement gains. Future studies should monitor students' longer-term STEM achievement and aspirations to gauge such impact that goes beyond immediate training effects. Further, our findings support previous results (Osterhaus et al., 2017; van der Graaf et al., 2016) that cognitive preconditions affect CVS mastery in children. Thus, it is even more relevant to meet children in their diversity of

cognitive preconditions and create appropriate learning opportunities already in elementary school. Our training intervention managed to meet some cognitive preconditions rather well; future studies should build on the present findings to further refine classroom interventions that strengthen CVS as the basis for more advanced scientific reasoning.

References

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint https://arxiv.org/abs/1406.5823

Bichler, S., Schwaighofer, M., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2020). How working memory capacity and shifting matter for learning with worked examples—A replication study. *Journal of Educational Psychology*, *112*(7), 1320–1337. https://doi.org/10.1037/edu0000433

Bohrmann, M. (2017). *Zur Färderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht*. Logos.

Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2015). The importance of being able to deal with variables in learning science. *International Journal of Science and Mathematics Education*, *13*(1), 145–163. https://doi.org/10.1007/s10763-013-9469-x

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood* (pp. 183–208). Psychology Press.

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study* (pp. 173–197). Psychology Press.

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 38–54). Cambridge University Press.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31–42. https://doi.org/10.1037/0022-0663.96.1.31

Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, *6*(4), 544–574. https://doi.org/10.1016/0010-0285(74)90025-5

Case, R., & Fry, C. (1973). Evaluation of an attempt to teach scientific inquiry and criticism in a working class high school. *Journal of Research in Science Teaching*, *10*(2), 135–142. https://doi.org/10.1002/tea.3660100205

Chase, C. C., & Klahr, D. (2017). Invention versus direct instruction: For some content, it's a tie. *Journal of Science Education and Technology*, *26*(6), 582–596. https://doi.org/10.1007/s10956-017-9700-6

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Chen, Z., & Klahr, D. (2008). Remote transfer of scientific-reasoning and problem-solving strategies in children. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 36, pp. 419–470). Elsevier Academic Press. https://doi.org/10.1016/S0065-2407(08)00010-4

Chinn, C. A., & Rinehart, R. W. (2016). Epistemic cognition and philosophy: Developing a new framework for epistemic cognition. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 472–490). Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

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

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

Cruz Neri, N., Guill, K., & Retelsdorf, J. (2021). Language in science performance: Do good readers perform better? *European Journal of Psychology of Education*, *36*(1), 45–61. https://doi.org/10.1007/s10212-019-00453-5

Dean, D., & Kuhn, D. (2007). Direct instruction versus discovery: The long view. *Science Education*, *91*(3), 384–397. https://doi.org/10.1002/sce.20194

D-EDK. (2016). *Lehrplan 21—natur, mensch, gesellschaft bereinigte fassung vom 29.02.2016*. Deutschschweizer Erziehungsdirektoren-Konferenz. http://v-ef.lehrplan.ch/downloads.php

Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: A large-scale quantitative study in elementary school. *Learning and Individual Differences*, *66*, 38–53. https://doi.org/10.1016/j.lindif.2018.02.003

Edelsbrunner, P. A., Schumacher, R., & Stern, E. (2022). Children's scientific reasoning in light of general cognitive development. In O. Houdé & G. Borst (Eds.), *The Cambridge handbook of cognitive development* (pp. 585–605). Cambridge University Press.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Gopnik, A., & Schulz, L. E. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176. https://doi.org/10.1037/0012-1649.40.2.162

Gottfredson, L. S., & Lapan, R. T. (1997). Assessing gender-based circumscription of occupational aspirations. *Journal of Career Assessment*, *5*(4), 419–441. https://doi.org/10.1177/106907279700500404

Greene, J. A., Sandoval, W. A., & Bråten, I. (2016). An introduction to epistemic cognition. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 1–16). Routledge.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning*, *21*(3), 356–382. https://doi.org/10.1080/13546783.2014.989263

Grimm, H., Edelsbrunner, P. A., & Möller, K. (2023). Accommodating heterogeneity: The interaction of instructional scaffolding with student preconditions in the learning of hypothesis-based reasoning. *Instructional Science*, *51*(1), 103–133. https://doi.org/10.1007/s11251-022-09601-9

Hall, A., & Miro, D. (2016). A study of student engagement in project-based learning across multiple approaches to STEM education programs. *School Science and Mathematics*, *116*(6), 310–319. https://doi.org/10.1111/ssm.2016.116.issue-6

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology*, *98*(2), 307–326. https://doi.org/10.1037/0022-0663.98.2.307

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, revision: KFT 4-12+ r*. Beltz-Test.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence*. Basic Books. https://doi.org/10.1037/10034-000

Kalyuga, S. (2009). *The expertise reversal effect*. IGI Global. https://doi.org/10.4018/978-1-60566-048-6.ch003

Klahr, D. (2005). Early science instruction. *Psychological Science-Cambridge*, *16*(11), 871–873. https://doi.org/10.1111/j.1467-9280.2005.01629.x

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667. https://doi.org/10.1111/j.0956-7976.2004.00737.x

Klahr, D., Triona, L., Strand-Cary, M., Siler, S. (2008). Virtual versus physical materials in early science instruction: Transitioning to an autonomous tutor for experimental design. In J. Zumbach, N. Schwartz, T. Seufert, & L. Kester (Eds.), *Beyond knowledge: The legacy of competence* (pp. 163–172). Springer. https://doi.org/10.1007/978-1-4020-8827-8_23

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development*, *86*(1), 327–336. https://doi.org/10.1111/cdev.12298

Koerber, S., & Osterhaus, C. (2019). Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *Journal of Cognition and Development*, *20*(4), 510–533. https://doi.org/10.1080/15248372.2019.1620232

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

Kuensting, J., Kempf, J., & Wirth, J. (2013). Enhancing scientific discovery learning through metacognitive support. *Contemporary Educational Psychology*, *38*(4), 349–360. https://doi.org/10.1016/j.cedpsych.2013.07.001

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Blackwell Publishing.

Kuhn, D. (2005). *Education for thinking*. Harvard University Press.

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. Academic Press.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866–870. https://doi.org/10.1111/j.1467-9280.2005.01628.x

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60*(4), i–157. https://doi.org/10.2307/1166059

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 17, pp. 1–44). Academic Press. https://doi.org/10.1016/S0065-2407(08)60356-0

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366

Lee, G., & Byun, T. (2012). An explanation for the difficulty of leading conceptual change using a counterintuitive demonstration: The relationship between cognitive conflict and responses. *Research in Science Education*, *42*(5), 943–965. https://doi.org/10.1007/s11165-011-9234-5

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction*, *11*(4-5), 357–380. https://doi.org/10.1016/S0959-4752(00)00037-2

Long, J. A. (2019). *interactions: Comprehensive, user-friendly toolkit for probing interactions* (Version 1.1.0). R package. https://cran.r-project.org/package=interactions

Lorch, R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, *102*(1), 90–101. https://doi.org/10.1037/a0017972

Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology*, *106*(1), 18–35. https://doi.org/10.1037/a0034375

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2022). g's little helpers—VOTAT and NOTAT mediate the relation between intelligence and complex problem solving. *Intelligence*, *95*, Article 101685. https://doi.org/10.1016/j.intell.2022.101685

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634. https://doi.org/10.1007/s11251-012-9248-z

Mayer, D. (2012). *Die Modellierung des wissenschaftlichen Denkens im Grundschulalter* [Doctoral dissertation]. Ludwig-Maximilians-Universität.

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction*, *29*, 43–55. https://doi.org/10.1016/j.learninstruc.2013.07.005

Mercer, N. (2013). The social brain, language, and goal-directed collective thinking: A social conception of cognition and its implications for understanding how we think, teach, and learn. *Educational Psychologist*, *48*(3), 148–168. https://doi.org/10.1080/00461520.2013.804394

National Research Council. (2012). *A framework for k-12 science education: Practices, cross-cutting concepts, and core ideas*. National Academies Press.

Osterhaus, C., Koerber, S., & Sodian, B. (2015). Children's understanding of experimental contrast and experimental control: An inventory for primary school. *Frontline Learning Research*, *3*(4), 56–94. https://doi.org/10.14786/flr.v3i4.220

Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology*, *53*(3), 450–462. https://doi.org/10.1037/dev0000260

Osterhaus, C., Koerber, S., & Sodian, B. (2020). The Science-P Reasoning Inventory (SPR-I): Measuring emerging scientific-reasoning skills in primary school. *International Journal of Science Education*, *42*(7), 1087–1107. https://doi.org/10.1080/09500693.2020.1748251

Osterhaus, C., Magee, J., Saffran, A., & Alibali, M. W. (2019). Supporting successful interpretations of covariation data: Beneficial effects of variable symmetry and problem context. *Quarterly Journal of Experimental Psychology*, *72*(5), 994–1004. https://doi.org/10.1177/1747021818775909

Peteranderl, S. (2019). *Experimentation skills of primary school children* [Doctoral dissertation]. ETH Zurich.

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children*, *85*(2), 248–264. https://doi.org/10.1177/0014402918802803

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learning and Instruction*, *59*, 34–45. https://doi.org/10.1016/j.learninstruc.2018.09.006

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119. https://doi.org/10.1037/0012-1649.32.1.102

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2007). *Lgvt 6-12: Lesegeschwindigkeits- und-verständnistest für die Klassen 6-12*. Hogrefe Göttingen.

Schroeder, S. (2011). What readers have and do: Effects of students' verbal ability and reading time components on comprehension with and without text availability. *Journal of Educational Psychology*, *103*(4), 877–896. https://doi.org/10.1037/a0023731

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *International Journal of Science Education*, *38*(2), 216–237. https://doi.org/10.1080/09500693.2015.1137651

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. https://doi.org/10.1016/j.dr.2015.12.001

Schwichow, M., Osterhaus, C., & Edelsbrunner, P. A. (2020). The relation between the control-of-variables strategy and content knowledge in

physics in secondary school. *Contemporary Educational Psychology*, *63*, Article 101923. https://doi.org/10.1016/j.cedpsych.2020.101923

Schwichow, M., Zimmerman, C., Croker, S., & Härtig, H. (2016). What students learn from hands-on activities. *Journal of Research in Science Teaching*, *53*(7), 980–1002. https://doi.org/10.1002/tea.21320

Siegler, R. S., Liebert, D. E., & Liebert, R. M. (1973). Inhelder and Piaget's pendulum problem: Teaching preadolescents to act as scientists. *Developmental Psychology*, *9*(1), 97–101. https://doi.org/10.1037/h0035073

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology*, *11*(3), 401–402. https://doi.org/10.1037/h0076579

Siler, S., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010). Predictors of transfer of experimental design skills in elementary and middle school children. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems*. ITS 2010, Lecture notes in computer science, Vol. 6095. Springer. https://doi.org/10.1007/978-3-642-13437-1_20

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, *57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753–766. https://doi.org/10.2307/1131175

Song, J., & Black, P. J. (1992). The effects of concept requirements and task contexts on pupils' performance in control of variables. *International Journal of Science Education*, *14*(1), 83–93. https://doi.org/10.1080/0950069920140108

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, *23*(4), 488–511. https://doi.org/10.1016/j.cogdev.2008.09.005

Studhalter, U. T., Leuchter, M., Tettenborn, A., Elmer, A., Edelsbrunner, P. A., & Saalbach, H. (2021). Early science learning: The effects of teacher talk. *Learning and Instruction*, *71*, Article 101371. https://doi.org/10.1016/j.learninstruc.2020.101371

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, *33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w

Vaci, N., Edelsbrunner, P., Stern, E., Neubauer, A., Bilalić, M., & Grabner, R. H. (2019). The joint influence of intelligence and practice on skill development throughout the life span. *Proceedings of the National Academy of Sciences*, *116*(37), 18363–18369. https://doi.org/10.1073/pnas.1819086116

Van Boxtel, C., Van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction*, *10*(4), 311–330. https://doi.org/10.1016/S0959-4752(00)00002-5

Van der Graaf, J., Segers, E., & Verhoeven, L. (2016). Scientific reasoning in kindergarten: Cognitive factors in experimentation and evidence evaluation. *Learning and Individual Differences*, *49*, 190–200. https://doi.org/10.1016/j.lindif.2016.06.006

van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction*, *56*, 1–9. https://doi.org/10.1016/j.learninstruc.2018.03.005

Van der Graaf, J., van de Sande, E., Gijsel, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *International Journal of Science Education*, *41*(9), 1119–1138. https://doi.org/10.1080/09500693.2019.1594442

Van Schijndel, T. J. P., Visser, I., van Bers, B. M. C. W., & Raijmakers, M. E. J. (2015). Preschoolers perform more informative experiments after observing theory-violating evidence. *Journal of Experimental Child Psychology*, *131*, 104–119. https://doi.org/10.1016/j.jecp.2014.11.008

Van Vo, D., & Csapó, B. (2021). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: Evidence of validity and latent predictors of item difficulty. *International Journal of Science Education*, *43*(13), 2185–2205. https://doi.org/10.1080/09500693.2021.1957515

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences*, *29*, 123–130. https://doi.org/10.1016/j.lindif.2013.01.003

Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2015). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science*, *43*(3), 365–379. https://doi.org/10.1007/s11251-014-9334-5

Ziegler, R., Hedder, I. R., & Fischer, L. (2021). Evaluation of science communication: Current practices, challenges, and future implications. *Frontiers in Communication*, *6*, Article 669744. https://www.frontiersin.org/articles/10.3389/fcomm.2021.669744

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

(*Appendix follows*)

Appendix

Teacher-Dependence of Learning Gains Within Control Group

To examine how strongly gains in CVS within the control group depended on the teachers, we implemented a multilevel model within students from the control group. This model had a typical setup of a one-way repeated measures analysis of variance: The dependent variable was students' overall CVS scores at pre- and posttest, predicted by Time (pre- vs. posttest) and a random intercept across students (this term is typical for a RMANOVA; Field et al., 2012). We further added a random intercept as well as a random slope of time, across teachers to cover the multilevel structure. As the results from this model in Table A1 show, whereas there was substantial variation across students (visible in the random intercept across students), there was less variation in the intercept across teachers, and even much less so for the effect of time (i.e., the learning gains). A likelihood ratio test corroborated this impression, showing that the random effect of time across teachers was not significant ($p = .471$). These results show that learning gains were very similar across teachers within the control group.

Table A1
Results of Multilevel Regression, Modeling Dependence of CVS Gains Within Control Group on Teachers

Parameter	Estimate	*SE*	*t*	*p*
Intercept	0.45	0.02	22.97	<.001
Time	0.10	0.01	10.41	<.001
Random effect	σ			
Intercept_student	0.042			
Intercept_teacher	0.007			
Time_teacher	0.001			

Note. Intercept stands for pretest mean-score on CVS, time represents CVS learning gains (difference between pre- and posttest). Standard error estimates, *t*-, and *p*-values are not available for random effect terms. CVS = control-of-variables strategy.

Table A2
Distributional Information for Central Study Variables

Variable	*M*	*SD*	Skewness	Kurtosis
CVS pretest	0.50	0.24	−0.08	−0.96
CVS posttest	0.65	0.26	−0.43	−0.94
CVS follow-up	0.69	0.24	−0.55	−0.77
Ide pretest	4.96	3.15	−0.09	−1.21
Ide posttest	6.48	3.15	−0.60	−0.86
Ide follow-up	6.93	2.95	−0.80	−0.50
Int pretest	7.37	2.59	−1.16	0.78
Int posttest	8.18	2.25	−1.53	2.11
Int follow-up	8.60	1.75	−1.52	2.34
Pla pretest	1.22	1.59	0.82	−1.02
Pla posttest	2.14	1.83	−0.17	−1.82
Pla follow-up	2.27	1.81	−0.30	−1.75
Und pretest	1.07	1.13	1.12	1.25
Und posttest	2.02	1.72	0.52	−1.00
Und follow-up	2.26	1.74	0.30	−1.23
Reasoning	31.8	9.53	−0.8	−0.12
Reading	17.74	10.47	0.6	0.03

Note. CVS = control-of-variables strategy overall score; Ide = identification; Int = interpretation; Pla = planning; Und = understanding.

Table A3
Descriptive Statistics of Central Study Variables by Condition

		M		*SD*			
Variable	Time	IG	CG	IG	CG	*Omega* ω	
CVS overall	Pretest	0.43	0.44	0.24	0.23	0.96	
	Posttest	0.65	0.55	0.28	0.26	0.98	
	Follow-up	0.67	0.61	0.27	0.26	0.97	
Reading comprehension		18.00	17.45	10.59	10.36	0.80	
						Figural	Numeric
Reasoning abilities		31.33	32.31	9.63	9.42	.92	.88

Note. IG = intervention group; CG = control group; CVS = control-of-variables strategy.

(*Appendix continues*)

Table A4

Intercorrelations Between Study Variables at Pre-, Post-, and Follow-Up Tests

Variable	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Age	−.04	.04	.10	.10	.09	.02	.12	.08	.09	.04	.04	.10	.08	.08	−.01	.02	.13	.05
2. Gender	—	.06	.08	−.02	−.04	−.02	.02	−.03	.02	−.01	−.02	.08	−.01	.01	−.06	−.04	.04	.04
3. Reasoning		—	.40	.49	.38	.46	.37	.32	.54	.39	.50	.47	.40	.57	.41	.49	.49	.47
4. Reading			—	.55	.44	.49	.42	.36	.54	.37	.46	.45	.45	.51	.37	.42	.44	.42
5. CVS pretest				—	.74	.86	.80	.70	.78	.59	.68	.63	.66	.73	.53	.65	.61	.60
6. Int pretest					—	.66	.36	.39	.56	.59	.51	.40	.41	.52	.48	.53	.41	.37
7. Ide pretest						—	.52	.49	.72	.57	.73	.55	.55	.68	.51	.66	.55	.53
8. Pla pretest							—	.44	.63	.36	.48	.59	.56	.59	.36	.46	.55	.51
9. Und pretest								—	.49	.33	.38	.35	.53	.45	.29	.38	.35	.46
10. CVS posttest									—	.73	.85	.86	.80	.86	.57	.73	.75	.72
11. Int posttest										—	.72	.45	.43	.61	.60	.62	.47	.42
12. Ide posttest											—	.60	.54	.76	.59	.77	.62	.58
13. Pla posttest												—	.60	.73	.42	.54	.74	.60
14. Und posttest													—	.67	.35	.50	.56	.72
15. CVS follow-up														—	.73	.85	.87	.82
16. Int follow-up															—	.79	.46	.44
17. Ide follow-up																—	.60	.56
18. Pla follow-up																	—	.62

Note. CVS = control-of-variables strategy overall score; Int = interpretation; Ide = identification; Pla = planning; Und = understanding; gender: 0 = male, 1 = female.

(*Appendix continues*)

Table A5

Linear Mixed-Effects Models Predicting Students' Identification and Interpretation-Scores at Follow-Up

	Identification			Interpretation		
Parameter	*B*	*t*	*p*	*B*	*t*	*p*
Intercept	0.69	54.80	<.001	0.86	99.42	<.001
Condition	0.01	0.42	.677	0.00	0.06	.952
Prior knowledge	0.17	10.47	<.001	0.08	7.28	<.001
Reasoning abilities	0.07	4.31	<.001	0.04	3.67	<.001
Reading comprehension	0.03	1.77	.077	0.02	1.92	.056
Condition × Prior Knowledge	−0.02	−1.07	.283	−0.04	−2.47	.014
Condition × Reasoning Abilities	0.00	−0.14	.886	0.01	0.42	.675
Condition × Reading Comprehension	−0.01	−0.27	.788	0.01	0.40	.691

Table A6

Linear Mixed-Effects Models Predicting Students' Planning and Understanding-Scores at Follow-Up

	Planning			Understanding		
Parameter	*B*	*t*	*p*	*B*	*t*	*p*
Intercept	0.51	25.21	<.001	0.38	23.77	<.001
Condition	0.12	4.22	<.001	0.15	6.52	<.001
Prior knowledge	0.22	9.50	<.001	0.13	7.35	<.001
Reasoning abilities	0.11	4.89	<.001	0.09	4.79	.001
Reading comprehension	0.07	2.84	.005	0.06	3.35	.010
Condition × Prior Knowledge	−0.08	−2.53	<.012	−0.03	−1.17	.242
Condition × Reasoning Abilities	0.03	0.81	.421	0.04	1.67	.096
Condition × Reading Comprehension	0.02	0.46	.644	0.00	−0.04	.969

Table A7

Confidence Intervals (90% Lower and Upper Bounds) for Models Predicting Posttest-Scores

	Overall CVS		Identification		Interpretation		Planning		Understanding	
Parameter	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Intercept	0.51	0.54	0.62	0.66	0.80	0.84	0.38	0.44	0.29	0.34
Condition	0.08	0.12	−0.01	0.05	−0.03	0.02	0.21	0.29	0.14	0.21
Prior knowledge	0.15	0.19	0.18	0.23	0.06	0.11	0.26	0.33	0.12	0.17
Reasoning	0.02	0.05	0.03	0.08	0.02	0.06	0.04	0.11	0.03	0.09
Reading comprehension	0.00	0.03	−0.01	0.03	0.00	0.04	0.02	0.08	0.02	0.08
Condition × Prior Knowledge	−0.04	0.01	−0.06	0.00	0.01	0.06	−0.22	−0.12	−0.04	0.03
Condition × Reasoning	0.00	0.05	−0.03	0.04	−0.04	0.02	0.03	0.12	−0.02	0.06
Condition × Reading Comprehension	−0.01	0.04	−0.01	0.06	−0.03	0.02	−0.00	0.09	0.01	0.09

Note. CVS = control-of-variables strategy.

Table A8

Confidence Intervals (90% Lower and Upper Bounds) for Models Predicting Scores at Follow-Up

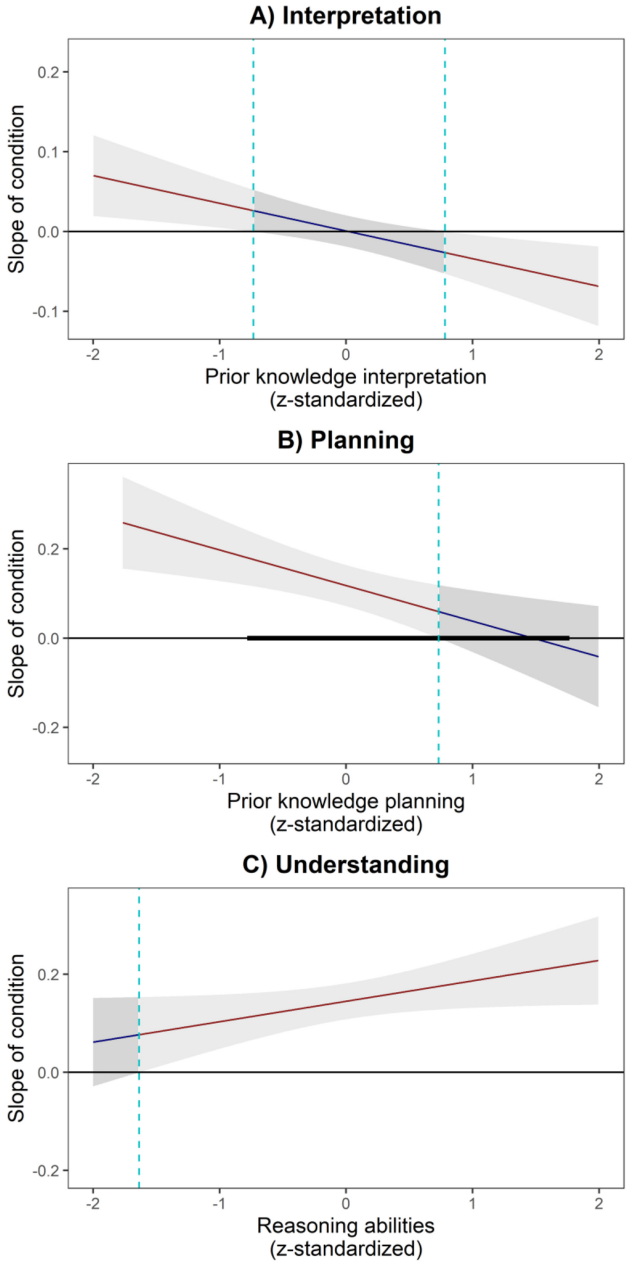
	Overall CVS		Identification		Interpretation		Planning		Understanding	
Parameter	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Intercept	0.59	0.62	0.70	0.71	0.85	0.87	0.48	0.54	0.35	0.41
Condition	0.05	0.09	−0.02	0.04	−0.02	0.02	0.07	0.16	0.11	0.18
Prior knowledge	0.15	0.19	0.14	0.19	0.06	0.09	0.18	0.25	0.10	0.15
Reasoning	0.03	0.07	0.04	0.09	0.02	0.05	0.08	0.15	0.06	0.12
Reading comprehension	0.00	0.04	0.00	0.05	0.00	0.04	0.03	0.10	0.03	0.09
Condition × Prior Knowledge	−0.07	−0.02	−0.06	0.01	−0.06	−0.01	−0.13	−0.03	−0.07	0.01
Condition × Reasoning	0.00	0.06	−0.04	0.03	−0.02	0.03	−0.03	0.08	0.00	0.08
Condition × Reading Comprehension	−0.02	0.03	−0.04	0.03	−0.02	0.03	−0.04	0.07	−0.04	0.04

Note. CVS = control-of-variables strategy.

(*Appendix continues*)

Figure A1

Depictions of Interaction Effects at Follow-Up



Note. See the online article for the color version of this figure.

Received August 15, 2022
Revision received December 8, 2022
Accepted January 23, 2023 ■