Improved Application of the Control-of-Variables Strategy as a Collateral Benefit of Inquiry-

Based Physics Education in Elementary School

Lennart Schalk[1,2], Peter A. Edelsbrunner[1], Anne Deiglmayr[1], Ralph Schumacher[1], Elsbeth Stern[1]

Affiliations:

[1]ETH Zurich, Zurich, Switzerland

[2]PH Schwyz, Goldau, Switzerland

*Corresponding author Email: lennart.schalk@ifv.gess.ethz.ch

Address: ETH Zurich, Clausiusstrasse 59, RZ H 23, 8092

Zurich, Switzerland

Phone: +41 44 632 55 67

Abstract

In a quasi-experimental classroom study, we longitudinally investigated whether inquiry-based, content-focused physics instruction improves students' ability to apply the control-of-variables strategy, a domain-general experimentation skill. Twelve third grade elementary school classes (Mdn_{age} = 9 years, N = 189) were randomly assigned to receive either four different physics curriculum units (intervention) or traditional instruction (control). Experiments were frequent elements in the physics units; however, there was no explicit instruction of the control-of-variables strategy or other experimentation skills. As intended, students in the intervention classes strongly increased their conceptual physics knowledge. More importantly, students in the intervention classes also showed stronger gains in their ability to apply the control-of-variables strategy correctly in novel situations compared to students in the control classes. Thus, a high dose of experimentation had the collateral benefit of improving the transfer of the control-of-variables strategy. The study complements lab-based studies with convergent findings obtained in real classrooms.

1. Introduction

Gaining competence in science requires learners to develop domain-specific content knowledge, as well as domain-general experimentation skills, across educational levels (National Research Council, 2012; Sandoval, Sodian, Koerber, & Wong, 2014). Laboratory studies have indicated that these two competence components can bootstrap one another (Schauble, 1990, 1996). We investigated whether an aspect of this mutual benefit can be exploited in real classroom instruction. Specifically, we implemented four basic physics curriculum units in elementary school classrooms. These units were designed to support the acquisition of conceptual content knowledge through numerous experimentation activities in a guided inquiry approach. All experiments were designed to allow for valid inferences (i.e., they instantiated the control-of-variables strategy). Students were guided through the process of setting up experiments, making predictions, performing the experiment, observing and recording data, and drawing conclusions. However, the underlying strategies of valid experimental design, particularly the control-of-variables strategy, were not explicitly taught, and the learners were not confronted with any violations of this strategy. We longitudinally investigated whether content-focused instruction for elementary school students has collateral benefits (through its strong reliance on valid experiments) for the development of their ability to apply the control-of-variables strategy in novel contexts.

1.1 Control-of-variables strategy and science education

The control-of-variables strategy (CVS) is a central domain-general principle of scientific reasoning. It specifies that causal data inferences obtained in an experiment can only be drawn if only one variable has been manipulated at a time (Strand-Cary & Klahr, 2008; Tschirgi, 1980). Understanding the CVS is necessary to generate and test causal hypotheses; that is, to design

conclusive and valid experiments and to critically evaluate the outcomes of experiments (D. Mayer, Sodian, Koerber, & Schwippert, 2014; National Research Council, 2012; Zimmerman, 2007). A first grasp of the CVS gradually emerges during childhood as a consequence of cognitive development and learning opportunities provided in school (Osterhaus, Koerber, & Sodian, 2017; Sandoval et al., 2014). Some kindergartners (van der Graaf, Segers, & Verhoeven, 2018) and first-graders (Sodian, Zaitchik, & Carey, 1991) can already recognize confounded hypothesis testing as being inappropriate. In elementary school, the ability to think scientifically, which includes the understanding of the CVS, constantly increases (Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015). Nevertheless, many secondary school students (and even adults) struggle when asked to evaluate and design conclusive experiments (Bullock, Sodian, & Koerber, 2009; Zimmerman, 2007). This issue is concerning because understanding the CVS is an important predictor of competence development in science (Bryant, Nunes, Hillier, Gilroy, & Barros, 2015).

Deliberate training can benefit students' understanding of the CVS. According to a recent meta-analysis (Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016), this training has typically been short-term interventions that focused on teaching the CVS or on teaching the CVS and additional content. Such explicit training is most effective if it includes demonstrations of valid and invalid (confounded) experiments and induces cognitive conflict (e.g., by challenging student conceptions with anomalous outcomes of a confounded experiment). Explicit training can also enable students to apply the CSV to new problems and in novel contexts (Chen & Klahr, 1999, 2008; Lorch Jr. et al., 2010; Lorch Jr. et al., 2014; Strand-Cary & Klahr, 2008). Importantly, by describing a training as "explicit", we do not maintain that it involves direct instruction of or lecturing about the CVS (e.g., a teacher explaining the logic of the CVS standing in front of the class). Rather, we use the term to distinguish previous trainings of the CVS in which the CVS was

the focus of instruction (e.g., by explicitly contrasting valid and invalid experiments or by providing explanations about the CVS in demonstration experiments) from our "implicit" training in the present study. That is, we investigated whether a student's ability to apply the CVS can implicitly benefit from a guided inquiry instruction designed to develop physics content knowledge.

1.2 Guided inquiry and abstraction of the CVS through structural alignment

With guided inquiry, we refer to instructional techniques that combine discovery learning with strong scaffolding from the teacher and the learning materials (Hmelo-Silver, Golan Duncan, & Chinn, 2007; R. E. Mayer, 2004). Guided inquiry is an effective instructional approach in education (Lazonder & Harmsen, 2016). Researchers have provided evidence for its benefits in domain-specific conceptual knowledge development in science education throughout preschool (Leuchter, Saalbach, & Hardy, 2014), elementary school (Hardy, Jonen, Möller, & Stern, 2006), and secondary school (Hanauer et al., 2006; Linn et al., 2014).

In guided inquiry-based instruction, students engage in active and self-directed exploration of complex phenomena and situations. For example, they create, test, and evaluate their own hypotheses in experimentation activities. However, this process is not discovery learning. Instead, the teacher and the instructional material provide guidance to direct the student's attention toward the learning goals. For example, the materials prompt students to write down the expectations, observations, and outcomes of the experiments. Teachers pre-plan and structure the experiments, provide hints if students struggle, and secure understanding by synthesizing and discussing the students' findings after the experimentation activities. This type of guidance is beneficial for acquiring conceptual content knowledge (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011).

Inquiry serves not only as a method of supporting understanding of domain-specific contents but also developing inquiry skills is itself an important instructional outcome (Abd-El-Khalick et al., 2004). In accordance with this conceptual duality of inquiry, guided inquiry may have collateral benefits beyond supporting content knowledge acquisition. Indeed, Schauble (1990, 1996) has demonstrated a mutual relation between the development of domain-specific scientific content knowledge and domain-general experimentation skills in small-scale lab-based experiments. Corroborated by intensive case studies, Schauble showed that scientific content knowledge benefitted the development of learners' understanding of experimental strategies, such as the CVS and their ability to apply these strategies, and that, in turn, strategies improved content knowledge development. Recently, Edelsbrunner, Schalk, Schumacher, and Stern (in press) have provided additional empirical support for one direction of this mutual relation. In their large-scale study on guided inquiry-based instruction, elementary school students' understanding of the CVS positively predicted conceptual change in the domain of floating and sinking. Specifically, a better understanding of the CVS increased the probability that students gained scientifically correct conceptual knowledge and decreased the prevalence of misconceptions. However, the other direction of the mutual relation (i.e., how content-focused guided inquiry instruction may benefit the application of the CVS) is less well understood.

We suggest that research on learning by structural alignment and analogical reasoning (e.g., Alfieri, Nokes-Malach, & Schunn, 2013; Gentner, 2010; Richland & Simms, 2015) might provide an explanation for how the ability to apply the CVS might benefit from content-focused guided inquiry instruction. When humans compare two or more situations or instances, the result can be abstraction. That is, learners create a knowledge representation that contains only the structural similarities between the two situations. This abstraction sets the stage for flexible application in

6

novel problems or contexts. In this sense, abstraction provides the basis for knowledge transfer (Chi & VanLehn, 2012; Gentner & Hoyos, 2017; Nokes-Malach & Mestre, 2013). Imagine a student who conducts several valid experiments (i.e., the experiments manipulate only one factor at a time) in various domains. Put differently, the student interacts with several instances of the CVS. If the student aligns these instances, this might support abstraction of the CVS because the CVS is a common structural feature across the valid experiments. Such spontaneous abstraction is rare in experimental laboratory settings with (rather) short interventions and a small number of examples; learners typically need specific scaffolds to abstract knowledge from few examples and to apply this knowledge in novel contexts (for overviews, see Gentner & Hoyos, 2017; Goldwater & Schalk, 2016). However, there is also evidence that analogical reasoning and abstraction are more frequent in naturalistic settings (e.g., Chan & Schunn, 2015; Dunbar, 2001) and when students encounter various examples over longer time periods, as in the studies by Schauble (1990, 1996). The reasons for these conflicting findings gained from laboratory studies and studies in naturalistic settings are not entirely clear. One plausible explanation is that naturalistic settings typically provide more opportunities and learning resources compared to the resources provided in laboratory studies (Hofer, Schumacher, Rubin, & Stern, in press). Therefore, we assumed that if students conduct many experiments over extended time in guided inquiry-based instruction, this experience might support them in structurally aligning the experiments, hence, in abstracting the CVS as a domain-general principle. If students abstract the CVS, it would improve their ability to apply it in novel contexts, that is, to transfer their knowledge.

1.3 The present study

We aimed to scale up one aspect of Schauble's findings (1990, 1996) to real classrooms. Specifically, we investigated whether the ability to apply the CVS increases as a collateral benefit of inquiry-based physics education in elementary school.

Beginning in 3rd grade, we implemented guided inquiry-based curriculum units to convey basic conceptual physics content knowledge. Crucially, understanding of the CVS increases at this age (Bullock & Ziegler, 1999). Thus, we could test whether and to what extent this development additionally benefits from the content-focused curriculum units. The units encompassed the broad topics *Floating & Sinking*, *Air & Atmospheric Pressure*, *Sound & Spreading of Sound*, and *Stability of Bridges*. Within each unit, students engaged in multiple guided experimentation activities designed to highlight to-be-learned physics concepts. All of these experiments were conclusive; they exemplified the CVS (see Appendix A for additional information on the curriculum and for examples of these activities). Thus, students enacted the CVS in their guided inquiry activities. However, they were never informed of this strategy. Thus, in contrast to studies that explicitly trained the CVS (e.g., Lorch Jr. et al., 2010; Lorch Jr. et al., 2014; Schwichow, Zimmerman, Croker, & Härtig, 2016; Strand-Cary & Klahr, 2008), none of the four curriculum units involved explicit instruction or specific training of the CVS.

In Switzerland, where the present study was conducted, teachers have a high degree of freedom in selecting topics related to science, history or geography when teaching the elementary school subject "Human Beings and their Environment". Thus, they can select topics from natural sciences, such as physics, they can focus on local geography (e.g., learning about areas of Switzerland, rivers, mountains, and so on) or the natural environment (e.g., learning about the native fauna and flora), or they can teach about the history of humankind (e.g., the Stone Age). Systematic analyses on classroom practice in Swiss elementary schools have revealed that teachers

rarely choose physics topics (Metzger & Schär, 2008). In their traditional instruction, they prefer topics dealing with local geography and natural environments. If the teachers choose physics topics, the topics are rarely instructed by implementing guided experimentation activities. Rather, teachers present phenomena, without delving into scientific reasoning and explanation.

We recruited elementary school teachers who were eager to teach physics but felt unable to do so without undergoing training. In a waiting list group design, teachers were randomly assigned to either the intervention or control condition. The teachers in the latter group were asked to continue teaching their classes in their traditional manner. Their training was postponed until the end of this study. Thereafter, they received the same training as teachers in the intervention classes.

We aimed to compare students who received instruction with the guided inquiry-based physics curriculum units in the subject "Human Beings and their Environment" (i.e., the intervention classes) to students who received their traditional instruction in this subject (i.e., the control classes). Given these different learning experiences, we expected a double advantage for the intervention classes. First, we expected them to show stronger gains in the content knowledge taught in the four physics curriculum units (a result to be considered as an implementation check because the control classes did not learn about these physics topics). Second, and this was the central hypothesis examined in this paper, we expected that the students in the intervention classes would also show stronger gains in their ability to apply the CVS to novel problems in novel contexts. This advantage would be a collateral benefit from the physics curricula if the strong dose of experimentation did indeed support the abstraction of the unifying principle underlying the experiments (i.e., the CVS).

2. Method

2.1 Sample and design

The participating students represent the first cohort of 3rd grade classes that joined the Swiss MINT Study. The Swiss MINT Study is a large-scale study in which early science instruction (as preparation for future learning) is examined longitudinally. The sample comprises students attending elementary schools in Zurich and surrounding German-speaking cantons of Switzerland. This area is densely populated, and approximately 80% of the elementary students' parents are Swiss nationals. The sample spans the full socioeconomic status range; however, the number of welfare recipients is generally low in Switzerland (approximately 3%). The research ethics committee of the [blinded] approved the study.

We only analyzed data from students whose parents provided written consent. In total, twelve 3rd grade elementary school classes with 189 students participated. A multilevel simulation study for this sample size indicated a statistical power above .90 for finding a small intervention effect (2% of total variance explained) on the level of school classes (see Appendix B for details of the power analysis). The median age of the students was 9 years (range, 8-11 years) at the beginning of the study. Of the 12 participating school classes, 6 randomly selected classes ($n = 81$, 37 girls) served as the intervention group, in which the teachers implemented the four physics curriculum units. The other 6 classes ($n = 108$, 58 girls) served as a waiting list control group. In these classes, the teachers continued with their traditional instruction. In the six intervention classes, there were 6, 7, 14, 14, 19, and 21 students per class. In the six control classes, there were 9, 10, 21, 22, 23, and 23 students per class. The number of participants and class sizes in the two conditions were unequal because not all parents provided consent and because class sizes generally vary in Switzerland.

2.2 Materials

2.2.1 Learning materials

In the intervention classes, the teachers implemented four early physics curriculum units developed by a team of science education experts at the University of Munster, Germany (Spectra Materials - KiNT-Boxes 1-4): *Floating & Sinking*, *Air & Atmospheric Pressure*, *Sound & Spreading of Sound*, and *Stability of Bridges*. These units focus on the instruction of domain-specific conceptual content knowledge of basic physics concepts. The *Floating & Sinking* unit introduces the concepts of water displacement, object density, and buoyancy force. The *Air & Atmospheric Pressure* unit introduces air as nonvisible matter that has weight and needs space, and how air pressure can be used. The *Sound & Spreading of Sound* unit introduces the concepts of pitch, frequency, and sound wave movement. The *Stability of Bridges* unit introduces basic types of forces and principles of stable construction design. Hardy and colleagues (2006) have provided an extensive exemplary description of one unit (*Floating & Sinking*). Each unit includes all experimentation materials and necessary information (e.g., worksheets, theoretical background information about the content) for implementation by the teacher.

The four curriculum units use the same core educational principles. Students frequently engage in experimentation to explore the different physics concepts (see Appendix A for examples of experimentation activities). The lessons emphasize instructional guidance and scaffolding to support learning. For example, prior knowledge is activated in a teacher-led classroom discussion and in paper-based exercises before experimentation. Students write down their assumptions concerning the outcomes of the experiments, and in a research notebook, they provide justifications for their expectations. After having conducted or observed an experiment, the

students write down the outcomes and compare them to their expectations. The research notebook also contains content-related information and exercises. The teacher concludes the lesson by securing students' understanding of the physics concept in a teacher-led classroom discussion. Importantly, the CVS or any other domain-general scientific reasoning skills are not explicitly mentioned in any of the instructional materials.

The four physics curriculum units encompass 60 lessons in total. The teachers of the intervention classes received a half-day of training for each unit provided by the study authors. To ensure high implementation fidelity, the teachers learned and practiced all instructional sequences and experiments in the trainings. Following the trainings, the teachers implemented the units within the elementary school subject "Human Beings and their Environment".

In the control classes, the teachers continued with their traditional instruction in the subject "Human Beings and their Environment". In Swiss elementary schools, this subject encompasses 3-4 lessons per week, and it includes, among others, topics from Natural Sciences, History, and Geography. However, as previously mentioned, teachers focus on the (social) geographical and the local environment and rarely on physics or other natural sciences (Metzger & Schär, 2008). If they chose topics on natural sciences, the focus is on demonstrating phenomena rather than working out explanations by performing controlled experiments. After the end of the present study, the teachers of the control classes received the same training as the intervention class teachers so that they could later implement the curriculum units.

2.2.2 Assessments

Intervention and control classes answered the same tests, which assessed the students' content knowledge on the four early physics curriculum units and their ability to apply the control-of-variables strategy (CVS).

The content knowledge assessments primarily served as an implementation check. We used paper-and-pencil tests to assess the students' content knowledge. The four tests, one for each curriculum unit, measured the students' domain-specific conceptual understanding of the physics concepts with multiple-choice questions (see Fig. 1 for examples). We constructed two different item orders for each test to prevent students from copying from their neighbors. We summed the number of correct answers to yield an indicator of students' content-specific conceptual knowledge for the topics Air & Atmospheric Pressure, Sound & Spreading of Sound, and Stability of Bridges. For the Floating & Sinking topic, we used a score that indicates how often the students choose a correct concept to explain whether an object floats or sinks without choosing a complementary misconception, as suggested by prior research on this unit (see Hardy et al., 2006; Kleickmann, Tröbst, Jonen, Vehmeyer, & Möller, 2016). To estimate reliability, we used McDonald's omega (ω), a more robust measure than Cronbach's alpha (Dunn, Baguley, & Brunsden, 2014; McNeish, advance online publication). The *Floating & Sinking* test included 11 questions (posttest reliability estimate: $\omega_{posttest} = .70$); the *Air & Atmospheric Pressure* test included 15 questions ($\omega_{posttest} = .67$); the *Sound & Spreading of Sound* test included 17 questions ($\omega_{posttest} = .75$), and the *Stability of Bridges* test included 18 questions ($\omega_{posttest} = .62$). Each curriculum unit and its respective test covered various physics concepts; therefore, we believe that these estimates indicate adequate reliability.

--- insert Fig. 1 here ---

We also assessed the students' ability to apply the CVS with a paper-and-pencil test. Importantly, the tasks of the CVS test are not related to the content knowledge of the curriculum units. Instead, tasks are situated in novel physics or in biology contexts (and, therefore, can be viewed as transfer tasks). The test tasks were analogous to well-established CVS tasks from the research literature, for example, the mouse task (Sodian et al., 1991), ramp task (Chen & Klahr, 1999), and airplane task (Bullock et al., 2009).

The CVS test contained 16 multiple-choice and open answer tasks. Following Bryant and colleagues (2015), we designed two kinds of multiple choice tasks. In the 7 creation tasks, students had to choose which experiment to conduct to examine the potential causal influence of a focal variable (see Fig. 2A for an example of a creation task). In the 4 evaluation tasks, students had to value a given experimental design (i.e., whether it is a good experiment). The students received 1 point for each correct answer and 0 points for incorrect answers. In the 5 open answer tasks, students had to decide whether a given experimental design allows drawing a definite conclusion, and they had to write down a justification for their decision. These tasks can also be classified as evaluation tasks; therefore, there were 9 evaluation tasks in total (see Fig. 2B for an example of an open answer evaluation task). Students received 0 points if their justification did not indicate any understanding of the CVS, 1 point if it referred to a single detail of the experimental design that was (or was not) properly controlled, and 2 points if the justification referred to two or more critical design features or if they explained the rationale underlying the CVS. Two independent raters coded all answers of the open answer tasks in the pre- and posttest, with a median interrater reliability of Spearman's *rho* = .92 (range across items: .72 - .95). Disagreements were resolved through discussion. Overall, the 16 tasks yielded a maximum score of 21 points (with satisfactory

reliabilities at pretest $\omega = .72$ and posttest $\omega = .83$). We used the percentage of points as students' CVS score for the analyses.

---insert Fig. 2 here---

2.3 Procedure

All students first answered the CVS assessment in a pretest. Then, the teachers in the intervention classes implemented the four physics curriculum units and assessed the respective content knowledge immediately before and after each unit. In the control classes, we guided the teachers to administer the content knowledge tests in comparable time intervals. After finishing the fourth unit, the students answered the CVS assessment again as a posttest, with a mean interval of 16 months between CVS pre- and posttest. At the beginning of the study all students were in the 3rd grade, at the end, all were in the 4th grade.

3. Results

3.1 Content knowledge gains

As an implementation check, we first analyzed whether students gained content knowledge from the curriculum units. The intervention and control classes showed comparable pretest performance in the Floating & Sinking, Air & Atmospheric Pressure, and Stability of Bridges assessments (see Table 1). In the Sound & Spreading of Sound assessment, students from the intervention classes performed moderately better than students from the control classes (difference: $d = 0.51$, $p = .001$); it is unclear where this difference originates. On the other content knowledge assessment, there were no statistically significant differences between the intervention and control classes (all ds < 0.23, all ps > .050).

---insert Table 1 here---

To examine students' gains in content knowledge on the four curriculum units, and in their understanding of the CVS, we conducted multilevel analyses. Within school classes, individual development is indistinguishable from the intervention. In a generic regression model (e.g., a repeated measures ANOVA), individual development is confounded with the development of the whole school class. Despite the limited number of school classes, we modeled a multilevel structure of the data. This statistical approach prevents the underestimation of standard errors due to the data dependencies within school classes (McCoach, 2010).

In the multilevel models, we regressed the posttest score on the pretest score on the individual level, and on both the average pretest score and the intervention variable on the classroom level. Between 2 and 25 of the 189 students missed one of the content knowledge assessments per curriculum unit (i.e., either the pre- or posttest). We chose full information maximum likelihood estimation, accounting for the missing values and correcting for deviations from nonnormality with robust estimation (using the Mplus software version 7.1, Muthén & Muthén, 2010), and then we applied the Holm-Bonferroni procedure to account for the multiple dependent variables. The multilevel models showed that students in the intervention classes had higher learning gains than students in the control classes in content knowledge (see Table 1). Students in the intervention classes showed strong learning gains in all four curriculum units, while students in the control classes showed a moderate improvement on one test only (i.e., on the Stability of Bridges test). We speculate that this improvement, which was much smaller than the gains in the intervention classes, reflects a retest effect or some unknown learning opportunities. Overall, these results confirmed the successful implementation of the physics curriculum units as

students in the intervention classes showed much stronger gains in content knowledge for all four units than students in the control classes.

3.2 Control-of-variables strategy assessment

There were also some missing values for the CVS assessments. Eleven students (5.8%) were absent from school at the pretest, and 14 students (7.4%) were absent at the posttest. Class sizes differed (as described in the Participants section), however, the class size did not significantly correlate with the students' CVS scores, neither at the pretest ($r = .05$, $p = .527$) nor at the posttest ($r = .10$, $p = .190$). Fig. 3 shows the performance in the CVS test. Students in the control classes scored $M = 29\%$ ($SD = 20\%$) at the pretest and $M = 38\%$ ($SD = 22\%$) at the posttest ($d = 0.50$); students in the intervention classes scored $M = 30\%$ ($SD = 18\%$) at the pretest and $M = 47\%$ ($SD = 28\%$) at the posttest ($d = 0.75$). Students from the intervention and control classes did not differ on the CVS pretest score ($d = 0.05$, $p = .733$).

---insert Fig. 3 here---

The multilevel model for the CVS assessment (see Fig. 4) indicates that the pretest explained 23% of the posttest variance on the individual within level (see Appendix C). Whether the ability to apply the CVS additionally benefitted from the intervention can be examined on the between level. The intraclass correlation coefficient, that is, the variance in the posttest CVS score explained by classroom differences, was 11%. Of this variance, differences that already existed between school classes at the pretest explained 56%, and the intervention explained 40%. A significant positive regression weight for the intervention variable ($b = 0.09$, $p < .05$, $R^2 = .40$) confirms the improved performance of students in the intervention classes on the CVS posttest compared to students in the control classes. In accordance with McNeish, Stapleton, and Silverman

(2017), who caution against using multilevel modeling with few level 2-units (i.e., school classes), we also estimated a model with cluster-robust standard errors instead of a full two-level-model (Appendix D). However, the model estimates, standard errors, and *p*-values remained highly similar, indicating that the two-level-model is appropriate. Supporting our central hypothesis, students in the intervention classes showed stronger gains on the CVS score than students in the control classes.

---insert Fig. 4 here---

4. Discussion

Students who received four inquiry-based physics curriculum units demonstrated strong gains in physics content knowledge, and they also improved their ability to apply the control-of-variables strategy (CVS) when designing and evaluating experiments in contexts not treated in the curricula. Their strong advantage over the control classes in content knowledge development is not surprising given that the control classes did not receive instruction on these physics topics. However, the higher gains on the CVS assessment in the intervention classes are noteworthy. The CVS was never explicitly instructed in the intervention classes, and these students never encountered experiments with confounded variables (i.e., invalid experimental designs). Nevertheless, they had a small but significant advantage in creating and evaluating experimental designs in novel contexts, as required by the tasks of the CVS assessment.

4.1 Abstracting the control-of-variables strategy

Based on the results of small lab-based studies (Schauble, 1990, 1996) and on research on learning by structural alignment and analogical reasoning (Alfieri et al., 2013; Gentner, 2010; Richland & Simms, 2015), we hypothesized that the strong dose of experimentation in the intervention classes

could result in the collateral benefit of supporting students' emerging understanding of the CVS. The various guided experimentation activities in the four curriculum units exemplified the appropriate application of the CVS in manifold contexts over a long time period. Our results suggest that these multiple examples helped students to abstract a domain-general understanding of the CVS.

To the best of our knowledge, abstraction of a domain-general experimentation strategy has never been empirically corroborated in longitudinal research in real classrooms. Researchers have intensively investigated structural alignment processes in highly controlled laboratory studies or in short classroom intervention studies (Alfieri et al., 2013). In these studies, it proved necessary to support the abstraction of a general principle. For example, without prompting students to describe commonalities and differences between situations, they fail to identify the structural similarity between the situations. However, theories about structural alignment and analogical reasoning (the overarching framework) consistently emphasize that the ability to identify structural and relational similarities across instances (e.g., examples, situations, contexts) is at the core of human cognition (Gentner, 2010; Holyoak, 2005), and the use of structural analogies is indeed more frequent in natural settings (Dunbar, 2001). In our study, the dose of valid experiments (i.e., conclusive experiments implementing the CVS) in the intervention classes was high. Thus, the contexts and situations in which the CVS occurred were variable and diverse, which provided students with an opportunity to discover and abstract this domain-general strategy on their own, without explicit support.

4.2 Strengths, limitations, and directions for future research

We realized this study as a classroom-based quasi-experiment. After receiving an introduction to the learning materials, teachers implemented the curriculum units in their classes, substituting them for the topics that they usually teach. This design has high ecological validity. However, it is also a limitation because it reduces control over how the teachers implemented the curriculum units. However, the curriculum units include all necessary teaching materials, and, thus, provide strong guidance not only for students but also for teachers. The various experiments allowed students to inquire phenomena on their own, scaffolded by prompts to describe expectations, observations, and outcomes of the single experiments in their research notebook. Obviously, the content-focused units worked: Students showed strong gains in conceptual content knowledge for every unit. We did also not monitor the instruction in the control classrooms, which would hardly have been possible given the length of the intervention. Typical education in the subject "Human Beings and their Environment" in Switzerland, however, rarely involves guided experimentation activities. Instead, teachers focus their instruction on delivering facts about geography and the local environment (Metzger & Schär, 2008). Using a waiting list group design, we could avoid selection bias because all teachers were eager to teach physics using guided inquiry; however, they did not feel prepared to do this without receiving training. Thus, we are confident in attributing the students' success in applying the CVS in novel contexts to the strong dose of guided experimentation in the intervention classes.

Importantly, we do not claim that the CVS should not be explicitly trained or that such explicit training should not be combined with an intervention, such as the one that we implemented. Researchers have shown that explicit training can be effective (e.g., Lorch Jr. et al., 2014; Schwichow, Zimmerman, et al., 2016; Strand-Cary & Klahr, 2008). Therefore, we urge

researchers to investigate combinations of intensive content knowledge instruction and explicit CVS trainings in the future.

Another limitation of our design is that we do not exactly know how the physics units led to the improved ability to apply the CVS. Whether it was an overall effect of the large dose of experimentation across the four curriculum units, or whether the units differed in their impact on fostering understanding of the CVS remain open questions. Moreover, experimentation skills are not limited to CVS but encompass various other facets, which we did not assess in our study (Koerber et al., 2015; Kuhn, Iordanou, Pease, & Wirkala, 2008; D. Mayer et al., 2014; National Research Council, 2012; Zimmerman, 2007). For example, recent research has indicated that experimentation skills are predicted by a broad epistemological understanding of science (Osterhaus et al., 2017). Future research should aim to continue disentangling the interplay of content-knowledge development and the development of various facets of experimentation skills.

5. Conclusion

The physics curriculum units had a double benefit. Elementary school students improved their content knowledge, as well as their ability to apply the CVS in novel contexts compared to students who received their traditional instruction. The effect size for this transfer effect was not particularly strong, however, given that CVS understanding in turn predicts content learning in science (Bryant et al., 2015), even a small advantage might profoundly impact learning with continuing education. Our results complement previous lab-based research on the interrelation of domain-specific and domain-general knowledge development (Schauble, 1990, 1996) and scale them up to real classrooms: guided inquiry-based instruction focused on content knowledge development can have

the collateral benefit of supporting the ability to transfer the CVS, a domain-general experimentation skill.

Acknowledgments

References

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., . . . Tuan, H.-L. (2004). Inquiry in science education: International perspectives. *Science Education, 88*, 397–419. doi:https://doi.org/10.1002/sce.10118

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18. doi:http://dx.doi.org/10.1037/a0021017

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist, 48*(2), 87-113. doi:http://dx.doi.org/10.1080/00461520.2013.775712

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*, S145-S163. doi: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: Finding from a 20 year longitudinal study* (pp. 173-197). New York, NY: 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). New York, NY: Cambridge University Press.

Chan, J., & Schunn, C. D. (2015). The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science, 39*, 126-155. doi:https://doi.org/10.1111/cogs.12127

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. doi: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). Amsterdam: Elsevier.

Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47*(3), 177-188. doi:https://doi.org/10.1080/00461520.2012.695709

Dunbar, K. (2001). The analogial paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind* (pp. 313-334). Cambridge, MA: MIT Press.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*, 399-421. doi:10.1111/bjop.12046

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

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*, 752-775. doi:10.1111/j.1551-6709.2010.01114.x

Gentner, D., & Hoyos, C. (2017). Analogy and abstraction. *Topics in Cognitive Science, 9*, 672-693. doi:https://doi.org/10.1111/tops.12278

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin, 142*(7), 729-757. doi:http://dx.doi.org/10.1037/bul0000043

Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., & Hatfull, G. F. (2006). Teaching scientific inquiry. *Science, 314*(5807), 1880-1881. doi:10.1126/science.1136796

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. doi:http://dx.doi.org/10.1037/0022-0663.98.2.307

Hmelo-Silver, C. E., Golan Duncan, R., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99-107. doi:https://doi.org/10.1080/00461520701263368

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (in press). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology*. doi:http://dx.doi.org/10.1037/edu0000266

Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 117-142). Cambridge, UK: Cambridge University Press.

Kleickmann, T., Tröbst, S., Jonen, A., Vehmeyer, J., & Möller, K. (2016). The effects of expert scaffolding in elementary science professional development on teachers' beliefs and motivations, instructional practices, and student achievement. *Journal of Educational Psychology, 108*(1), 21-42. doi:http://dx.doi.org/10.1037/edu0000041

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. doi:10.1111/cdev.12298

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development, 23*, 435-451. doi:https://doi.org/10.1016/j.cogdev.2008.09.006

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(6), 681-718. doi:https://doi.org/10.3102/0034654315627366

Leuchter, M., Saalbach, H., & Hardy, I. (2014). Designing science learning in the first years of schooling. An intervention study with sequenced learning material on the topic of 'floating and sinking'. *International Journal of Science Education, 36*(10), 1751-1771. doi:http://dx.doi.org/10.1080/09500693.2013.878482

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science, 344*(6180), 155-156. doi:10.1126/science.1245980

Lorch Jr., 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. doi:http://dx.doi.org/10.1037/a0017972

Lorch Jr., 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. doi:http://dx.doi.org/10.1037/a0034375

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. doi:https://doi.org/10.1016/j.learninstruc.2013.07.005

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods in instruction. *American Psychologist, 59*(1), 14-19. doi:http://dx.doi.org/10.1037/0003-066X.59.1.14

McCoach, D. B. (2010). Hierarchical linear modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 123-140). New York: Routledge.

McNeish, D. (advance online publication). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*. doi:http://dx.doi.org/10.1037/met0000144

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierachical linear modeling. *Psychological Methods, 22*(1), 114-140. doi:https://doi.org/10.1037/met0000078

Metzger, S., & Schär, P. (2008). Rahmenbedingungen des NaTech-Unterrichts im quantitativen und qualitativen Vergleich: Vom Kindergarten bis zur Sekundarstufe 1 [Determining qualitative and quantitative factors of MINT instruction: From kindergarden to secondary

education]. In E. Stern, S. Metzger, & A. Zeyer (Eds.), *Expertise zu Naturwissenschaften und Technik in der Allgemeinbildung im Kanton Zürich [Expertise on science and technology in general education in the Canton Zurich]* (pp. 36-51). Zurich: Bildungsdirektion.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide: Statistical analysis with latent variables*. Los Angeles: Muthén & Muthén.

National Research Council (Ed.) (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

Nokes-Malach, T. J., & Mestre, J. P. (2013). Toward a model of transfer as sense-making. *Educational Psychologist, 48*(3), 184-207. doi:https://doi.org/10.1080/00461520.2013.807556

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. doi:http://dx.doi.org/10.1037/dev0000260

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 177-192. doi:http://dx.doi.org/10.1002/wcs.1336

Sandoval, W. A., Sodian, B., Koerber, S., & Wong, J. (2014). Developing children's early competencies to engage with science. *Educational Psychologist, 49*(2), 139-152. doi:http://dx.doi.org/10.1080/00461520.2014.917589

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for

    generating evidence. *Journal of Experimental Child Psychology, 49*, 31-57.

    doi:https://doi.org/10.1016/0022-0965(90)90048-D

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts.

    Developmental Psychology, 32(1), 102-119. doi:http://dx.doi.org/10.1037/0012-

    1649.32.1.102

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.

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

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. doi:DOI

    10.1002/tea.21320

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical

    beliefs from evidence. *Child Development, 62*, 753-766. doi:10.1111/j.1467-

    8624.1991.tb01567.x

Spectra Materials - KiNT-Boxes 1-4. Retrieved from http://www.spectra-

    verlag.de/englisch/SID=4h6lug0rnpijls081smu7t4n92/shop/start.php3

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional

    effectiveness and path independence. *Cognitive Development, 23*, 488-511.

    doi:https://doi.org/10.1016/j.cogdev.2008.09.005

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development,*

    51(1), 1-10. doi:http://www.jstor.org/stable/1129583

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. doi:https://doi.org/10.1016/j.learninstruc.2018.03.005

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

Table 1. Domain-specific content knowledge scores and gains across the four curriculum units.

Unit	Condition	Pretest M (SD)	Posttest M (SD)	Gain d	Effect of intervention (classroom level, controlling for pretest scores)
Floating & Sinking	Control	3.79 (1.46)	3.86 (1.38)	0.06	$t(175) = 7.48, p < .001, R^2 = .12$
(max. 16)	Intervention	3.43 (1.81)	8.46 (3.25)	1.80***	
Air & Atmospheric	Control	6.69 (2.07)	6.64 (1.88)	-0.04	$t(186) = 12.41, p < .001, R^2 = .76$
Pressure (max. 15)	Intervention	6.92 (2.37)	11.08 (2.03)	1.78***	
Sound & Spreading	Control	7.66 (2.49)	8.14 (2.68)	0.19	$t(186) = 5.97, p < .001, R^2 = .40$
of Sound (max. 17)	Intervention	9.00 (2.79)	12.50 (2.62)	1.28***	
Stability of Bridges	Control	9.06 (1.80)	10.06 (2.07)	0.41***	$t(177) = 5.78, p < .001, R^2 = .66$
(max. 18)	Intervention	9.31 (1.96)	12.47 (1.73)	1.42***	

The gain per condition is presented as Cohen's *d* for the repeated measures. R^2 is the estimated explained variance in the posttest that is attributable to the intervention controlling for pretest scores. *** indicates $p < .001$

A

This metal plate is immersed into water. What do you expect to happen?

The metal plate…

☐ sinks   ☐ floats

Why?

☐ …because it lies on the water.

☐ …because it is metal.

☐ …because its weight is evenly spread out.

☐ …because the displaced water weighs less than the plate.

☐ …because it is not pushed up strongly enough by the water.

☐ …because the plate is so heavy.

B

You stuff a tissue into a glass. Then, you turn the glass upside-down and immerse it into a bucket that is filled with water. What do you expect to happen?

☐ The tissue gets wet but dries quickly when the glass is lifted.

☐ The tissue falls out of the glass and floats away.

☐ The tissue gets wet because water enters the glass.

☐ The tissue does not get wet because air is in the glass between the water and the tissue.

C

What do you expect to happen when air is pumped out of a bell jar covering a ringing alarm clock?

☐ Nothing happens.

☐ The alarm becomes louder and louder.

☐ It gets darker in the glass until one cannot see the clock anymore.

☐ The alarm becomes quieter and quieter.

D

These two clotheslines are almost the same. But clothesline 1 is attached to the top of the posts, while clothesline 2 is attached to the ground. You hang up two similar baskets and loaded them with bricks. Can one of the two clotheslines hold more bricks than the other?

Clothesline 1

Clothesline 2

☐ Clothesline 1 can hold more.

☐ Clothesline 2 can hold more.

☐ Both can hold the same amount.

Fig. 1. Example items from the content knowledge assessments. The figure shows one example for each of the assessments for the different curriculum units: (A) *Floating & Sinking*, (B) *Air & Atmospheric Pressure*, (C) *Sound & Spreading of Sound*, and (D) *Stability of Bridges*. Items are translated from the original German versions of the assessment.

Fig. 2. Example items from the control-of-variables (CVS) assessment. (A) represents an example of a creation task; (B) represents an example of an open-ended evaluation task. These items were adapted from Bullock and colleagues (2009) and Chen and Klahr (1999). The items are translated from the German versions used in the present study.
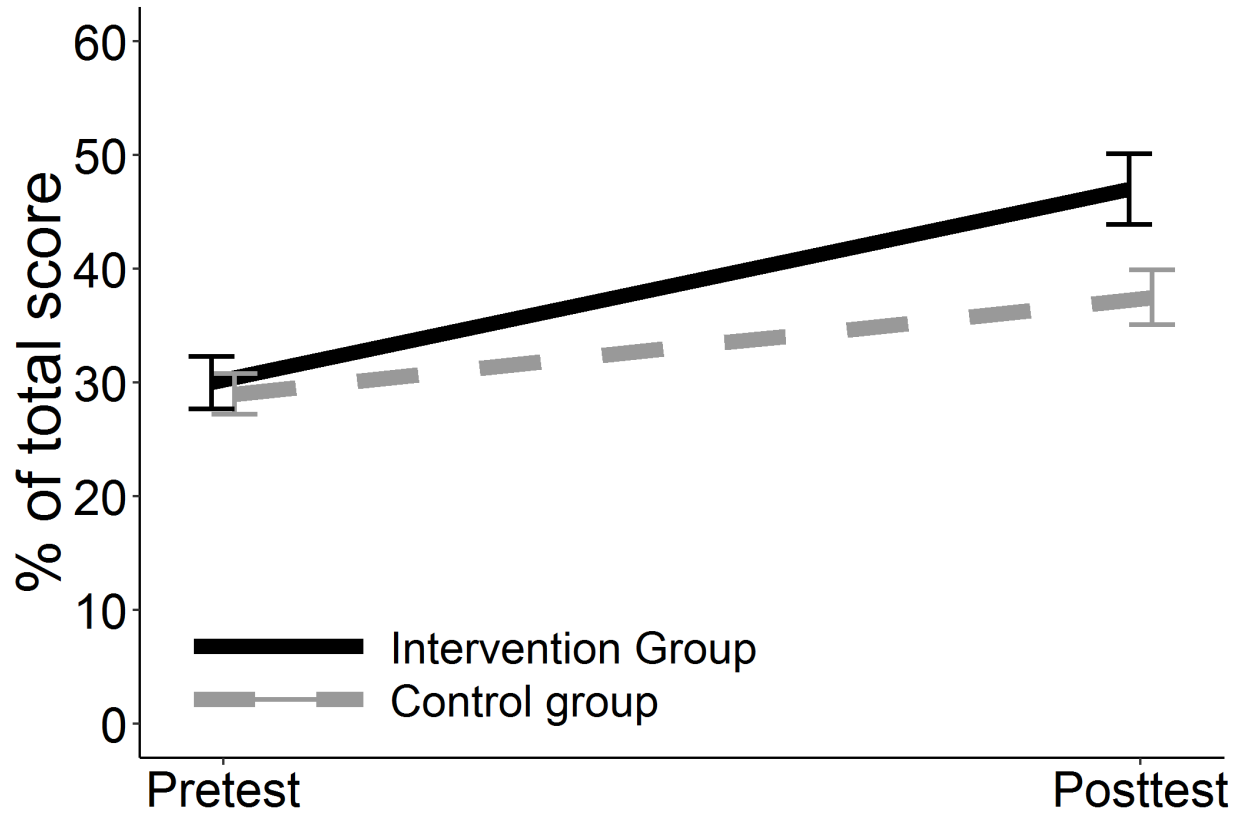
Fig. 3. Students' mean scores on the control-of-variables strategy (CVS) assessment before and after instruction. The error bars represent the standard error of the mean.
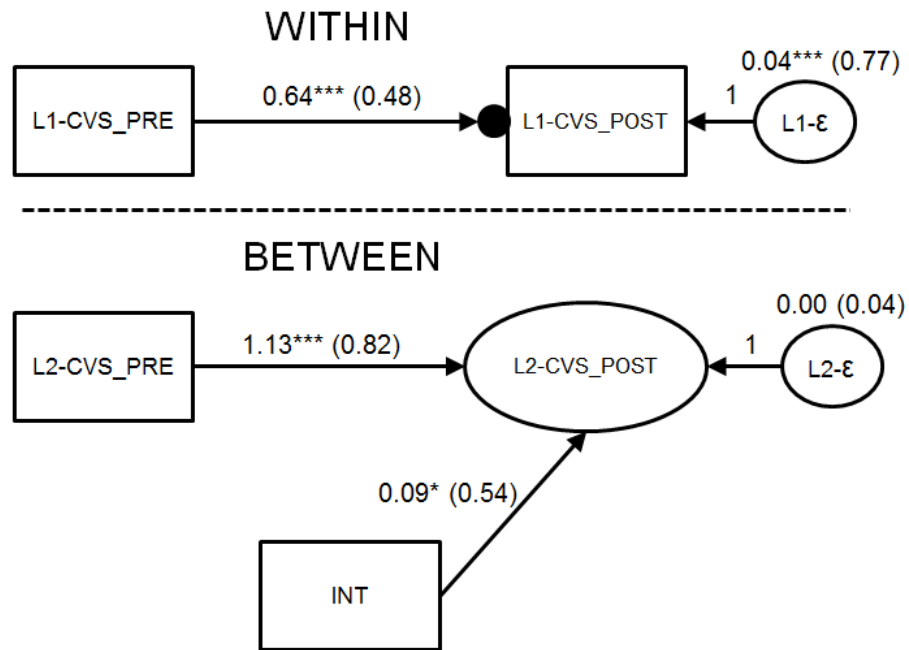
Fig. 4. A multilevel regression model predicting the control-of-variables strategy (CVS) posttest score. On the WITHIN level, the individual posttest score (L1-CVS_POST) is regressed only on the individual pretest score (L1-CVS_PRE); on the BETWEEN level, the class average posttest score (L2-CVS_Post) is regressed on the class average pretest score (L2-CVS_PRE) and on the intervention (INT). The dot in the WITHIN-level model indicates random intercepts across the school classes. Raw parameter estimates and significance levels are presented outside the brackets, and standardized parameter estimates are presented within the brackets. Standardized parameters of residuals indicate the percentage of nonexplained variance on the level of students (L1-ε) and school classes (L2- ε). * indicates $p < .05$, and *** indicates $p < .001$.

Appendix A: Additional information on the curriculum units

In the intervention classes, we implemented four basic physics curriculum units: (1) Floating & Sinking, (2) Air & Atmospheric Pressure, (3) Sound & Spreading of Sound, and (4) Stability of Bridges. Each unit included various guided experimentation activities with the aim to improve the students' content knowledge. In the following, we describe the primary content knowledge learning goals and describe several exemplary experiments for each unit. In the tables that compile the exemplary experiments (see Table A.1-A.4), we provide information about the children's typical naïve conception ("What children initially think"), the specific learning goal ("What children are supposed to learn"), the setup of the experiment ("Guided experiment"), and how the control-of-variables strategy was implemented in the experiment ("Variables controlled and varied").

A.1 Floating & Sinking

The primary learning goal of this curriculum unit is that children can explain and predict why objects float or sink. Scientifically appropriate explanations are based on the concepts of water displacement, object density, and buoyancy force. It is expected that children can acquire a prequantitative understanding of these concepts when they overcome naïve explanations, such as "light things float while heavy ones sink", or "a ship made of iron floats because the air inside pulls it up". A prequantitative understanding means knowing that objects float if the amount of displaced water has more weight than the object itself and that objects sink if the displaced water is lighter than the objects themselves. A basic understanding of the relation between the amount of displaced water and the amount of the buoyancy force, for example, is supported in the following experiment. Children are instructed to immerse pots of different sizes into water and to report which pot requires more effort to be immersed into water. Teachers are asked to refrain

from using the labels "buoyancy force" and "density" because children at this age cannot be

expected to fully understand these concepts. The goal was to direct children's conceptual

understanding in the right direction. The three experiments described in Table A.1 are examples

of how children investigated the crucial variable for the outcomes of immersing objects into

fluids.

Table A.1 Exemplary experiments from the Floating & Sinking unit

What children initially think	What children are supposed to learn	Guided experiment	Variables controlled and varied
It depends on object characteristics, such as weight, size, or shape whether something floats or sinks in water. For example, light things float, and heavy ones sink; solid objects that float will sink if holes are inserted.	For solid bodies that do not enclose air, it only depends on the kind of material whether an object floats or sinks. Other factors, such as weight and size, have no influence on the floating and sinking of solid bodies.	Children handle and explore several solid objects of different material, weight, size, and shape. In a first step, they must predict for each object whether it will float or sink. In a second step, children have to immerse the object into water and observe the results.	Characteristics, such as material, weight, size, and shape, are varied within pairs of objects. For example, a wooden and a metallic plank of the same size and shape must be compared (i.e., the objects differ only in the material that they are made from).
The amount of water displaced by a solid object fully immersed into water depends on the weight, the material, or the shape of an object. A cube made of iron is expected to displace more water than a Styrofoam cube of the same size. When one identical piece of plasticine is modeled as a ball, it will displace more water than if it is modeled as a slice.	The amount of displaced water only depends on a solid object's volume. However, the term "volume" is not yet explicitly used. To prepare a basic understanding of the relationship between an object's volume and the amount of displaced water, children are instructed, for example, to observe the amount of water displaced by cubes of the same size but different material.	In a series of four experiments, children systematically explore the influence of different factors: material, shape, weight, and volume on the amount of water displaced by an object. In particular, they observe the increase in the water level when an object is immersed into a measuring jug. In the case of shape, for example, children are asked to change a ball of plasticine into a slice without losing material.	The material, shape, weight, and volume of objects are systematically varied, with all other factors remaining constant.
Most children are not aware that floating and sinking not only depend on the object but also on the fluid.	An object that floats in water may sink in a fluid which is lighter than water (e.g., oil). An object that sinks in water may float in a heavier fluid (e.g., salt water).	Children are presented with a boiled egg that sinks in water. They are instructed to think of possible activities to make the egg float (e.g., "add salt").	The object is held constant, and the density of the liquid is varied.

A.2 Air & Atmospheric Pressure

The primary learning goal of this curriculum unit is to support children in developing an appropriate concept of air. That is, children learn that air – although invisible – has a material nature and, thus, interacts with the physical environment. Many children think, for example, that air has no material nature; therefore, they predict that an inflated ball has the same weight as a flat ball. Some even predict that the inflated ball has less weight because they presume that air has a negative weight. To prove otherwise, children, for example, perform an experiment with a flat and an inflated ball on a scale (see Table A.2 for further exemplary experiments). In addition, children learn that air can propel sailing ships and slow down parachutes, that heated air expands and, therefore, rises and that air can exert pressure on objects in the earth's atmosphere due to its weight.

Table A.2 Exemplary experiments from the Air & Atmospheric Pressure unit

What children initially think	What children are supposed to learn	Guided experiment	Variables controlled and varied
As air in everyday situations can neither be seen nor felt, children assume that air is nothing and, thus, needs no space.	Air has a material nature; that is, air is something that needs space.	In a series of four experiments, children determine that air needs space. For instance, they try to inflate a balloon that is inserted in a bottle which either has a hole or not. Only if the bottle has a hole it is possible to inflate the balloon.	In all four experiments of this series, the volume and material of the various containers (e.g., a bottle) are kept constant. The only variable is whether the air can leave the container (e.g., through a hole).
Parachutes with a small surface are as good as large ones.	It depends on the size of a parachute's surface how strongly it is slowed down by the air.	Children build their own parachutes and perform experiments with them to determine which factors influence how strongly a parachute slows down.	The material and shape of the parachute, as well as the size of its surface, are systematically varied. In addition, the weight of the parachutist and the height of the drop point are also varied.

Vacuum cups adhere on flat surfaces, such as glazed tiles, because the vacuum somehow sucks them onto the tiles.	The difference of the air pressure under the vacuum cup and the air pressure outside the vacuum cup is responsible for vacuum cups adhering on flat surfaces.	Children use intact vacuum cups and cups with holes to explore the role that the difference in air pressure plays in the cup's adherence on flat surfaces.	The material and the size of the vacuum cups, as wells as the flat surface, are kept constant. The only variable is whether the cups are intact or have a hole.

A.3 Sound & Spreading of Sound

The primary learning goal of this unit is to support children in developing an appropriate concept of sound. In particular, children are supposed to understand that sounds are produced by vibrations and that sounds spread by waves that need a medium (e.g., a gas, a liquid, or a solid body). For example, students observe that they can hear an alarm clock ringing under a glass only if there is air in the glass in a demonstration experiment. If the air is removed with the help of a vacuum pump, the alarm clock can no longer be heard. Furthermore, in a series of systematic experiments, children investigate which variables affect the pitch and loudness of sounds (see Table A.3 for further exemplary experiments). In this unit, children also learn about the structure of the human ear and the different functions of its parts.

Table A.3 Exemplary experiments from the Sound & Spreading of Sound unit

What children initially think	What children are supposed to learn	Guided experiment	Variables controlled and varied
Before the instruction, children are typically unaware of the physical mechanisms underlying the pitch and loudness of sounds.	Differences in the pitch of sounds can be explained in terms of the frequency of vibrations, whereas differences in the loudness of sounds have to be explained in terms of the intensity of vibrations.	Children are instructed to perform experiments with a ruler on a table and a small guitar to investigate which factor influence the pitch and loudness of sounds.	The material and size of the rulers and guitars are kept constant. Only the length of the vibrating part of the ruler or the guitar string, and the intensity of the vibrations, are varied.
Without instruction, children usually do not know that sound consists	Sound consists of waves transmitted by the air.	Children perform different experiments to examine how sound waves are transmitted by	The materials are kept constant. Children can vary the distance between the source of

of waves transmitted by the air.		the air. For example, they hold a balloon in their hands and feel it vibrate when they shout. They strike a drum, and the sound waves, bundled and intensified by a funnel-shaped tube, move a little ball.	sound and the receiver of the sound waves.

A.4 Stability of Bridges

The primary learning goal of this unit is to prepare a basic understanding of some simple principles of mechanics. All mechanics concepts are introduced through concrete examples of different kinds of bridges. For instance, the concept of counterbalance is illustrated by a simple bridge consisting of wooden blocks that needs additional blocks as counterbalances to be stabilized (see Table A.4 for exemplary experiments). Furthermore, the mechanical principle that forces can be split into vertical and horizontal forces, is illustrated by Roman arch bridges that need lateral counter bearings for stability. Children also learn about the stability of triangles and of different profiles.

Table A.4 Exemplary experiments from the Stability of Bridges unit

What children initially think	What children are supposed to learn	Guided experiment	Variables controlled and varied
Prior to instruction, the children are typically unaware of the lateral forces at arch bridges.	To stabilize arch bridges, lateral counter bearings are needed to compensate for the horizontal forces.	Children investigate the lateral forces through experiments with a wooden model of a roman arch bridge.	The material and size of the bridge model are kept constant. Only the weights on top of the bridge and the lateral counter bearings are varied.
It only depends on the number of stiffeners to make a frame bridge stable.	To stabilize frame bridges, a certain orientation of stiffeners (i.e., the so-called stable triangle) is decisive, not the number of stiffeners.	Children perform different experiments to examine how to stabilize a frame bridge.	Children can choose between stiffeners of different size, and they try out stiffeners in different orientations.

Beam bridges and suspension bridges have similar stability.	Suspension bridges are significantly more stable than beam bridges.	Children design an experiment to compare the stability of beam bridges to the stability of suspension bridges.	The bridged distance, the material of the bridges and the weights are kept constant; only the kind of bridge construction is varied.

Appendix B: Power analysis

A simulation study was conducted using the Mplus software package, version 7.11, to estimate

the statistical power the sample offered for finding a small effect (2% of overall variance

explained) of the intervention, which was implemented on the level of school classes. Data were

simulated to stem from twelve school classes similar to those in our sample. The simulated

model was the exact two-level model reported in the article. The result was that in 90.9% (i.e.,

909) of 1000 simulated data sets, the parameter estimate for the intervention effect on the level

of school classes showed a p-value $< .05$, our chosen significance level. This result indicates high

statistical power for our study.

Appendix C: Multilevel model assumptions for the control-of-variables (CVS) assessment

The primary reason to apply the multilevel model was to control for correlated residuals on the level of school classes. Linearity and homoscedasticity of effects were investigated for the continuous predictor variable (i.e., pretest score on the CVS assessment). Visual inspection of the plot indicates linearity and homoscedasticity (see Fig. C.1).
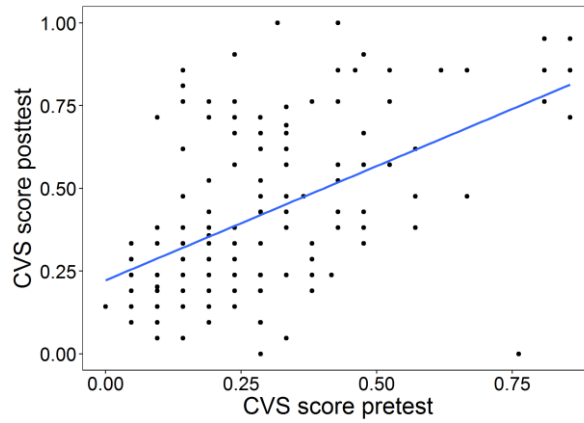

Fig. C.1 A scatterplot of the pre- and posttest scores on the control-of-variables assessment

The assumptions of normality (see Table C.1) and homogeneity of variances across the two conditions were slightly violated (variance homogeneity tested in Mplus: $\chi^2_1 = 5.05$, $p = .025$), therefore, we used the robust estimator, which can handle such deviations up to a moderate degree.

Table C.1 Skewness and kurtosis of the control-of-variables strategy assessment

	Skewness (pretest)	Kurtosis (pretest)	Skewness (posttest)	Kurtosis (posttest)
Control	1.16	0.97	0.82	-0.01
Intervention	1.62	2.8	0.23	-1.32

Appendix D: Path model with cluster-robust standard error estimates

Recent research indicates that multilevel modeling is not always the preferred analytical approach when the number of level 2-units (i.e., school classes) is low (McNeish et al., 2017). Therefore, as a complementary analytical approach to the multilevel model, we estimated a path model with cluster-robust standard errors (*type = complex* option in the Mplus software package version 7.11, indicating school classes as cluster variable). The results from this model are presented in Fig. D.1.
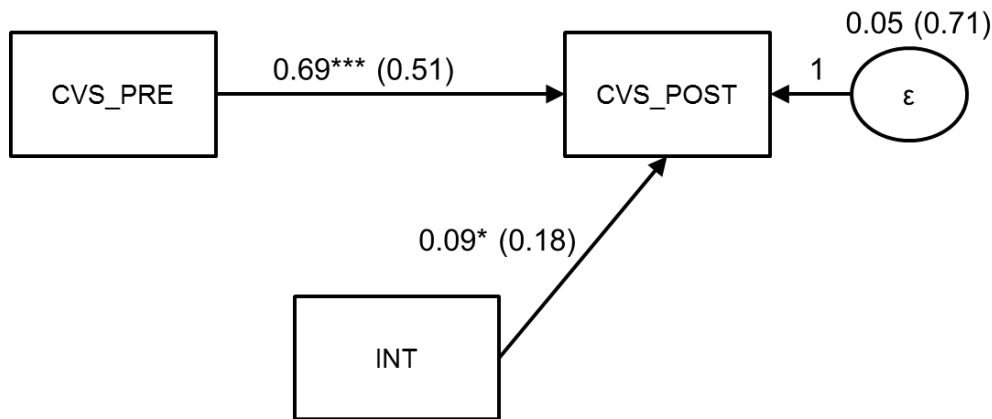

Fig. D.1 A regression model predicting the control-of-variables strategy (CVS) posttest score. The posttest score (CVS_POST) is regressed on the pretest score (CVS_PRE) and on the intervention (INT). Raw parameter estimates and significance levels are presented outside the brackets, and standardized parameter estimates are presented within the brackets. Standardized parameter of residual (ε) indicates the percentage of nonexplained variance. * indicates $p < .05$, and *** indicates $p < .001$.