

# Journal of Educational Psychology

## **Preparation for Future Conceptual Learning: Content-Specific Long-Term Effects of Early Physics Instruction**

Peter A. Edelsbrunner, Ralph Schumacher, Brigitte Hanger-Surer, Lennart Schalk, and Elsbeth Stern

Online First Publication, September 9, 2024. <https://dx.doi.org/10.1037/edu0000887>

### CITATION

Edelsbrunner, P. A., Schumacher, R., Hanger-Surer, B., Schalk, L., & Stern, E. (2024). Preparation for future conceptual learning: Content-specific long-term effects of early physics instruction.. *Journal of Educational Psychology*. Advance online publication. <https://dx.doi.org/10.1037/edu0000887>

# Preparation for Future Conceptual Learning: Content-Specific Long-Term Effects of Early Physics Instruction

Peter A. Edelsbrunner<sup>1, 2</sup>, Ralph Schumacher<sup>1</sup>, Brigitte Hänger-Surer<sup>3</sup>, Lennart Schalk<sup>4</sup>, and Elsbeth Stern<sup>1</sup>

<sup>1</sup>Department of Humanities, Political and Social Sciences, ETH Zurich

<sup>2</sup>Department of Psychology, LMU Munich

<sup>3</sup>Institute for Secondary School I and II, Pädagogische Hochschule Fachhochschule Nordwestschweiz

<sup>4</sup>Institute for Research on Instruction and Subject-Specific Education, Pädagogische Hochschule Schwyz

This study used a quasirandomized within-classroom design to investigate whether prior knowledge about physics gained in elementary school prepares students for future learning in related content areas in secondary school. A total of 433 children (intervention group) received four basic curriculum units on physics from their elementary school teachers. The units dealt with floating and sinking, air and atmospheric pressure, the stability of bridges, and sound and the spreading of sound. These children entered 60 newly composed classes in early secondary school that completed an advanced curriculum unit on hydrostatic pressure and buoyancy force with their secondary school teachers. A total of 942 students (control group) in these classes had not received the four basic physics curriculum units. On a conceptual knowledge test about hydrostatic pressure and buoyancy force, the intervention group outperformed the control group in the pretest ( $d = 0.28$ ) and in the posttest ( $d = 0.25$ ). Students in the intervention group showed similar learning gains as those in the control group, but when controlling for pretest performance, they achieved higher learning outcomes. Regression analyses within the intervention group revealed that this advantage resulted from the content-specific transfer of conceptual knowledge from topically related basic curriculum units. The basic physics instruction also prepared male and female students equally for future learning.

## *Educational Impact and Implications Statement*


Our study shows that students who have received early physics instruction in elementary school benefit slightly more from later, more advanced physics instruction in secondary school. This finding provides first evidence that the idea of a spiral curriculum, in which learners first build basic knowledge that is later on expanded in more demanding instruction, can work, although it is yet to be further examined how this process can be optimized in school instruction.

**Keywords:** physics education, spiral curriculum, knowledge transfer, science learning, quasirandomized within-classroom design

Physics is one of the most challenging and unpopular subjects in high school for many otherwise capable learners (Hofer et al., 2018; Möller et al., 2006). Educational researchers agree that a main barrier to understanding scientific concepts and explanations in physics classes is learners' personal alternative conceptual frameworks about the functioning of the physical world. These naïve concepts (sometimes labeled misconceptions or intuitive conceptions) are derived and often overgeneralized from everyday experiences, and they conflict with scientific explanations (e.g., Carey, 2000). For

example, an appropriate understanding of the concept of density is impeded by children's belief that all light objects float in water while heavy objects sink. Likewise, the understanding of buoyancy force is hampered by children's belief that a ship made of steel floats because the air in the ship's body pulls it upward (Hardy et al., 2006). Many naïve concepts on matter, force, hydrostatic pressure, and other basic topics persist until university and beyond, although students study scientific explanations in physics classes (Loverude et al., 2010; Mazur, 2015; Tobin et al., 2023).

Olusola O. Adesope served as action editor.

Peter A. Edelsbrunner  <https://orcid.org/0000-0001-9102-1090>

The basic curriculum units were implemented within a study that was funded by the Jacobs Foundation from 2010 to 2015 under the title "Boosting Hidden Potential in Science Education." The data and analysis code for this article can be found at <https://osf.io/94rxq/>.

This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License (CC BY-NC-ND 4.0; <https://creativecommons.org/licenses/by-nc-nd/4.0/>). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license

is included in attribution. No derivative works are permitted under this license.

Peter A. Edelsbrunner conducted the formal analysis. Peter A. Edelsbrunner and Lennart Schalk contributed equally to data curation. Peter A. Edelsbrunner, Lennart Schalk, and Elsbeth Stern contributed equally to writing—original draft and writing—review and editing. Ralph Schumacher and Elsbeth Stern contributed equally to conceptualization and funding acquisition. Peter A. Edelsbrunner, Ralph Schumacher, Brigitte Hänger-Surer, and Lennart Schalk contributed equally to methodology.

Correspondence concerning this article should be addressed to Peter A. Edelsbrunner, Department of Humanities, Political and Social Sciences, ETH Zurich, RZ H16, Clausiusstrasse 59, 8092 Zurich, Switzerland. Email: [peter.edelsbrunner@gmail.com](mailto:peter.edelsbrunner@gmail.com)

In this article, we present the results of a longitudinal field trial in which we investigated whether early physics education in elementary school facilitates the understanding of related but more advanced physics concepts years later in secondary school. The implementation of such a study involves methodological challenges, but it also provides the opportunity for theoretical considerations that contribute to a better understanding of learning and knowledge transfer at school. Moreover, because the study was conducted with a large number of regular in-service teachers, it provides information about whether and why implementing early physics education may be worthwhile. To this end, we prepared learners to learn in secondary school about hydrostatic pressure, a topic for which the persistence of naïve conceptions has been shown (Loverude et al., 2010).

Physics educators and many teachers widely agree that the challenge of effective physics teaching is to productively use learners' prior knowledge and help them reconcile it with scientific ideas, a process often labeled conceptual change (diSessa & Minstrell, 1998; Hammer & Elby, 2003; Hofer et al., 2018; Vosniadou, 2019). In recent decades, various theoretical models and approaches for initiating and supporting this process of knowledge reconstruction have been developed (for an overview, see Potvin et al., 2020). Traditional approaches to conceptual change rely on sudden corrections initiated by a cognitive conflict that demonstrates the shortcomings of students' knowledge (Posner et al., 1982). Alternatively, more recent approaches emphasize the importance of providing sufficient time for revising knowledge and integrating unconnected pieces of knowledge into a coherent conceptual network. For example, Vosniadou (2019) emphasizes the importance of intermediate steps of learning that must be accompanied by an understanding that ideas and explanations should be continuously checked for their agreement with facts. This process can be initiated in elementary school by introducing science as a method of inquiry (Shtulman & Walker, 2020). There is also growing evidence that contradictory and incompatible scientific conceptions may actually coexist in a learners' minds, even if the scientifically correct one prevails (Potvin & Cyr, 2017; Shtulman & Harrington, 2016). Early exposure to fundamental science concepts may help children to question intuitive frameworks that may later interfere with learning scientific theories (Shtulman & Walker, 2020).

Hardy et al. (2006) showed that many third graders who had completed teacher-guided inquiry-based teaching units in physics no longer expressed misconceptions about floating and sinking objects in water (e.g., the air in the ship pulls it upward) and retained appropriate conceptual understanding 1 year later (e.g., the water pushes the ship upward). These effects can also be achieved by regular elementary school teachers who are trained to implement teaching units in basic physics (Edelsbrunner et al., 2018; Schalk et al., 2019). These studies indicate that early physics instruction may be valuable to prevent the consolidation of intuitive conceptual frameworks and prepare students for more advanced instruction.

A further argument for an earlier start of physics instruction concerns the gender gap, which is particularly prevalent in physics. While scientific concepts in physics are difficult for everyone to understand, numerous studies worldwide have found that female students struggle more than their male peers (van den Hurk et al., 2019). For example, otherwise competent female students at the high school level are more likely to end up as underachievers in physics than male students (Hofer & Stern, 2016). There are various reasons for the gender gap in science, technology, engineering & mathematics (STEM) areas (Ceci et al., 2009), and its mitigation requires a variety of approaches. One approach is to start physics classes in elementary school because many studies indicate

that the gender gap in academic preferences solidifies in early secondary school (Ceci et al., 2009; Legewie & DiPrete, 2012). If girls in elementary school already experience competence in explaining and predicting phenomena of the physical world, they may be less likely to regard physics as a "male" subject that is not for them (see, e.g., Steegh et al., 2019). Overall, starting physics instruction early may allow a gradual development of conceptual understanding and reduce the gender gap.

## Preparation for Future Learning by Implementing Spiral Curricula

The effort required to succeed in academic learning contrasts with the ease of acquisition of many other skills that are supported by innate learning systems (Stern, 2017). Reading and writing, mathematics, and the natural sciences are recently developed cultural achievements that are possible because of the flexibility of the human mind to reutilize and recycle innate learning systems in completely new ways (Dehaene, 2011). To maintain such cultural achievements (and further develop them; see, e.g., Tomasello, 2009), every generation must undergo time-consuming learning processes in which school curricula that build upon one another often play an important role. Mastering academic skills such as reading, writing, and mathematics can be facilitated through early educational programs starting in kindergarten. Promoting precursor skills of reading and writing (e.g., phonological awareness, Schneider et al., 1997) or mathematics (e.g., focusing on numerosity; Hannula-Sormunen et al., 2020) facilitates learning in elementary school because children have the opportunity to build knowledge and skills that form the basis for more complex future learning. The more precisely early training is tailored to the competencies to be acquired later, the greater the benefits. For instance, in a randomized intervention study, Siegler and Ramani (2008) showed that playing linear, but not circular, board games helped students to acquire the concept of a number line and use it for arithmetic understanding. These examples demonstrate that the learning of more advanced concepts and skills builds on more basic but essential concepts and skills.

The early content-sensitive preparation for more advanced science learning was expressed in Bruner's (1960) idea of a spiral curriculum. Content must be structured such that complex concepts can be understood at a simplified level first and then subsequently revisited at more complex levels. By revisiting, expanding, and refining previously learned concepts, children may achieve a more complete understanding of concepts and how they relate to one another (see, e.g., Chi & VanLehn, 2012; M. Schwartz, 2009) so that they make productive use of their acquired knowledge (De Corte, 2003). Following the idea of a spiral curriculum, implementing early science curricula in elementary school is considered a worthwhile approach to address naïve conceptions so that they do not hamper the acquisition of more advanced topics.

The benefit of implementing physics education in elementary school may go beyond limiting the damage caused by solidified misconceptions. Early guided inquiry-based curricula (for an overview, see Lazonder & Harmsen, 2016) may help children build simplified pre-concepts that are nonetheless compatible and that prepare them for scientific concepts and theories. Learners may use these preconceptions when confronted with related content areas in more advanced physics classes. For instance, the understanding of the physical concept of density could be prepared in elementary school by teaching students that different materials of the same size can have different weights and, vice versa, that different materials of different sizes can have

the same weight (Edelsbrunner et al., 2018). Once acquired, such pre-conceptions can be refined in later physics classes when students learn the Archimedean principle.

Although the concept of the spiral curriculum is intuitively plausible and fully compatible with theories on cognitive development and learning (Demetriou & Spanoudis, 2018; Ireland & Mouthaan, 2020; M. Schwartz, 2009), thoroughly testing its potential benefits is a complex and costly endeavor that has rarely been undertaken in a scientifically sound manner. Indirect evidence for the usefulness of spiral curricula comes from studies demonstrating the impact of prior knowledge on the explanation of achievement differences in a domain, which remains after controlling for general intelligence (Simonsmeier et al., 2022; Tricot & Sweller, 2014). This impact of prior knowledge has been confirmed in adult samples in studies on expertise in various domains (Charness, 1991; Schneider et al., 1989), in longitudinal studies across childhood (Schalk et al., 2019; Shing & Brod, 2016; Staub & Stern, 2002), and in long-term effects of intervention studies (Hannula-Sormunen et al., 2020; Hardy et al., 2006; Schneider et al., 1997). However, the results are less clear when learning gains are used as outcome measures instead of test scores. Meta-analytic evidence by Simonsmeier et al. (2022) indicates that learners do not generally gain more knowledge with higher prior knowledge. Brod (2021) discusses why and when prior knowledge may support future learning by emphasizing three important aspects. First, prior knowledge can only be beneficial for future learning if it is activated when learners are confronted with new information. Second, the activated prior knowledge must be relevant for (i.e., have a nonarbitrary association with) the content to be learned to avoid misleading students or detracting attention. Finally, prior knowledge can be expected to be particularly helpful if it is congruent (i.e., in agreement) with the new knowledge to be acquired.

The “preparation for future learning” paradigm introduced by Bransford and Schwartz (1999) emphasizes the importance of transfer and conceptual change processes in a broad sense since educators

are hopeful that students will show evidence of transfer in a variety of situations: from one problem to another within a course, from one course to another, from one school year to the next, and from their years in school to their years in the workplace (p. 61).

Furthermore, “[f]uture learning frequently requires ‘letting go’ of previous ideas, beliefs, and assumptions. Effective learners resist ‘easy interpretations’ by simply assimilating new information into their existing schemas; they critically evaluate new information and change their views (accommodate) when necessary” (p. 93). For the usefulness of this paradigm, it is not even crucial that learning in the first step was successful (e.g., D. Schwartz et al., 2005). However, it is of utmost importance that the opportunity for early learning creates awareness of the content challenges and the effort required to overcome them—as this may prepare for future learning.

Relevant and congruent prior knowledge may vary in its format. One type of relevant prior knowledge could be that learners draw on automated procedures or chunks of information when processing new learning material. This activation would free working memory resources that can be invested in encoding and thinking about new information. Another type of relevant and congruent prior knowledge is subject-specific strategies that are applicable to a broad variety of content within a domain. For example, monitoring strategies in mathematical problem solving (Mevarech & Fridkin, 2006) or the evaluation of evidence with the help of the control-of-variables strategy in

science (Edelsbrunner et al., 2022; Schalk et al., 2019) are broadly applicable strategies. The kind of overlapping knowledge we focus on in the present study concerns conceptual knowledge, specifically, the consecutive construction and refinement of concepts. We investigate whether early-acquired basic physics concepts that are compatible with scientific explanations but are age-appropriately simplified prepare students for and support future physics concept learning. This preparation and support may result from extending, restructuring, generalizing, or abstracting from the existing basic concepts when students are faced with more advanced concepts and explanations. Demonstrating this type of consecutive concept construction as a consequence of revisiting related content would provide direct evidence for the effectiveness of organizing learning opportunities according to the idea of a spiral curriculum. However, is it realistic to expect that students will remember what they learned years ago and be able to recognize similarities between content taught by different teachers? From what is empirically known about knowledge transfer, there is no reason for high expectations.

### Making Knowledge Usable for Transfer

Although prior knowledge is necessary for future learning, it is not sufficient. Various circumstances may stand in the way of activating the appropriate knowledge and transferring it to new contexts. While the transferability of knowledge is the prime goal of schooling, psychological research has repeatedly shown that transfer across contexts and over time is hard to achieve (Barnett & Ceci, 2002; Detterman & Sternberg, 1993; Lobato & Hohensee, 2021). Numerous studies indicate that learners rarely transfer what they have learned across different problems. While experts in a field seem to have a “vaster amount of small memory structures, in addition to high-level structures or holistic processing” (Sala & Gobet, 2017, p. 183), such as recognizing isomorphic conceptual structures of superficially dissimilar problems and scenarios, novices in a field (such as students) typically fail to see such deep isomorphisms (Chi & VanLehn, 2012; Goldstone & Day, 2012; Gray & Holyoak, 2021). Consequently, whereas transfer of knowledge to solve problems of similar kind and within shorter time lags is frequently observed, transfer of knowledge to solve problems with different surface features and over longer periods of time is a rarely observed performance.

Researchers have realized for some time that they were looking for transfer in short-term experimental studies, a setting in which it is rather unlikely to occur (Barnett & Ceci, 2002). Limited time may prevent adequate initial learning and explain the absence of far transfer. In less artificial learning environments, learners have more time, and the distribution of the allocated time may support initial learning and allow for subsequent transfer. When learning requires conceptual change, as is the case in physics, spacing the learning time may be beneficial. Most students oscillate between their naïve beliefs and scientific conceptions for a long period of time (Vosniadou & Brewer, 1992). This oscillation provides opportunities to experience situations in which naïve beliefs might be useful and those in which the application of scientific conceptions is more helpful and productive (Ohlsson, 2009, 2013). Only when students have the opportunity to repeatedly activate and rethink their knowledge can they construct concepts that are generalizable across contexts. Schools naturally offer spaced learning opportunities because the lessons allocated to a subject are spread over days and weeks, and it is not rare for topics to be revisited years later to refine students’ understanding.

To make acquired knowledge useful in new situations, it must be organized in ways that enable learners to recognize relevant similarities between their knowledge and the novel situation (e.g., Chi & VanLehn, 2012). In the case of physics, the required knowledge should be organized according to conceptual principles rather than according to surface features (e.g., Koponen & Kokkonen, 2014). For example, simply learning facts about what kind of material sinks and what kind floats (e.g., solid bodies made of Styrofoam float, while those made of steel sink, or that hollow objects are pulled upward by the air inside them) is unlikely to be productive for future learning (Hardy et al., 2006). An effective preparation would be to convey the more abstract idea that objects with less weight than the amount of water they displace float, whereas objects with more weight than the amount of displaced water sink. When students understand that the relationship between the volume and weight of the immersed object and the volume and weight of the replaced medium is decisive, they may understand the commonalities between a hot air balloon and a ship; because of the buoyancy force, they both rise because they are pushed upward by the heavier medium (cold air or water, respectively). To prepare students for future learning, the aim of science education should be to promote the acquisition of concepts that are abstract enough to allow for generalization and transfer to superficially different situations.

However, even if the initial learning is successful, transfer may still fail because learners cannot see the similarities between their knowledge and the novel content or problems presented and therefore do not activate and retrieve their relevant knowledge. Expecting the direct application of previously acquired knowledge to new problems, as it is typically tested in short-term designs, is likely to overstrain even otherwise capable learners. At the same time, subtle hints about the similarity between prior knowledge and a novel problem can help learners recognize the similarity and increase the likelihood of transfer (e.g., Gray & Holyoak, 2021). For the aim of consecutive concept construction in school, combining the idea of the spiral curriculum with the notion of preparation for future learning (Bransford & Schwartz, 1999) leads to productive conceptualizations. A spiral curriculum describes normatively how content should be organized, typically, across several years of schooling. The notion of preparation for future learning takes the perspective of the learners. Regarding science education, early education should provide necessary first building blocks for discovering and understanding the laws of nature; it will take years of education in which scientific laws and concepts are repeatedly revisited to help learners grow and refine their conceptions toward a scientifically suitable conceptual understanding. Accepting, for example, that one's explanations and predictions need to be reconsidered because they are not in line with the outcomes of experiments remains crucial to learning in the sciences (Kuhn, 2010). Such small steps may prepare learners for following instructional steps, so that they can gradually build up knowledge across their science education, spanning several years in schools.

### **Design and Research Questions (RQs) of the Present Study**

From the previous section, we conclude that the overly pessimistic view on transfer derived from short-term experimental studies may not fully apply to school settings, particularly if a spiral curriculum targeted toward conceptual change is implemented. In the present study, we investigated consecutive concept construction in physics education from elementary to secondary school in regular classroom

settings. Would students apply what they have learned in basic elementary school curriculum units on “floating and sinking” and “air and atmospheric pressure” (i.e., their prior knowledge) to an advanced curriculum unit entitled “hydrostatic pressure and buoyancy force” taught in secondary school?

Despite the narrow focus on the content examined, the design of our study allows for broader conclusions about the usefulness of early physics education in general. Many requirements are placed on elementary school, primarily on promoting reading, writing, and mathematics. In non-English speaking countries as well as in multilingual countries, time resources are required for foreign/second language education. Although science lessons have a long tradition in elementary school, in many countries, including Switzerland, their main focus is on topics on biology and local geography, which is, of course, also justified (e.g., D-EDK, 2016; National Research Council, 2013). Given this competition for lesson times in elementary school, adding physics to the timetable is likely to raise concerns about teaching this topic at the expense of other subjects and topics. The results of the present study may help to better evaluate the pros and cons of implementing physics curricula in elementary school.

For the present study, we analyzed data from the Swiss MINT Study (SMS), a longitudinal study that has been conducted at regular schools (Edelsbrunner et al., 2018; Schalk et al., 2019). MINT is the acronym for mathematics, informatics, natural science, and technology. It is used in German-speaking countries and corresponds largely to STEM. Within the SMS, elementary school teachers were trained to teach four curriculum units on basic physics concepts with a focus on conceptual understanding. In the present study, we investigated how learning the basic curriculum units in elementary school prepares students to learn an advanced physics curriculum unit in secondary school. Two of the basic curriculum units (floating and sinking, air and atmospheric pressure) share overlapping concepts with the advanced curriculum unit, which focuses on hydrostatic pressure and buoyancy force, while the other two basic curriculum units (stability of bridges and sound and the spreading of sound) do not.

In Switzerland, elementary school comprises 6 years. Secondary school starts at Grade 7 with newly composed classes and new teachers. This regulation allows the realization of a quasi-experimental research design to investigate the question of whether students can take advantage of early physics education. In the newly composed secondary classes, only some of the students had undergone the basic physics curriculum units in elementary school (intervention group), while others had not (control group). We recruited secondary school teachers who taught science classes in Grade Levels 7 and 8 to implement the advanced curriculum unit on hydrostatic pressure and buoyancy force. If children from the intervention group activated and used their prior knowledge, they should have been better prepared for learning the advanced curriculum unit than those who did not complete the basic curriculum units in elementary school. Importantly, the secondary school teachers did not know which of their students belonged to which group. Moreover, the advanced curriculum did not make any direct and explicit reference to the basic curriculum units; that is, it was designed to be taught as a stand-alone curriculum unit. Designing the advanced curriculum in this way gave all students a fair chance to benefit from the instruction. Even though it is well known that hints about relevant prior knowledge support knowledge transfer, we decided to not provide any references to avoid actively disadvantaging children from the control group and to circumvent the critique that these references would



be decisive (i.e., to circumvent the accusation of implementing a strawman design).

We expected that children from the intervention group would benefit more from the advanced curriculum unit than children from the control group because they could make use of their knowledge from the basic curriculum unit. That is, we predicted transfer performance (the precise RQs are presented below). Based on a broad range of empirical findings, Barnett and Ceci (2002) developed an influential taxonomy to help explain how likely it is that knowledge will be applied and extended to a new context. The taxonomy comprises the transfer dimensions of the knowledge domain, the physical, temporal, functional, and social contexts and the modality of testing. We used these dimensions to precisely characterize the similarities between the basic curriculum units and the advanced curriculum unit, that is, their distance with regard to the required transfer. Concerning the knowledge domain in the basic curriculum units and the advanced curriculum unit, the transfer distance was moderate. The basic curriculum units shared some overlap with the advanced curriculum unit, particularly the context of water and buoyancy, which was part of the floating and sinking unit and the advanced unit. The curriculum units differed in many other surface features of the domain. For example, to achieve transfer from the conceptual understanding of air pressure in the Earth's atmosphere to the advanced curriculum unit, students had to recognize structural similarities of air pressure with hydrostatic pressure that are not visible on the surface level. Overall, despite some similarities, transfer of the knowledge domain appears demanding. Regarding the physical and social context, instruction and tests take place at school, but the teachers and the locations (elementary and secondary schools) are different.

The temporal context of our study reduced the likelihood of transfer because there was a time lag of 3 years on average between learning the basic curriculum units and learning the advanced curriculum unit (more details about the timing follow in the Method section). Two perspectives can be taken regarding the likelihood of transfer given this temporal context. On the one hand, based on the spacing effect, it can be useful to have a gap of at least some days, weeks, or even months between related learning units so that learners' knowledge has time to consolidate between the units (Carpenter, 2012). On the other hand, fadeout must be expected after a delay of multiple years (Bailey et al., 2017), although research employing the floating and sinking unit has shown that elementary school students can maintain a good share of their newly acquired knowledge for at least a year (Hardy et al., 2006). Overall, the temporal gap in our study made it likely that students could not retrieve all of their acquired knowledge from elementary school; at the same time, the knowledge that they could still retrieve might be strengthened because of the time gap. It could be consolidated and, as such, more easily transferred and related to the new content. Further aspects of the functional and social contexts as well as the modality of teaching and testing were quite similar between the basic curriculum units and the advanced curriculum unit. All units could be described as academic learning in a classroom setting (functional context). All teaching units implemented a guided inquiry-based approach with small-group activities, and all students were tested individually within classrooms (social context). In elementary and secondary school, multiple choice tests were applied; therefore, the modality of testing was the same across units. According to the taxonomy, the greatest challenges for transfer in the present study were the knowledge domain, particularly the

temporal context with its relatively large time gap between the basic curriculum units and the advanced curriculum unit.

Figure 1 presents the design of this study. The learning gains of each curriculum unit were measured by pre- and posttests. The order of the basic curriculum units in this diagram is merely exemplary. Elementary school teachers were free to choose the order in which they taught the four units.

The following three RQs were addressed:

*RQ1:* When does prior conceptual knowledge unfold its potential?

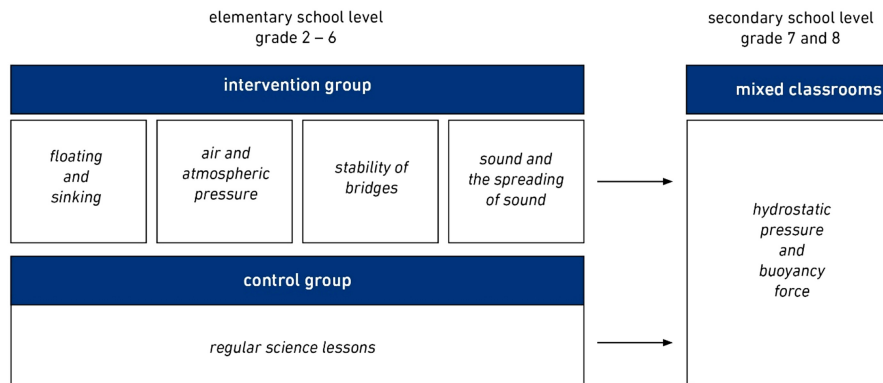
The advanced curriculum unit was preceded and followed by an identical pre- and posttest, both of which could reveal the hypothesized between-group differences in favor of the intervention group. The intervention group could outperform the control group in the pretest because their prior knowledge allowed them to draw conclusions when working on the pretest. It is also plausible that the advantage of the intervention group was revealed only while students completed the learning unit on hydrostatic pressure (in the sense of preparation for future learning). Given the learning experience with the basic curriculum units, students may have had an advantage in developing conceptual knowledge based on the content of the advanced curriculum unit. Specifically, we investigated the following five pathways of consecutive concept construction as a consequence of having received the basic curriculum units:

1. No effect: The intervention and the control group do not differ in the pretest or in the posttest.
2. Attenuation effect: The intervention group outperforms the control group in the pretest, but in the posttest, the difference declines or disappears because the advanced curriculum unit compensates for differences in prior knowledge (since it was designed as a stand-alone unit).
3. Learning-only effect: The intervention and the control group do not differ in the pretest, but the intervention group outperforms the control group in the posttest.
4. Constant effect: The intervention group outperforms the control group in the pretest and in the posttest to a similar extent.
5. Boosting effect: The intervention group outperforms the control group in the pretest and in the posttest, but the between-group difference in favor of the intervention group is larger in the posttest.

Investigating these five possible pathways can provide an answer to RQ1. We also considered the effects of gender in an additional exploratory analysis in which we extended this model to control for gender and to add its interaction with the early physics intervention. We did not have a central RQ or hypotheses regarding students' gender. However, since gender differences are an ongoing issue in STEM education research and specifically in physics education (Hofer & Stern, 2016), we examined potential effects to obtain initial insight into the role of early physics education in shaping gender differences and to generate hypotheses for future research.

In addition to comparing absolute learning gains between the intervention and control groups to answer RQ1, we aimed to investigate whether learners in the intervention group achieved similar or higher learning outcomes at the posttest when they started with the same knowledge as those in the control group at the pretest. This question is different from the question of the similarity of learning

**Figure 1**  
*Design of the Present Study*



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

gains posed in RQ1 (Köhler et al., 2021). Comparisons of learning gains are usually analyzed with models in the tradition of repeated-measures analyses of variance (ANOVAs), whereas comparisons of outcomes for learners starting with similar scores are usually analyzed with analysis of covariance (ANCOVA) models (e.g., Köhler et al., 2021; Lüdtke & Robitzsch, 2022). After using a repeated-measures model to investigate RQ1, we used an ANCOVA-type model to investigate the following question:

*RQ2:* Do students from the intervention group achieve higher learning outcomes than students from the control group after controlling for knowledge at pretest?

For this RQ, we employed a regression model with the pretest as a covariate, which resembled an ANCOVA structure. We examined whether, when controlling for the pretest, a variable indicating the treatment (intervention vs. control group) could explain further variation in students' learning outcomes at the posttest. An effect of the treatment beyond the pretest would indicate that learners in the intervention group achieved higher learning outcomes at the posttest that could not be explained by potential knowledge differences between the two groups that already existed at the pretest. After examining the main effect of condition, we again, similar to RQ1, added a main effect of gender as well as its interaction with the condition for an exploratory analysis of differential effects for male and female learners.

Learning in each of the four basic curriculum units was measured with tests on conceptual understanding. Due to the conceptual overlap with the target curriculum unit on hydrostatic pressure and buoyancy force, achievement in tests on floating and sinking and air and atmospheric pressure was expected to have greater predictive power than achievement in tests on the stability of bridges and on sound and the spreading of sound, leading to the third RQ.

*RQ3:* Can transfer effects be traced back to overlapping conceptual knowledge between the basic curriculum units and the advanced curriculum unit?

Only learners from the intervention group were included in the analysis of this RQ. Their posttest scores in the four curriculum units served as predictors for their achievement in the advanced curriculum on hydrostatic pressure and buoyancy force in a hierarchical

multiple regression model (Tabachnick et al., 2013). In the first model, we included posttest achievement in the topics of floating and sinking and air and atmospheric air pressure as predictors. Both of these topics showed an overlap of concept knowledge with hydrostatic pressure and buoyancy force. Consequently, we expected that learners' knowledge on these two topics would predict their posttest achievement on hydrostatic pressure and buoyancy force when controlling for pretest knowledge in this unit. In a second step, we added students' posttest achievement on the other two topics from the basic curriculum units (stability of bridges, sound and the spreading of sound) to the model. We expected that performance on these topics would have little or no predictive value for students' performance on the test on hydrostatic pressure and buoyancy force when controlling for the other two topics.

## Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All data, research materials, and analytic scripts are available on the Open Science Framework (OSF; <https://osf.io/94rxq/>). Data were analyzed using the R software environment Version 4.0.2 (R Core Team, 2021) via the tidyverse and brms packages (Bürkner, 2017; Wickham et al., 2019). The design and analyses were not preregistered. For statistical tests, we used 90% credible intervals (CIs) (see Method section for further details).

## Method

We randomly selected public elementary schools from urban and rural areas in the German-speaking part of Switzerland. Via the principals of the schools, teachers were asked to implement the four basic curriculum units. When a considerable portion of the students had finished elementary school, we started by contacting the neighboring secondary schools to which the students transitioned. We also invited the secondary school teachers via the school principals. Since these secondary schools were attended by students from several elementary schools, only a share of the students had received the basic curriculum unit in elementary school. This approach realized a quasi-experimental randomized design. Some of the students

had received the basic curriculum units (intervention group). The other students had not; they had received the standard science instruction (control group), which typically focuses on topics other than physics (predominantly biology and geography).

## Participants

The sample encompassed  $N = 1,375$  secondary school students (age at pretest = 13.64 years;  $SD = 0.77$ ; 679 male, 49.31%; 677 female, 49.16%; 18 missing values for gender) from 60 classes. The sample size was determined by ongoing recruitment that was finished when this article was written. Twenty-four classes received the advanced curriculum unit in Grade 7 ( $n = 501$ , 282 [56.29%] belonged to the control group), and 36 classes received it in Grade 8 ( $n = 874$ , 660 [75.51%] belonged to the control group). Overall, 942 students belonged to the control group ( $M_{\text{age}}$  at pretest = 13.69 years,  $SD = 0.78$ ,  $n = 459$  [48.73%] male), and 433 students belonged to the intervention group ( $M_{\text{age}}$  at pretest = 13.53 years,  $SD = 0.75$ ,  $n = 218$  [50.34%] male). Ethical approval was granted by Peter A. Edelsbrunner's institution. Parents gave consent to use the data for scientific purposes. They received a written description of the study, and the second author of this article (Ralph Schumacher) visited parent assemblies at the schools to describe the study.

The basic curriculum units contained teaching materials suitable for all elementary school grade levels and could thus be implemented in Grade Levels 1–6. The teachers were free to choose the grade level and the order in which they implemented the four basic curriculum units, but the large majority implemented them in Grade Levels 3 and 4. Due to unforeseen events in some classes, not all units could be implemented. Fluctuations and missing data also occurred at the student level because some students may have left school while the units were applied, or they missed either the pre- or the posttests because of sickness. Our criteria for being considered in the intervention group was having received at least one of the four units. The number and percentage of students from the intervention group who completed different numbers of units were as follows: 79 (18.24%) completed one unit, 84 (19.40%) completed two units, 127 (29.33%) completed three units, and 166 (38.34%) completed all four units. Overall, 326 (75.28%) of the intervention group students received instruction on the topic of air and atmospheric pressure, 331 (76.44%) received instruction on the topic of floating and sinking, 358 (82.68%) received instruction on the topic of sound and the spreading of sound, and 269 (62.12%) received instruction on the topic of the stability of bridges. The risk for selection bias within this quasiexperimental design was minimal, as school principals obliged their teachers to participate in our study, in Switzerland all teachers receive similar education and the same payment independently of school, and schools are generally diverse regarding socioeconomic background. No students opted out from the study in secondary school. The consent received from the intervention group in primary school covered their participation in the study in secondary school. Students and parents had the opportunity to opt out at any time point, but only a few parents or students did so in elementary school.

## Procedure: The Curriculum Units and Their Implementation

The basic curriculum units as well as the advanced curriculum unit were developed by science educators employed as researchers at universities. All curriculum units promoted the active construction

and restructuring of conceptual knowledge in an inquiry-learning setting. Means of cognitive activation, such as prompts for self-explanations, comparing and contrasting cases, and metacognitive questions, were included. Before experiments were conducted, students were asked to explain and discuss their own predictions (in the sense of predict–observe–explain cycles). This guided inquiry gave them the opportunity to realize the limits of their knowledge when their hypotheses did not match the outcome of an experiment. Subsequently, teachers promoted active knowledge construction by structuring and scaffolding children's discussions in search of the best explanation. Typically, several different experiments were used to explore the same physics concept, and the children were prompted to describe what the experiments had in common.

In elementary school and in secondary school, the units were taught by regular teachers who had undergone thorough trainings (one afternoon for each topic, delivered by Ralph Schumacher). The teachers were informed not only about the physics concepts and theories but also about typical intuitive conceptual frameworks and students' misconceptions that impede learning. In addition, they were provided with elaborated teaching materials for all lessons of a curriculum unit (e.g., detailed descriptions of the lessons, worksheets, texts for reading), equipment, and instruments for all experiments. They also learned how to conduct the hands-on experiments. Although the elementary school teachers were free to choose the order in which they provided instruction on the different topics, many received training on the topic of air and atmospheric pressure first. Consequently, approximately half of the students ( $n = 233$ , 54%) received instruction on this topic first, 53 students (12.24%) received instruction on floating and sinking first, 79 students (18.24%) received instruction on sound and the spreading of sound first, and 67 students (15.47%) received instruction on the stability of bridges first. A regression model testing for interactions of the first topic that was taught with the posttest achievement on the advanced curriculum unit indicated only very small effects (Table A1).

## Four Basic Curriculum Units for Elementary School Physics Education

The four basic curriculum units were developed at the University of Münster (Möller et al., 2006). The units underwent several scientific evaluations (e.g., Hardy et al., 2006; Kleickmann et al., 2010; Schalk et al., 2019). Each basic curriculum unit contained teaching materials for 14–16 lessons that were typically spread over several weeks. In each unit, the children underwent structured predict–observe–explain trials on their own and in small groups to experience phenomena and develop conceptual understanding of the underlying physics concepts under guidance of their teachers. The units are described elsewhere in detail (Hardy et al., 2006; Schalk et al., 2019). Here, we provide only a short overview of the central topics, aims, and examples of the instructional contents for each unit.

1. Floating and sinking: The main learning goal of this curriculum unit is for children to predict and explain why certain solid and hollow objects float while others sink. The conceptual understanding that children are supposed to develop is related to the concepts of water displacement and buoyancy force. Technical terms such as “density” and “buoyancy” are not used explicitly to avoid technical jargon at this age. Learners must overcome intuitive but naïve



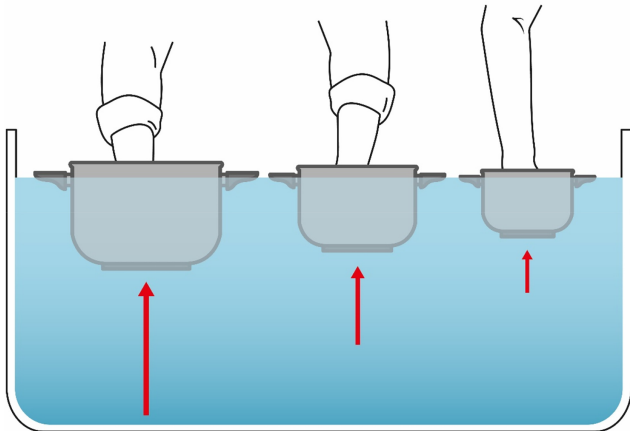
conceptions such as “light things float while heavy ones sink” or “hollow things are pulled upward by the air inside them.” Finally, the children also learn that objects that displace more water than they weigh float (and if they displace less water, then they sink). This conceptual understanding is supported through various teacher-guided experiments. For example, the children immerse three pots of different sizes into water and report which pots require the most effort to be pushed fully into the water (Figure 2). In this experiment, the children experience and reflect upon the fact that pushing larger objects down into water requires more effort than pushing down smaller objects. Whereas understanding buoyancy force is one of the central learning goals in this unit, its origin is not explained or explored.

2. Air and atmospheric pressure: Many children think that air has no or even negative weight (Hardy et al., 2006). The major learning goal of this unit is for children to develop an adequate conceptual understanding of air as being composed of matter that has weight, needs space, and interacts with its physical environment in specific ways. To this end, children conduct experiments in which they experience the material nature of air and its specific characteristics. In an experiment with a ball, they must predict whether the weight of the ball will be the same, increase, or decrease when it is inflated with air (Figure 3). After providing their predictions and observing on a scale that the weight of the ball increases after it has been inflated, they discuss reasons for this phenomenon, supported by their teachers who guide them in finding the appropriate explanations. In another experiment that is important for the present study, the children predict what happens when air is pumped out of a closed glass container that has a balloon filled with air inside it (Figure 3). After observing that the balloon increases in size, the children discuss with the teacher that this occurs because of a reduction in air pressure in the glass. Afterward, they learn that air also exerts pressure within the Earth’s atmosphere and that this pressure decreases with increasing altitude. In further

experiments, the children learn that air can slow falling toy parachutes and propel sailing ships and that it expands and rises when it is heated.

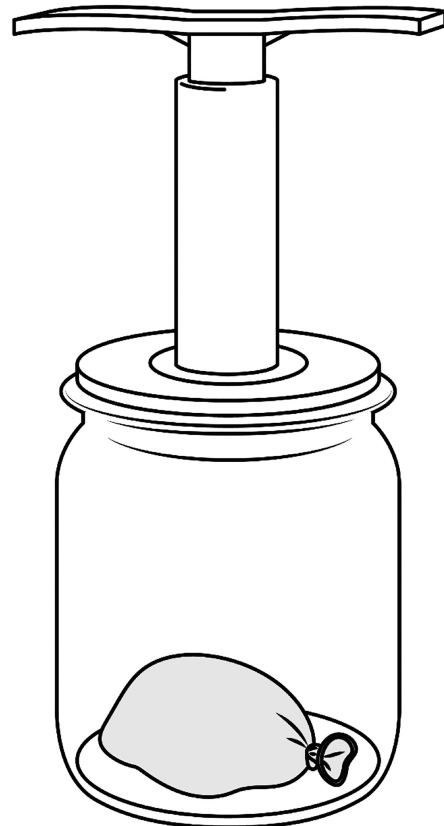
3. Sound and the spreading of sound: The goals of this unit are for children to understand that sound is produced by vibrations, and these vibrations are waves in air or other solid, liquid, or gaseous media. For example, students see and hear a ringing alarm clock in a glass container and predict what happens to the sound when air is pumped out of the container. After perceiving a decrease in the sound volume, the children discuss with their teachers that without a medium such as air, sound cannot travel, followed by a discussion of the central characteristics of sound waves. In another experiment, the students determine what variables affect the pitch and volume of sounds. They also acquire knowledge about the anatomy and functions of the human ear with regard to sound perception.
4. Stability of bridges: This teaching unit is aimed at developing initial knowledge about forces and mechanics. These concepts are introduced by building bridges and investigating the factors that affect the stability of bridges. In an experiment exemplifying counterbalance, the children add blocks to a wooden bridge that requires stabilization. The mechanical principle that vertical forces can be split into vertical and horizontal forces is illustrated with an arch bridge that needs

**Figure 2**  
*Depiction of an Exemplary Exercise in the Basic Curriculum Unit on Floating and Sinking*



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

**Figure 3**  
*Depiction of an Exemplary Exercise in the Basic Curriculum Unit on Air and Atmospheric Pressure*



lateral counter bearings to be stabilized. The children also acquire a conceptual understanding of profiles and how they contribute to stability.

### Advanced Curriculum Unit on Hydrostatic Pressure and Buoyancy Force

The advanced curriculum unit for secondary school encompassed the topics of hydrostatic pressure and buoyancy force. It was developed by the MINT-Learning Center of ETH Zurich under the supervision of the physicist Brigitte Hänger-Surer (coauthor of this article). Over the course of six lessons, students learned to explain how the buoyancy force in water is caused by differences in hydrostatic pressure. To give all students from the intervention and control groups a fair chance to benefit from this teaching unit, the first two lessons introduced the concept of buoyancy force with three simple experiments and made students aware that they lacked a causal explanation of what gives rise to this force. In Lessons 3–5, students experimentally explored hydrostatic pressure. For example, by measuring with simple manometers, they learned that hydrostatic pressure constantly increases with depth (see Figure 4).

Students learned from this instructional unit that the buoyancy force is caused by differences in hydrostatic pressure: Since hydrostatic pressure increases constantly with increasing depth, the hydrostatic pressure on an object under water from below is higher than the hydrostatic pressure from above (see Figure 5).

This difference in pressure results in an upward force in water, the buoyancy force. Students are guided stepwise by different instructions and activities (e.g., by completing an incomplete version of Figure 5). Based on this conceptual knowledge about the pressure difference, students are asked to construct an explanation of how the different hydrostatic pressures on an object under water cause the buoyancy force in the sixth lesson.

### Relations Between the Basic Curriculum Units and the Advanced Curriculum Unit

Two of the basic curriculum units, floating and sinking as well as air and atmospheric pressure, introduce concepts that are, from a physics perspective, further developed in the advanced curriculum unit. Specifically, the basic curriculum unit on floating and sinking in elementary school offers alternatives to the naïve belief that hollow objects are pulled upward. The unit rather supports a prequantitative understanding of the Archimedean principle: If an object is immersed in water, it is pushed up by the replaced water. Therefore, if an object has less weight than the amount of replaced water, it floats, whereas if it has more weight, it sinks. Buoyancy force is merely assumed as a fact in the basic curriculum unit without explaining its cause. The basic curriculum unit on air and atmospheric pressure focuses on the material nature of air. In addition, children learn that atmospheric air pressure decreases with altitude respectively increases with depth. This is illustrated by the analogy that “we all live on the ground of a sea of air” (Figure 6).

The physics concepts of these two basic curriculum units are embedded in different contexts while their conceptual contents are both related to the advanced curriculum unit. In line with the preparation for future learning paradigm, students who undergo the basic curriculum units experience the challenge of revising their often deeply rooted intuitive concepts and explanations. An example

**Figure 4**

*Manometer Used in the Advanced Curriculum Unit on Hydrostatic Pressure and Buoyancy Force*

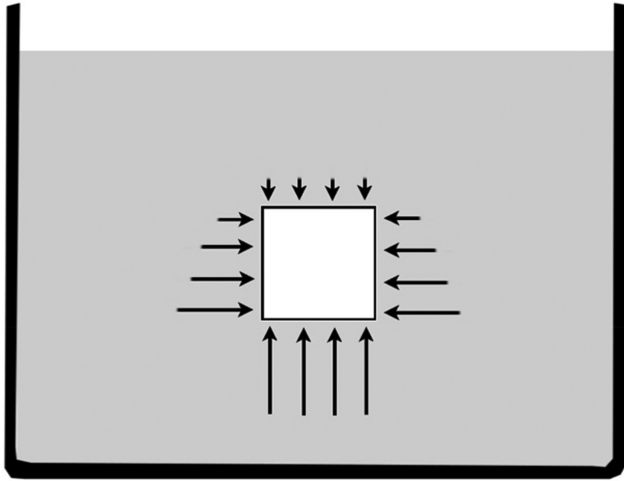


*Note.* ETH MINT = logo of the ETH MINT learning center. See the online article for the color version of this figure.

for triggering this challenge is an experiment in the basic curriculum unit in which students immerse pots of different sizes into water (see Figure 2). They experience that varying effort is needed to push objects of different volume into water: Pushing larger objects down into water requires more effort than pushing down smaller objects. Understanding that buoyancy force depends on the amount of displaced water contrasts with children’s typical intuition that objects float because the air contained in them pulls them upward. The conceptual change is thus based on the insight that objects float in water, not because they are pulled upward by the air contained in them, but because they are pushed upward by the displaced water. In this and similar experiments, children’s initial intuition is

**Figure 5**

*Exemplary Image Used for Explaining Hydrostatic Pressure and Buoyancy Force*



repeatedly challenged, but learners experience this challenge in a supportive environment. In collaborative discussions with their teachers and peers, learners are encouraged to express any beliefs without embarrassment. Undoubtedly, conceptual change is a tedious process, but the many opportunities offered in the multiweek basic physics units help to revise naïve or intuitive ideas of air pulling things upward through the concept of water pushing things upward. Developing an understanding that water is pushing immersed objects upward may prepare learning from the advanced unit, which focuses on establishing an understanding why differences in water pressure (as water pressure increases with depth) give rise to the buoyancy force. Moreover, having learned in the basic curriculum that atmospheric pressure decreases with altitude can help abstract and generalize the concept of buoyancy, so it becomes less bound to the water context.

### Assessments

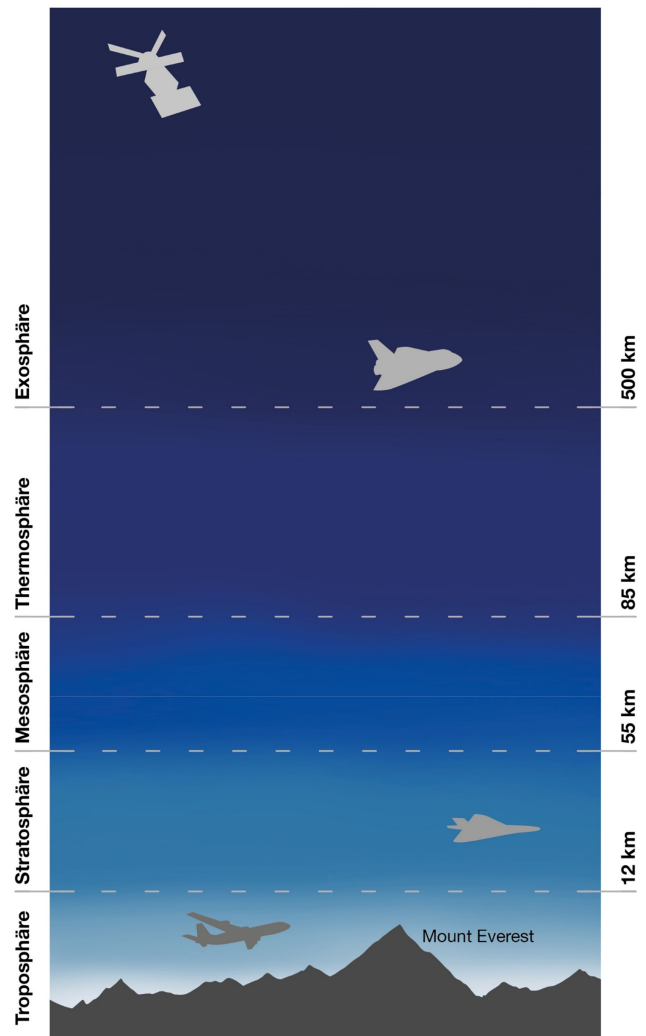
#### 1. Test of hydrostatic pressure and buoyancy force

Before and after the advanced curriculum unit, students completed a test on their conceptual understanding of hydrostatic pressure and buoyancy force, which contained 13 multiple choice questions (for an exemplary question about hydrostatic pressure, see Figure 7; for an exemplary question about buoyancy force, see Figure 8). The questions in the pre- and posttest were identical. Pre- and posttest solution rates served as dependent variables for the analyses.

Many distractor options in the questions covered misconceptions about hydrostatic pressure and buoyancy force (see Figures 7 and 8). Therefore, a decrease in selected distractors and a simultaneous increase in correct answers provided a proxy for conceptual restructuring. For each of the 13 questions, two points were given for a completely correct answer (all correct options marked with no incorrect option). If either correct option was left out or one incorrect option was marked, one point was given; otherwise, the score for the question was zero. Accordingly, the maximum score for the test was 26. The test was developed by our team under the supervision of the physicist Brigitte Hänger-Surer. It was administered as a

**Figure 6**

*Depiction Used to Explain That “We All Live on the Ground of a Sea of Air” in the Air and Atmospheric Pressure Basic Curriculum Unit (Labeling of Spheres in German)*



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

paper-pencil test by the teachers according to our instruction. The teachers returned the tests to us by mail, and they received feedback about the mean achievement gains in their classrooms. The estimated internal consistencies were  $\alpha = .53$ ,  $\omega = .54$  (see Dunn et al., 2014 for a description of  $\omega$ ) at pretest and  $\alpha = .64$ ,  $\omega = .65$  at posttest. Considering that the tests encompassed knowledge about multiple topics and concepts that were treated in the unit, these internal consistencies appeared adequate (Stadler et al., 2021; Taber, 2018). Note that internal consistency is not equated with reliability. Tests can be reliable and valid when their internal consistency is low (Neubauer & Hofer, 2022), particularly knowledge tests (Stadler et al., 2021; Taber, 2018).

#### 2. Tests on the four basic curriculum units

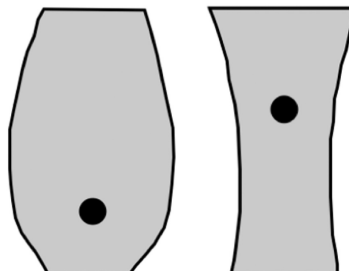
To address RQ3 on the specificity of the transfer, we used the scores of the multiple choice tests for the basic curriculum units (floating and

**Figure 7**

*Example Question From the Test on Hydrostatic Pressure*

Two differently shaped jars are filled with water. The hydrostatic pressure is measured at the marked two spots.

(1) At which of these two marked spots do we measure a higher hydrostatic pressure?



(2) Why is that so? Select all correct answers.

- because at the narrow part of the jar the water is compressed
- because at the wider part of the jar more water exerts pressure
- because the hydrostatic pressure increases with depth
- because the hydrostatic pressure is higher near the water surface
- because the water exerts more pressure, the deeper one gets

*Note.* The correct answer for the first part of the question is at the lower spot in the left jar, and for the second part, the third and fifth options are correct.

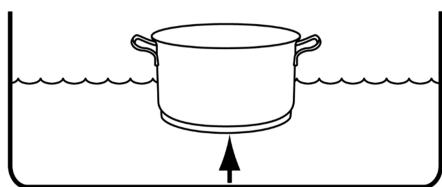
sinking: 11 items; air and atmospheric pressure: 15 items; sound and the spreading of sound: 17 items; stability of bridges: 18 items) as predictors. The tests were presented at the end of each of the four curriculum units to the children of the intervention group. The tests are described elsewhere in detail (Hardy et al., 2006; Schalk et al., 2019). Estimates of internal consistencies were  $\alpha = .91$ ,  $\omega = .92$  for the test on floating and sinking,  $\alpha = .70$ ,  $\omega = .72$  for air and

atmospheric pressure,  $\alpha = .79$ ,  $\omega = .80$  for sound and the spreading of sound, and  $\alpha = .87$ ,  $\omega = .87$  for stability of bridges. Considering the breadth of topics and concepts within the different tests, all internal consistencies appeared acceptable (Stadler et al., 2021; Taber, 2018). The average time gap between the posttests on the four basic curriculum units and the pretest on hydrostatic pressure and buoyancy force varied between 0.39 and 6.10 years ( $M = 3.19$  years,  $SD = 1.00$ ). In accordance with our assumption that the time gap represents a challenge to knowledge transfer, a model presented in the Appendix indicates that a longer gap between the basic curriculum units and the advanced curriculum unit predicts lower achievement on the hydrostatic pressure and buoyancy-force posttest within the intervention group (Table A2). Since this time gap only existed within the intervention group, we did not further consider it in our analyses.

**Figure 8**

*Example Question From the Test on Buoyancy Force*

Why does water exert an upward force on objects immersed into water?



- because the replaced water wants to get back to its place
- because the objects are light
- because the objects contain air
- because the hydrostatic pressure increases with depth
- because some materials float
- because water exerts a higher pressure from below than from the top

*Note.* The fourth and sixth options are correct.

### Analytic Approach

For statistical analyses, we employed multilevel modeling because the students in our sample were nested in school classes. Taking this nesting into account in multilevel models avoided bias (Barr et al., 2013; Kéry & Schaub, 2011; McElreath, 2020). We chose Bayesian estimation, which is less prone to estimation issues than frequentist analyses (König & van de Schoot, 2018), is bias-free in smaller samples (Birgé, 2015), and provides an intuitive integration and interpretation of multilevel parameters (Kéry & Schaub, 2011). For these and further reasons, Bayesian estimation has been implemented increasingly frequently in recent educational research (see, e.g., Berweger et al., 2023; Geary et al., 2021; Hausen et al., 2022; Merk et al., 2023; Schmidt et al., 2023). Note that the Bayesian approach mostly affects model estimation, not how the model is specified. Readers can



adopt their regular knowledge about general linear models and multi-level regression modeling in interpreting the parameters presented below.

We included a random intercept across learners' secondary school classrooms in all models. We also examined whether there was additional systematic variation stemming from students' elementary school classrooms. The intraclass correlation coefficient estimates for learners' elementary school classrooms were only between 0% and 2% across the four basic curriculum unit topics when controlling for dependence stemming from their secondary school classrooms, indicating that modeling these residual dependencies was not necessary. The random intercept of the posttest showed an intraclass correlation coefficient of .17 across secondary school classrooms, indicating that approximately 17% of variance in the posttest could be attributed to systematic differences between learners' classrooms for the more advanced unit. In addition to the random intercept, we added correlated random slopes for all further regression parameters apart from the effects of the pretest (which did not show variation across classrooms). We also extended the models to allow for heterogeneous residual variances across classrooms and for skewness in the posttest. With these model specifications, posterior predictive checks and residual plots, including qq plots, indicated appropriate model fit and adherence to statistical assumptions. Note that the parameter estimates from models that implemented multilevel structure can deviate from descriptive statistics despite good model fit.

All models converged without issues, as indicated by Rhat estimates of 1.00 for all parameters, the absence of divergent transitions in the estimation process, effective sample sizes above 500, and visual inspection of posteriors and mixing in posterior trace plots (for explanations of these indices, see Bürkner, 2017). The posteriors of all central model parameters were strongly unimodal. We implemented the models in the brms package (Bürkner, 2017). For the Hamiltonian Monte Carlo estimation of each model, we used four chains with 4,000 draws, of which 1,000 were treated as a warm-up, with no thinning and a high target acceptance rate of .99. We used default priors for model parameters apart from regression weights because these were the parameters of interest regarding our RQs. These default priors were student  $t$  distributions with 3  $df$ , locations of 0 and scales of 10 for variance parameters, as well as a Lewandowski–Kuwowicka–Joe (LKJ) distribution with one degree of freedom for random effect correlations. We did not deviate from these default priors as these specifications mapped well onto the expected score ranges of our measures. The LKJ-prior with one degree of freedom does not restrict random effect correlations, which was appropriate given that we did not have any expectations regarding plausible values for these parameters. For regression weights (i.e., slope parameters), we defined Gaussian priors with locations representing small effect sizes and variances of 10. A prior robustness check confirmed that these prior choices were weakly informative, implying that they did not noticeably influence the resulting model estimates.

We present and interpret the means of posterior distributions as parameter estimates. Bayesian models do not provide typical  $p$  values, but they have an informative alternative called CIs, which are akin to confidence intervals from frequentist statistics, but are less prone to typical misinterpretations (Hoekstra et al., 2014). For statistical inference, we interpreted 90% CIs (defined by highest density posterior intervals; McElreath, 2020) as follows. If the estimated CI

of a model parameter did not include zero, we interpreted this as evidence that the respective effect deviated from zero. If the estimated interval included zero, we interpreted its range but did not exclude the possibility that the respective parameter was zero (Sorensen & Vasishth, 2015). There is no standard range for the reporting of CIs (McElreath, 2020). We decided to use 90% CIs because they provide a rather high certainty of 90% for the unobserved parameters to fall within the respective interval. Note that this is a simple and intuitive interpretation that is not possible with classical, frequentist confidence intervals. An advantage of 90% intervals is that their limits are usually estimated more reliably than those of broader intervals, for example, 95% intervals (McElreath, 2020). We set up regression models akin to models from the  $t$  test and ANOVA families (these models are further described in the Results section). We decided to present the results in regression parameters instead of ANOVA-like sum of squares values because we prefer the interpretation of regression parameters (for a discussion of the use of sum of squares within a Bayesian framework, see Marsman et al., 2019).

There were 1%–3% missing data on students' posttest scores on the assessments of the basic curriculum units. We imputed the missing values directly in the models, an approach that is generally recommended in Bayesian modeling (McElreath, 2020). For imputation, we used the information of the remaining predictor variables for the estimation of missing data on each respective predictor variable. Imputation was only necessary for the two models evaluating RQ3; in all other models, there were no missing data.

## Results

### RQ1: When Does Prior Conceptual Knowledge Unfold Its Potential?

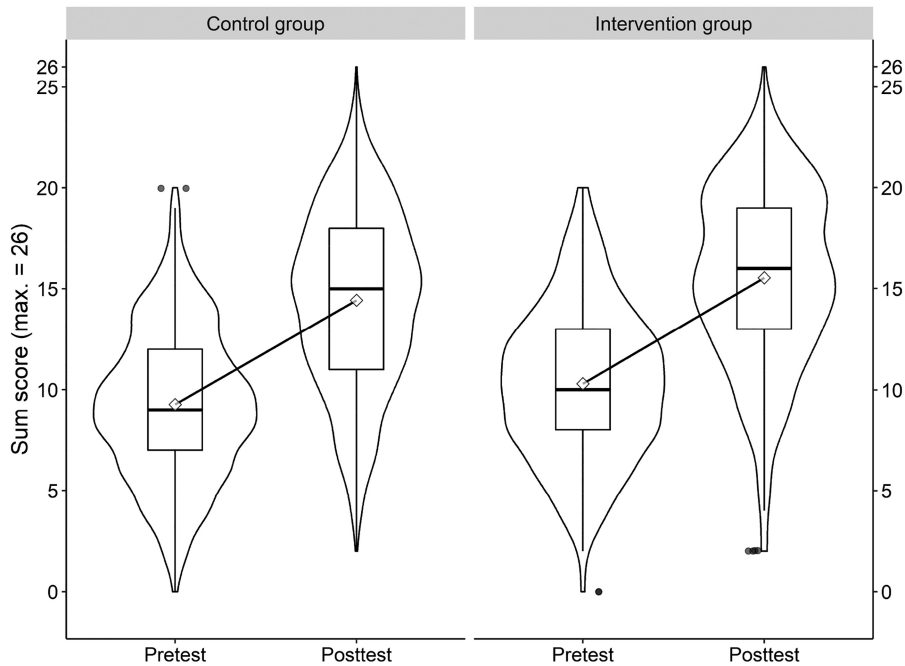
Regarding the question of whether the potential of the intervention unfolds before or after the advanced curriculum unit, Figure 9 depicts descriptive statistics and the full distributions of the pre- and posttests measuring students' knowledge about hydrostatic pressure and buoyancy force. At the pretest, students in the control group achieved a mean score of  $M = 9.27$  ( $SD = 3.70$ ), and those in the intervention group achieved a mean score of  $M = 10.29$  ( $SD = 3.76$ ). At the posttest, students in the control group achieved a mean score of  $M = 14.43$  ( $SD = 4.35$ ), and those in the intervention group achieved a mean score of  $M = 15.54$  ( $SD = 4.42$ ). The results of Bayesian multilevel regression models comparing these scores at pre- and posttest between conditions are presented in Table 1. The implemented models were akin to simple  $t$  tests, comparing scores on the pre- or posttest between the intervention and control groups but with a multilevel structure to correct for nesting in classrooms.

As shown in Figure 9 and confirmed by the regression results in Table 1, the students in the intervention group outperformed those in the control group at both pre- and posttest. Descriptively, the students in the intervention group showed an advantage of  $d = 0.28$ , 90% CI [0.18, 0.37] at pretest and a similar advantage of  $d = 0.25$ , 90% CI [0.15, 0.35] at posttest. The CIs of the effect of condition in Table 1 were clearly above 0 at both time points, corroborating a positive effect. Regarding our first RQ, we concluded that the potential of the intervention (i.e., having received the basic curriculum units in elementary school) was revealed at both the pre- and posttest.

Another aspect of RQ1 was to examine whether students in the two conditions differed in their learning gains from pre- to posttest.



**Figure 9**  
*Distributions of Scores on Hydrostatic Pressure and Buoyancy Force Pre- and Posttests in Control Group and Intervention Group*



*Note.* Violin shapes indicate densities of score distributions, overlaid with boxplots. Points above and below distributions indicate outliers. Squared points represent mean values, and lines connecting squared points visualize changes between pre- and posttests. max. = maximum.

Descriptively, students in the control group yielded a gain of  $d = 1.12$ , 90% CI [1.06, 1.17], and those in the intervention group yielded a gain of  $d = 1.15$ , 90% CI [1.07, 1.23]. To compare these gains between conditions, we implemented a Bayesian hierarchical repeated-measures ANOVA. In addition to the main effects of time (pre- vs. posttest) and condition (control vs. intervention group) and the interaction of these two variables, we added a random intercept for each student. This model structure represented a typical repeated-measures ANOVA (Field et al., 2012). We also added additional random effects as described in the Analytic Approach section to model the multilevel structure.

**Table 1**  
*Results From Bayesian Multilevel *t* Tests Comparing Hydrostatic Pressure and Buoyancy Force Pretest and Posttest Scores Between Conditions*

Parameter	Estimate	Error	90% CI
Pretest			
Intercept	9.18	.21	[8.86, 9.52]
Condition	0.78	.30	[0.29, 1.25]
Posttest			
Intercept	14.15	.31	[13.64, 14.65]
Condition	1.14	.37	[0.55, 1.76]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). The intercept represents the estimate in the control group. Random effects capturing the multilevel structure are available from the additional online materials which are available at the OSF page (<https://osf.io/94rxq/>). CI = credible interval; OSF = Open Science Framework.

The results from this model are presented in Table 2. As apparent from Figure 9 and confirmed by the model estimates in Table 2, students in both conditions gained similarly. Whereas the positive effect of condition confirmed that the students in the intervention group showed generally higher scores than those in the control group, this effect was similar at both time points. This similarity is indicated by the CI for the interaction between time and condition, which clearly includes 0 and excludes larger estimates. These results support the “continuous effect” hypothesis (Figure 1), according to which students in the intervention group would outperform those in the control group at both time points to a similar extent.

**Table 2**  
*Results From Bayesian Repeated-Measures Multilevel Model Regressing Hydrostatic Pressure and Buoyancy Force Test Scores on Time, Condition, and Their Interaction*

Parameter	Estimate	Error	90% CI
Intercept	9.17	.40	[8.84, 9.52]
Time	4.99	.33	[4.46, 5.54]
Condition	0.81	.31	[0.32, 1.34]
Time × Condition	0.31	.39	[-0.33, 0.95]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). Intercept represents estimate in control group at pretest. Random effects capturing the multilevel structure are available from the additional online materials which are available at the OSF page (<https://osf.io/94rxq/>). CI = credible interval; OSF = Open Science Framework.

The multilevel nature of the model allows us to determine whether the interaction between time and condition differed across teachers and their school classes. The model estimate indicated variation in the interaction effect across teachers and their classrooms, with an estimated variance of  $\sigma = 1.56$ . This value indicates rather large variation in the intervention effect among classrooms. Translating this value into Cohen's  $d$  scaled at the pooled standard deviation from the pretest, the model predicts treatment effects to encompass a range from  $d = -0.58$  to  $d = 0.75$  for 95% of classrooms.

We also conducted an additional exploratory analysis because the violin plots (Figure 9) indicated between-condition differences in the distribution of the posttest. The top end seemed to be more populated in the intervention group than in the control group. To descriptively examine this impression, we compared the proportion of students reaching a value in the highest quartile at posttest between conditions. For the entire sample, the percentile rank of 75 in the posttest on hydrostatic pressure and buoyancy force corresponded to a score of 18. Nineteen percent of the participants in the control group and 42% of the participants in the intervention group scored above 18. This post hoc analysis suggests that prior knowledge gained from the basic curriculum units may have helped students attain high achievement.

The results of gender differences are presented in the Appendix. In both groups, males slightly outperformed females in the pre- and the posttest on hydrostatic pressure ( $d = 0.10$ – $0.20$ ). There was no interaction between gender, condition, and time, with confidence intervals excluding large effects. Thus, no gender-specific effects of early physics education on more advanced learning could be detected.

### RQ2: Do Students From the Intervention Group Achieve Higher Learning Outcomes Than Students From the Control Group After Controlling for Knowledge at Pretest?

In addition to comparing learning gains with the repeated-measures ANOVA presented for RQ1, we examined whether students from both conditions with similar knowledge about hydrostatic pressure and buoyancy force at pretest showed comparable or different performance at posttest. To this end, we fitted a regression model in which we controlled for students' pretest knowledge and added a main effect of condition.

As shown in Table 3, the results indicated a positive effect of condition when controlling for the pretest. That is, students from the intervention group who started with the same knowledge at pretest as those in the control group could be expected to end up with an advantage of 0.81 points (a Cohen's  $d$  of 0.18) on the posttest. Despite students in both conditions gaining comparable amounts of knowledge from pre- to posttest (see results from RQ1), students in the intervention group who started with the same knowledge score as those in the control group achieved higher knowledge scores at posttest. In the discussion, we will go into more detail about how these diverging results regarding RQs 1 and 2 can be interpreted. As seen in the Appendix (Table A3), this analysis also did not reveal an interaction between gender and condition.

### RQ3: Can Transfer Effects Be Traced Back to Overlapping Conceptual Knowledge Between the Basic Curriculum Units and the Advanced Curriculum Unit?

To examine whether the effects of the early physics instruction could be traced back to overlapping conceptual knowledge (of the

**Table 3**

*Results From Bayesian Multilevel Model Regressing Hydrostatic Pressure and Buoyancy Force Posttest Score on Pretest Score and Condition*

Parameter	Estimate	Error	90% CI
Intercept	10.13	.40	[9.49, 10.81]
Pretest	0.44	.03	[0.39, 0.48]
Condition	0.81	.35	[0.24, 1.37]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). Random effects capturing the multilevel structure are available from the additional online materials which are available at the OSF page (<https://osf.io/94rxq/>). CI = credible interval; OSF = Open Science Framework.

floating and sinking and the air and atmospheric pressure units with the advanced curriculum unit), we implemented two multilevel multiple regression models to model the performance of the 433 participants of the intervention group.

The results from the two models are presented in Table 4. In the first model, we predicted students' posttest scores on the test on hydrostatic pressure and buoyancy force from their posttest scores from the basic curriculum units on floating and sinking and air and atmospheric pressure. In addition, we controlled for students' pretest scores on hydrostatic pressure and buoyancy force. The model showed that students' scores on the tests from these two basic curriculum units could indeed predict their learning achievement on the advanced curriculum unit beyond their pretest scores in the advanced curriculum unit. Both tests had standardized regression weights of 0.13, with CIs excluding zero (Table 4).

In a second step, we added the posttest scores on the other two basic curriculum units to the model to see whether these would have lower regression weights. As shown in Table 4, these tests did not further predict students' achievement on the hydrostatic pressure and

**Table 4**

*Results From Stepwise Standardized Bayesian Multiple Regression Model, Regressing Hydrostatic Pressure and Buoyancy Force Posttest Score on Pretest Score and on Posttest Achievement in the Floating and Sinking and Air and Atmospheric Pressure Curriculum Units (Step 1), With Additional Control for Posttest Achievement in the Sound and the Spreading of Sound and Stability of Bridges Units (Step 2)*

Parameter	Estimate	Error	90% CI
Step 1			
Intercept	0.18	.08	[0.04, 0.32]
Pretest	0.32	.05	[0.25, 0.40]
Floating and sinking posttest	0.13	.06	[0.03, 0.24]
Air and atmospheric pressure posttest	0.13	.06	[0.02, 0.23]
Step 2			
Intercept	0.11	.11	[−0.07, 0.29]
Pretest	0.25	.07	[0.14, 0.37]
Floating and sinking posttest	0.15	.11	[−0.01, 0.33]
Air and atmospheric pressure posttest	0.11	.10	[−0.05, 0.27]
Sound and the spreading of sound posttest	0.04	.09	[−0.11, 0.18]
Stability of bridges posttest	0.02	.09	[−0.13, 0.16]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). All variables are  $z$ -standardized. CI = credible interval.

buoyancy force posttest. These results indicate that content-specific transfer was the reason for the effect of the basic curriculum units.

### Discussion

Are basic physics units in elementary school a means to prepare secondary school students' understanding of more advanced physics concepts? On the basis of the results of this quasi-experimental randomized intervention study conducted in 60 classrooms of early secondary schools, the answer is a yes; even 3 years later, on average, students benefit from early physics education. Our design was based on the idea of a spiral curriculum (Ireland & Mouthaan, 2020), in which the content needs to be structured such that complex concepts can be understood at a simplified level first and then revisited at more complex levels later. In our study, a spiral curriculum was implemented to promote conceptual understanding of floating and sinking and air and atmospheric pressure. When the children were in secondary school, concepts were revisited in an advanced unit on hydrostatic pressure and buoyancy force presented by new teachers who were not aware of the basic curriculum units and did not know which of their students had received them. Our pre-post design allowed us to address RQ1, in which we asked when prior knowledge unfolds its potential. The intervention group outperformed the control group in the pretest. This was the case even though the pretest did not contain any items that were identical to the materials or tests applied in elementary school. It seems that the problems presented in the tests activated prior knowledge and stimulated reasoning processes that increased performance. The posttest results showed that both groups benefited considerably from the hydrostatic pressure and buoyancy force unit, indicating that prior knowledge had a constant effect on future learning. This result fits the "continuous effect" hypothesis, which states that learners in the intervention group started at a higher level because their prior knowledge (even without particular activation by hints about the basic curriculum units) allowed them to draw conclusions when working on the pretest.

A different perspective on the data were pursued with RQ2, in which we asked whether students from the intervention group achieve higher learning outcomes than students from the control group after controlling for knowledge at pretest (Lüdtke & Robitzsch, 2022). To answer this question, we varied the analytic approach by modeling whether a significant amount of variance in learning outcomes of the advanced curriculum could be traced back to receiving the basic curriculum units in elementary school after controlling for the pretest on hydrostatic pressure and buoyancy force. The results indicated that learners who started with similar knowledge in the intervention group outperformed their peers from the control group at posttest.

The apparent difference between the outcomes of the two statistical approaches (repeated measures vs. pretest-as-covariate) is well known and is labeled Lord's paradox (Pearl, 2016). Recent work has made advances in reconciling the apparent contradictions of the two approaches (Köhler et al., 2021; Lüdtke & Robitzsch, 2022; Pearl, 2016). Which statistical model is appropriate depends on the RQ and the assumptions made in the models. Since in our design the first intervention (i.e., the basic curriculum units) had taken place before the pretests of the more advanced curriculum unit, pretest differences were not the result of a sampling bias but rather characterized different learning opportunities. Therefore, it can be assumed that both approaches deliver informative and unbiased results (Lüdtke & Robitzsch, 2022), but they differ because they answer different questions. To better understand the differences in the results, it is helpful to

refer to the concepts of mediation analysis. As Pearl (2016) notes, the repeated-measures approach indicates the total effect of the early physics curriculum units on learners' gains on the advanced curriculum unit. The total effect worked partly indirectly through differences at pretest and partly directly on the posttest. Both effects together yielded an overall (i.e., total) effect close to zero that was visible in our repeated-measures model.

The ANCOVA approach, in contrast, models only the direct effect of the basic units on posttest achievement, controlling for effects via the pretest (Pearl, 2016). This direct effect was clearly positive, as evident from our analyses regarding RQ2. What do these insights from comparing the two models imply regarding the efficacy of our intervention? Overall, the models show that despite leading to largely comparable learning gains during the intervention, early physics instruction helped learners achieve higher learning outcomes at posttest. Thus, what it means to achieve a positive treatment effect can be interpreted differently, as shown in the two statistical approaches, but overall, early physics education clearly benefitted the learners. Although the two analyses revealed different benefits of the treatment effect (the difference at pretest vs. the pretest-controlled outcome), the finding that both analytic approaches showed positive effects shows robustness (Schweinsberg et al., 2021). This also concerned the effects of our early physics education on gender differences. In neither of the two analyses did we find specific beneficial effects for female or male students in the intervention group; male and female students were equally prepared for future learning by early physics instruction.

The ANCOVA applied to answer RQ2 indicates the importance of prior knowledge for future learning. If both groups start with the same knowledge at pretest, children who receive the basic curriculum units are likely to benefit more from the advanced curriculum unit in hydrostatic pressure and buoyancy force than those who do not. The longitudinal design of our study allows us to go beyond the impact of the pretest to pinpoint learning transfer based on conceptual knowledge. The source of transfer was addressed in RQ3. Our results revealed that we could trace transfer effects to overlapping conceptual knowledge. The intervention group received four basic physics curriculum units in elementary school, with two of them (floating and sinking, air and atmospheric pressure) sharing conceptual overlap with the advanced curriculum unit while the other two (stability of bridges, sound and the spreading of sound) did not. In our study, conceptual overlap was prepared by offering early learning opportunities that shifted children's attention from "objects being pulled upward" to the concept of mutual pressure between objects and the medium (e.g., air or water) in which they are located, and the resulting buoyancy force that pushes them upward. Regression analyses for the intervention group revealed that the superiority of the intervention group could be traced back to the content-specific transfer of concepts because only the tests on the overlapping units predicted achievement in the advanced curriculum unit. Studies have confirmed that overlapping knowledge constitutes transfer (Barnett & Ceci, 2002), but many studies have focused on common ground in procedures or facts. We provide evidence that learners can make use of the opportunity to transform basic conceptual understanding into more advanced knowledge.

### Implications and Limitations

The present results counter an overly pessimistic view on transfer of learning (e.g., De Bruyckere et al., 2020; Detterman & Sternberg, 1993). As discussed in the introduction, it has not been easy in

experimental short-term studies to show that previously acquired knowledge is used in new situations (Barnett & Ceci, 2002; Lobato & Hohensee, 2021). Expecting newly acquired knowledge to be immediately recognized as suitable for solving a new problem may be overly ambitious. In our longitudinal field study, learners were given extensive training in elementary school with the basic curriculum unit (14–16 lessons each) as well as in secondary school with the advanced curriculum unit (six lessons). With the lectures naturally spread over some time (since only a few lessons per week were devoted to physics education, especially in elementary school), the children had opportunities to frequently reactivate and restructure their knowledge. Based on our findings, we suggest that research on knowledge transfer has a chance of solid and realistic outcomes if longitudinal designs with more extensive training are applied. Such designs also have much higher external validity than short-term interventions given the way lesson plans and school curricula are designed.

The intervention group's advantage in the advanced curriculum unit was significant, but given  $d = 0.28$  at pretest and  $d = 0.25$  at posttest, it might not appear overwhelmingly large according to traditional standards (Lord, 1967). However, these standards have been frequently criticized (see, e.g., Bakker et al., 2019) since they do not reflect realistic expectations about the effect sizes found in (quasi) experimental designs (Schäfer & Schwarz, 2019), and they fail to consider that studies differ in designs and aims (Bakker et al., 2019; Kraft, 2020). In comparison to meta-analytic findings from studies examining the effects of inquiry-based approaches on students' learning, our effects appear promising. One meta-analysis reported larger average effect sizes of  $d = 0.65$  (Schroeder et al., 2007) and  $d = 0.65$  for teacher-guided inquiry settings (Furtak et al., 2012). These meta-analyses all focused on short-term elementary learning outcomes of a broad variety. Other meta-analyses reported more modest effect size estimates. For example, Hattie's (2008) meta-meta-analysis reported an average effect of  $d = 0.50$  on students' achievement, but with important nuances. That is, meta-analyses that distinguish between effects on different outcomes generally report stronger effects on process outcomes (i.e., inquiry skills;  $d = 0.52$  in Bredderman, 1983;  $d = 0.40$  in Shymansky et al., 1990) than on content knowledge (Bredderman, 1983:  $d = 0.16$ ; Shymansky et al., 1990:  $d = 0.26$ ). Hattie infers (2008, p. 377) that "[i]nquiry learning seems more successful for improving inquiry skills, but maybe other methods are needed to also add more positive effects on content and surface knowledge." Our study employed an innovative method to test this conjecture. Although our effects do not reach the extent of some well-known studies that have found large effects of guided inquiry on content knowledge (e.g., Hardy et al., 2006), it is important to consider that our effects capture knowledge transfer over multiple years and across different learning contexts and that they require transfer to an advanced topic; thus, they appear to be of notable magnitude for theory and practice.

The effects of our study must be further interpreted in light of the realistic context in which our study was conducted. All curriculum units were applied in real classrooms by the normal teachers. Although the teachers underwent the same training, our multilevel analyses showed that their potential in implementing the target units differed. Achievement differences at the classroom level may reflect teachers' characteristics or the student composition of the classroom. Future research is needed to identify the causes of the differences among classrooms. These differences might encompass factors such as teachers' pedagogical beliefs, their implementation of instructional techniques such as scaffolding, or the quality of their verbal

interactions with students, all of which have been found to affect teaching and learning in inquiry-based instruction in (early) physics (Hadley et al., 2022; Herrmann et al., 2021; Kleickmann, 2008; Kleickmann et al., 2010; Studhalter et al., 2021).

Moreover, to give the control group a fair chance, the target curriculum was designed as a stand-alone teaching unit. Accordingly, the prior knowledge acquired in elementary school could help students understand the content from the advanced curriculum unit (as shown by the present results), but the unit was designed to introduce all relevant aspects to avoid intrinsically disadvantaging children from the control group. Our goal was to determine whether having learned basic aspects at an earlier time would nonetheless be supportive. It is likely that group differences would have been larger if the teachers of the advanced curriculum unit had knowledge about the basic curriculum units to provide them with the opportunity to make direct references to their content and thus to provide hints for the students. However, this knowledge would have been in conflict with our goal of investigating the impact of a spiral curriculum under realistic school conditions. In Switzerland, and many other countries, elementary and secondary students are taught by different teachers who receive different trainings in their teacher education, and the two groups often have little opportunity for exchange. The visible yet moderate effects achieved in our study indicate that the idea of a spiral curriculum, which is, for example, visible in the Next Generation Science Standards (National Research Council, 2013), might contribute to improved learning outcomes. In addition, our post hoc analysis suggested that prior knowledge gained from the basic curriculum units may have helped students attain high achievement. In the posttest on hydrostatic pressure and buoyancy force, the percentage of high-performing students (percentile rank  $> 75$ ) was more than twice as high in the intervention group (42%) than in the control group (19%). Learners obviously differ in the extent to which they benefit from the early learning opportunities. Future research should explore the reasons for such interindividual variation. This requires taking cognitive and motivational characteristics as well as factors specific to each learning environment into account.

Overall, although our effects sizes appear moderate, they should be seen in light of the transfer distance across time and context (Barnett & Ceci, 2002). Furthermore, given the design characteristics of our study, they are in line or even above the effects that can be expected in larger field trials (Kraft, 2020; Schiefer et al., 2021). From these perspectives, investing in spiral curricula might be worth the effort, particularly when considering that factors other than content knowledge (e.g., attitudes, beliefs, and skills) might also be beneficial. The idea of spiral curriculum was proposed by Bruner (1960) before contemporary cognitive theories of learning by knowledge construction were developed. While Bruner concentrated on changes in the mode of representation (enactive, iconic, symbolic), later theories focused on the lengthy and arduous process of restructuring knowledge so it allows for complex problem solving and flexible transfer. Theories on conceptual change as well as the preparation for future learning approach emphasize that it is unrealistic to assume leaps in learning gains, no matter how stimulating the learning environments are. However, a stimulating learning environment for early science education can encourage learners to think further and question their beliefs as they are repeatedly exposed to topics with overlapping concepts embedded in different contexts. Future research should focus on the process of revising knowledge in the short term immediately after the teaching unit as well as in the long term. As the intervention group already showed better performance compared to the control group in the pretest on hydrostatic pressure,



at least part of the learners managed to preserve conceptual understanding developed in the basic curriculum units across several years. What exactly may have caused this long-term retention is an exciting direction for future research. Such research should include conative process factors likely to modulate science learning, such as self-concept, interest, and motivation, as well as moderating factors explaining variation in transfer such as the intellectual investment trait need for cognition (Liu & Nesbit, 2024).

By showing the content-specific long-term effects of early physics instruction, our study can inform decisions about the topics to be addressed in early science curricula. Many early science curricula aim to promote domain-general competencies such as the understanding of inquiry and the control of variable strategy, but choosing an area that fits a spiral curriculum can have a double benefit (i.e., benefiting also science content knowledge) in the long run.

## 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>
- Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics, 102*(1), 1–8. <https://doi.org/10.1007/s10649-019-09908-4>
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Berweger, B., Kracke, B., & Dietrich, J. (2023). Preservice teachers' epistemic and achievement emotions when confronted with common misconceptions about education. *Journal of Educational Psychology, 115*(7), 951–968. <https://doi.org/10.1037/edu0000792>
- Birgé, L. (2015). About the nonasymptotic behavior of Bayes estimators. *Journal of Statistical Planning and Inference, 166*(5), 67–77. <https://doi.org/10.1016/j.jspi.2014.07.009>
- Bransford, J., & Schwartz, D. (1999). Rethinking transfer. *Review of Research in Education, 24*(1), 61–100. <https://doi.org/10.3102/0091732X02400106>
- Bredderman, T. (1983). Effects of activity-based elementary science on student outcomes: A quantitative synthesis. *Review of Educational Research, 53*(4), 499–518. <https://doi.org/10.3102/00346543053004499>
- Brod, G. (2021). Toward an understanding of when prior knowledge helps or hinders learning. *npj Science of Learning, 6*(1), Article 24. <https://doi.org/10.1038/s41539-021-00103-w>
- Bruner, J. (1960). *The process of education*. Harvard University Press.
- Bürkner, P. C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology, 21*(1), 13–19. [https://doi.org/10.1016/S0193-3973\(99\)00046-5](https://doi.org/10.1016/S0193-3973(99)00046-5)
- Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283. <https://doi.org/10.1177/09637214124527>
- Ceci, S. J., Williams, W. M., & Barnett, M. (2009). Women's underrepresentation in science. Sociocultural and biological considerations. *Psychological Bulletin, 135*(2), 218–261. <https://doi.org/10.1037/a0014412>
- Charness, N. (1991). Expertise in chess: The balance between knowledge and search. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 39–63). Cambridge University Press.
- Chi, M. T., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47*(3), 177–188. <https://doi.org/10.1080/00461520.2012.695709>
- De Bruyckere, P., Kirschner, P. A., & Hulshof, C. D. (2020). If you learn A, will you be better able to learn B? *American Educator, 44*(1), 30–34.
- De Corte, E. (2003). Transfer as the productive use of acquired knowledge, skills, and motivations. *Current Directions in Psychological Science, 12*(4), 142–146. <https://doi.org/10.1111/1467-8721.01250>
- D-EDK. (2016). *Lehrplan 21—Natur, Mensch, Gesellschaft*. <https://v-ef.lehrplan.ch/downloads.php>
- Dehaene, S. (2011). The massive impact of literacy on the brain and its consequences for education. *Human Neuroplasticity and Education, 117*, 19–32.
- Demetriou, A., & Spanoudis, G. (2018). *Growing minds: A developmental theory of intelligence, brain, and education*. Routledge.
- Detterman, D. K., & Sternberg, R. J. (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Ablex Publishing.
- DiSessa, A. A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 155–187). Routledge.
- Dunn, T. J., Baguley, T., & Brunson, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- 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 skills 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.
- Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasiexperimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300–329. <https://doi.org/10.3102/0034654312457206>
- Geary, D. C., Hoard, M. K., Nugent, L., & Scofield, J. E. (2021). In-class attention, spatial ability, and mathematics anxiety predict across-grade gains in adolescents' mathematics achievement. *Journal of Educational Psychology, 113*(4), 754–769. <https://doi.org/10.1037/edu0000487>
- Goldstone, R. L., & Day, S. B. (2012). Introduction to “new conceptualizations of transfer of learning.” *Educational Psychologist, 47*(3), 149–152. <https://doi.org/10.1080/00461520.2012.695710>
- Gray, M., & Holyoak, K. J. (2021). Teaching by analogy: From theory to practice. *Mind, Brain, and Education, 15*(3), 250–263. <https://doi.org/10.1111/mbe.12288>
- Hadley, E. B., Barnes, E. M., Wiernik, B. M., & Raghavan, M. (2022). A meta-analysis of teacher language practices in early childhood classrooms. *Early Childhood Research Quarterly, 59*(2), 186–202. <https://doi.org/10.1016/j.ecresq.2021.12.002>
- Hammer, D., & Elby, A. (2003). Tapping epistemological resources for learning physics. *Journal of the Learning Sciences, 12*(1), 53–90. [https://doi.org/10.1207/S15327809JLS1201\\_3](https://doi.org/10.1207/S15327809JLS1201_3)
- Hannula-Sormunen, M., Nanu, C., Luomaniemi, K., Heinonen, M., Sorariutta, A., Södervik, I., & Mattinen, A. (2020). Promoting spontaneous focusing on numerosity and cardinality-related skills at day care with one, two, how many and count, how many programs. *Mathematical Thinking and Learning, 22*(4), 312–331. <https://doi.org/10.1080/10986065.2020.1818470>
- Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Why does a large ship of iron float? Effects of instructional support in 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>



- Hattie, J. (2008). *Visible learning*. Routledge.
- Hausen, J. E., Möller, J., Greiff, S., & Niepel, C. (2022). Students' personality and state academic self-concept: Predicting differences in mean level and within-person variability in everyday school life. *Journal of Educational Psychology, 114*(6), 1394–1411. <https://doi.org/10.1037/edu0000760>
- Herrmann, A., Bürgermeister, A., Lange-Schubert, K., & Saalbach, H. (2021). Die Bedeutung von Partizipation und Scaffolding für die Leistung im naturwissenschaftlichen Sachunterricht in Klassen mit hohem und niedrigem Anteil mehrsprachiger Schüler\*innen [The role of participation and scaffolding for science learning in primary school classes with high and low proportion of multilingual pupils]. *Zeitschrift für Grundschulforschung, 14*(2), 305–323. <https://doi.org/10.1007/s42278-021-00112-z>
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- Hofer, S., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology, 110*(8), 1175–1191. <https://doi.org/10.1037/edu0000266>
- Hofer, S., & Stern, E. (2016). Underachievement in physics: When intelligent girls fail. *Learning and Individual Differences, 51*, 119–131. <https://doi.org/10.1016/j.lindif.2016.08.006>
- Ireland, J., & Mouthaan, M. (2020). Perspectives on curriculum design: Comparing the spiral and the network models. *Research Matters, 30*, 7–12. <https://www.cambridgeassessment.org.uk/research-matters>
- Kéry, M., & Schaub, M. (2011). *Bayesian population analysis using WinBUGS: A hierarchical perspective*. Academic Press.
- Kleickmann, T. (2008). *Zusammenhänge fachspezifischer Vorstellungen von Grundschullehrkräften zum Lehren und Lernen mit Fortschritten von Schülerinnen und Schülern im konzeptuellen naturwissenschaftlichen Verständnis* [Relations of subject-specific beliefs of elementary school teachers on teaching and learning with students' gains in conceptual science understanding]. Doctoral dissertation, University of Münster.
- Kleickmann, T., Vehmeyer, J., & Möller, K. (2010). Zusammenhänge zwischen Lehrervorstellungen und kognitivem Strukturieren im Unterricht am Beispiel von Scaffolding-Maßnahmen [Relations of elementary teachers' subject-specific beliefs about teaching and learning with students' gains in conceptual science understanding]. *Unterrichtswissenschaft, 38*(3), 210–228.
- Köhler, C., Hartig, J., & Naumann, A. (2021). Detecting instruction effects—deciding between covariance analytical and change-score approach. *Educational Psychology Review, 33*(3), 1191–1211. <https://doi.org/10.1007/s10648-020-09590-6>
- König, C., & van de Schoot, R. (2018). Bayesian Statistics in educational research: A look at the current state of affairs. *Educational Review, 70*(4), 486–509. <https://doi.org/10.1080/00131911.2017.1350636>
- Koponen, I. T., & Kokkonen, T. (2014). A systemic view of the learning and differentiation of scientific concepts: The case of electric current and voltage revisited. *Frontline Learning Research, 2*(3), 140–166. <https://doi.org/10.14786/flr.v2i3.120>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 497–523). Wiley-Blackwell.
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681–718. <https://doi.org/10.3102/0034654315627366>
- Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review, 77*(3), 463–485. <https://doi.org/10.1177/0003122412440802>
- Liu, Q., & Nesbit, J. C. (2024). The relation between need for cognition and academic achievement: A meta-analysis. *Review of Educational Research, 94*(2), 155–192. <https://doi.org/10.3102/00346543231160474>
- Lobato, J., & Hohensee, C. (2021). Current conceptualizations of the transfer of learning and their use in STEM education research. In C. Hohensee & J. Lobato (Eds.), *Transfer of learning* (pp. 3–25). Springer.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304–305. <https://doi.org/10.1037/h0025105>
- Loverude, M. E., Heron, P. R. L., & Kautz, C. H. (2010). Identifying and addressing student difficulties with hydrostatic pressure. *American Journal of Physics, 78*(1), 75–85. <https://doi.org/10.1119/1.3192767>
- Lüdtke, O., & Robitzsch, A. (2022). *ANCOVA versus change score for the analysis of two-wave data: An overview of different modeling decisions*. PsyArXiv. <https://psyarxiv.com/ajf9d/>
- Marsman, M., Waldorp, L., Dablander, F., & Wagenmakers, E. J. (2019). Bayesian Estimation of explained variance in ANOVA designs. *Statistica Neerlandica, 73*(3), 351–372. <https://doi.org/10.1111/stan.12173>
- Mazur, E. (2015). *Principles and practice of physics*. Pearson.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
- Merk, S., Ophoff, J. G., & Kelava, A. (2023). Rich data, poor information? Teachers' perceptions of mean differences in graphical feedback from statewide tests. *Learning and Instruction, 84*(4), Article 101717. <https://doi.org/10.1016/j.learninstruc.2022.101717>
- Mevarech, Z., & Fridkin, S. (2006). The effects of IMPROVE on mathematical knowledge, mathematical reasoning and meta-cognition. *Metacognition and Learning, 1*(1), 85–97. <https://doi.org/10.1007/s11409-006-6584-x>
- Möller, K., Hardy, I., Jonen, A., Kleickmann, T., & Blumberg, E. (2006). Naturwissenschaften in der Primarschule: Zur Förderung konzeptuellen Verständnisses durch Unterricht und zur Wirksamkeit von Lehrerfortbildungen [Science in primary school. On the support of conceptual understanding through instruction and on the efficacy of teacher further education]. In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule: Abschlussbericht des DFG-Schwerpunktprogramms* [Studies on the instructional quality of schooling] (pp. 161–193). Waxmann.
- National Research Council. (2013). *Next generation science standards: For states, by states*.
- Neubauer, A. C., & Hofer, G. (2022). (Retest-) Reliable and valid despite low alphas? An example from a typical performance situational judgment test of emotional management. *Personality and Individual Differences, 189*, Article 111511. <https://doi.org/10.1016/j.paid.2022.111511>
- Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist, 44*(1), 20–40. <https://doi.org/10.1080/00461520802616267>
- Ohlsson, S. (2013). Beyond evidence-based belief formation: How normative ideas have constrained conceptual change research. *Frontline Learning Research, 1*(2), 70–85. <https://doi.org/10.14786/flr.v1i2.58>
- Pearl, J. (2016). Lord's paradox revisited—(oh Lord! Kumbaya!). *Journal of Causal Inference, 4*(2), Article 0021. <https://doi.org/10.1515/jci-2016-0021>
- Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211–227. <https://doi.org/10.1002/scs.v66:2>
- Potvin, P., & Cyr, G. (2017). Toward a durable prevalence of scientific conceptions: Tracking the effects of two interfering misconceptions about buoyancy from preschoolers to science teachers. *Journal of Research in Science Teaching, 54*(9), 1121–1142. <https://doi.org/10.1002/tea.21396>
- Potvin, P., Nenciovici, L., Malenfant-Robichaud, G., Thibault, F., Sy, O., Mahhou, M. A., Bernard, A., Allaire-Duquette, G., Sarrasin, J. B., Brault Foisy, L.-M., Brouillette, N., St-Aubin, A.-A., Charland, P., Masson, S., Riopel, M., Tsai, C.-C., Bélanger, M., & Chastenay, P. (2020). Models of conceptual change in science learning: establishing an exhaustive inventory based on support given by articles published in major journals. *Studies in Science Education, 56*(2), 157–211. <https://doi.org/10.1080/03057267.2020.1744796>
- R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>

- Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, 26(6), 515–520. <https://doi.org/10.1177/0963721417712760>
- Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10, Article 813. <https://doi.org/10.3389/fpsyg.2019.00813>
- 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(1), 34–45. <https://doi.org/10.1016/j.learninstruc.2018.09.006>
- Schiefer, J., Stark, L., Gaspard, H., Wille, E., Trautwein, U., & Golle, J. (2021). Scaling up an extracurricular science intervention for elementary school students: It works, and girls benefit more from it than boys. *Journal of Educational Psychology*, 113(4), 784–807. <https://doi.org/10.1037/edu0000630>
- Schmidt, K., Edelsbrunner, P. A., Rosman, T., Cramer, C., & Merk, S. (2023). When perceived informativity is not enough. How teachers perceive and interpret statistical results of educational research. *Teaching and Teacher Education*, 130, Article 104134. <https://doi.org/10.1016/j.tate.2023.104134>
- Schneider, W., Körkel, J., & Weinert, F. E. (1989). Domain-specific knowledge and memory performance: A comparison of high- and low-aptitude children. *Journal of Educational Psychology*, 81(3), 306–312. <https://doi.org/10.1037/0022-0663.81.3.306>
- Schneider, W., Küspert, P., Roth, E., Visé, E., & Marx, H. (1997). Short- and long-term effects of training phonological awareness in kindergarten: Evidence from two German studies. *Journal of Experimental Child Psychology*, 66(3), 311–340. <https://doi.org/10.1006/jecp.1997.2384>
- Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T. Y., & Lee, Y. H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, 44(10), 1436–1460. <https://doi.org/10.1002/tea.20212>
- Schwartz, D., Bransford, J., & Sears, D. (2005). Efficiency and innovation in transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–51). Information Age Publishing.
- Schwartz, M. (2009). Cognitive development and learning: Analyzing the building of skills in classrooms. *Mind, Brain, and Education*, 3(4), 198–208. <https://doi.org/10.1111/j.1751-228X.2009.01070.x>
- Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C., Van Assen, M. A., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Prasad, V. V., Bernstein, A., Robinson, E., Snellman, K., Sommer, S. A., Otner, S. M. G., Robinson, D., ... Schulte-Mecklenbeck, M. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, 165(4), 228–249. <https://doi.org/10.1016/j.obhdp.2021.02.003>
- Shing, Y. L., & Brod, G. (2016). Effects of prior knowledge on memory: Implications for education. *Mind, Brain, and Education*, 10(3), 153–161. <https://doi.org/10.1111/mbe.12110>
- Shtulman, A., & Harrington, K. (2016). Tensions between science and intuition across the lifespan. *Topics in Cognitive Science*, 8(1), 118–137. <https://doi.org/10.1111/tops.12174>
- Shtulman, A., & Walker, C. (2020). Developing an understanding of science. *Annual Review of Developmental Psychology*, 2(1), 111–132. <https://doi.org/10.1146/annurev-devpsych-060320-092346>
- Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60's on student performance. *Journal of Research in Science Teaching*, 27(2), 127–144. <https://doi.org/10.1002/tea.3660270205>
- Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science*, 11(5), 655–661. <https://doi.org/10.1111/j.1467-7687.2008.00714.x>
- 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>
- Sorensen, T., & Vasisht, S. (2015). *Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists*. arXiv. <https://arxiv.org/abs/1506.06201>
- Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, 60, Article 100832. <https://doi.org/10.1016/j.newideapsych.2020.100832>
- Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasiexperimental evidence from elementary mathematics. *Journal of Educational Psychology*, 94(2), 344–355. <https://doi.org/10.1037/0022-0663.94.2.344>
- Steegh, A. M., Höffler, T. N., Keller, M. M., & Parchmann, I. (2019). Gender differences in mathematics and science competitions: A systematic review. *Journal of Research in Science Teaching*, 56(10), 1431–1460. <https://doi.org/10.1002/tea.21580>
- Stern, E. (2017). Individual differences in the learning potential of human beings. *npj Science of Learning*, 2(1), Article 2. <https://doi.org/10.1038/s41539-016-0003-0>
- 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>
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics*. Pearson.
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tobin, R. G., Lacy, S. J., & Crissman, S. (2023). Does the dog in the car have kinetic energy? A multiage case study in the challenges of conceptual change. *Physical Review Physics Education Research*, 19(1), Article 010133. <https://doi.org/10.1103/PhysRevPhysEducRes.19.010133>
- Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard University Press.
- Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, 26(2), 265–283. <https://doi.org/10.1007/s10648-013-9243-1>
- van den Hurk, A., Meelissen, M., & van Langen, A. (2019). Interventions in education to prevent STEM pipeline leakage. *International Journal of Science Education*, 41(2), 150–164. <https://doi.org/10.1080/09500693.2018.1540897>
- Vosniadou, S. (2019). The development of students' understanding of science. *Frontiers in Education*, 4, Article 32. <https://doi.org/10.3389/educ.2019.00032>
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24(4), 535–585. [https://doi.org/10.1016/0010-0285\(92\)90018-W](https://doi.org/10.1016/0010-0285(92)90018-W)
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), Article 1686. <https://doi.org/10.21105/joss.01686>

(Appendix follows)

## Appendix

### Impact of the Order of Instruction of the Basic Curriculum Units on Achievement on the Advanced Curriculum Unit

To examine whether the order in which the basic curriculum units were instructed affected learners' achievement on the advanced curriculum unit, we implemented a multilevel regression model similar to that used for RQ2 (ANCOVA approach). As shown in Table A1, posttest achievement on the hydrostatic pressure and buoyancy force unit, controlling for pretest achievement on the same unit, was not affected by the order in which the basic curriculum units were instructed (i.e., which unit was instructed first). In comparison to students who received instruction on the topic of air and atmospheric pressure first, those who received instruction on the sound and the spreading of sound or the stability of bridges first descriptively had higher scores on the posttest, although with credible intervals including 0. One explanation for such an effect might be that when the basic curriculum units with lower relevance for the advanced curriculum unit are introduced first, the more relevant units will be introduced later and thus in closer proximity to the advanced curriculum unit.

### Impact of Time Gap on Achievement on the Advanced Curriculum Unit

To examine whether the average time gap between the basic curriculum units and the advanced curriculum unit affected learning transfer, we implemented a multilevel regression model similar to that used for RQ2 (ANCOVA approach). As shown in Table A2, posttest achievement on the advanced hydrostatic pressure and buoyancy force curriculum unit, controlling for pretest achievement on the same unit, appeared lower for students with a longer time gap between the basic units and the advanced curriculum unit. The credible interval of the effect of the time gap included 0, so a lack of effect of the temporal distance should not be ruled out, although the effect estimate was clearly in the expected direction (i.e., negative, indicating that a longer time gap diminished learning transfer).

### Gender Differences in Learning Transfer From the Basic to Advanced Curriculum Units

We present details of the analyses concerning gender differences in learning gains (extending the model from RQ1) as well as in achievement controlling for pretest differences across conditions (extending the model from RQ2). Before adding the effects of gender to the respective models, we first descriptively compared the mean achievement of males and females at pre- and posttest. Gender differences in the two conditions at pre- and posttest are depicted in Figure A1. At pretest, boys had an advantage of  $d = 0.14$  in the intervention group and  $d = 0.20$  in the control group. At posttest, boys had an advantage of  $d = 0.10$  in the intervention group and  $d = 0.10$  in the control group. These descriptive analyses indicate a minor advantage for boys in both conditions that remained mostly stable in both conditions, although with a slight decrease from pre- to posttest.

Next, we compared the effect of the intervention on students' learning gains between genders. With this approach, we wanted to determine whether learning with the basic curriculum units mitigates gender differences, leaves them unaffected, or even favors students of either gender in the sense of a rich-get-richer effect. To statistically examine this

**Table A1**

*Results From Bayesian Multilevel Model Regressing Hydrostatic Pressure and Buoyancy Force Posttest on Pretest and on the First Basic Curriculum Unit Instructed*

Parameter	Estimate	Error	90% CI
Intercept	11.99	.72	[9.81, 11.93]
Pretest	0.44	.05	[0.36, 0.53]
Floating and sinking first	0.02	.83	[-1.38, 1.35]
Sound and the spreading of sound first	0.33	.71	[-0.85, 1.47]
Stability of bridges first	1.05	.70	[-0.14, 2.17]

*Note.* The intercept indicates the estimated posttest score for students who received the air and atmospheric pressure curriculum unit first and had zero points of pretest knowledge. Error indicates standard deviation of the estimate (comparable to standard error). Random effects capturing the multilevel structure are available from the additional online materials which are available at the OSF page (<https://osf.io/94rxq/>). CI = credible interval; OSF = Open Science Framework.

question, we extended the repeated-measures ANOVA model from RQ1. We added main effects as well as two-way and three-way interaction terms of gender to this model. The three-way interaction term between time (pre- vs. posttest), condition (control- vs. intervention group) and gender (female vs. male) indicates whether learning with the basic curriculum unit affected genders similarly. As shown by the results in Table A3, the three-way interaction had a negative estimate, indicating a less strong intervention effect for males than for females. The credible interval of this effect did, however, include 0, indicating that an effect of zero cannot be excluded. Thus, being prepared with the basic curriculum units in elementary school did not have a clear effect on gender differences for the advanced curriculum unit, although the point estimate of the interaction indicated a slightly stronger intervention effect for female learners.

Finally, we also extended the model from RQ2 with the effects of gender to examine whether when controlling for pretest differences between conditions, posttest achievement was affected similarly by the condition for male and female learners. The results from this model, extending the Bayesian multilevel ANCOVA model with a main effect of gender and its interaction with condition, are presented in Table A4.

As shown in Table A3, male students tended to show higher posttest achievement than female students in the control condition, and this effect appeared to be lower in the intervention condition.

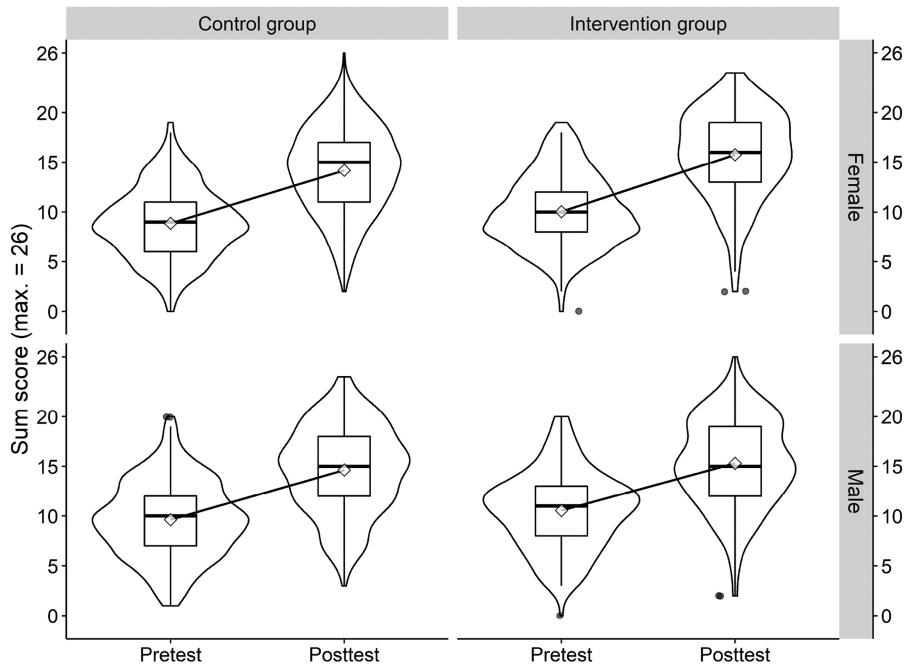
**Table A2**

*Results From Bayesian Multilevel Model Regressing Hydrostatic Pressure and Buoyancy Force Posttest on Pretest and Average Time Gap to Basic Curriculum Units*

Parameter	Estimate	Error	90% CI
Intercept	12.29	.96	[10.73, 13.88]
Pretest	0.43	.05	[0.34, 0.52]
Time gap	-0.31	.24	[-0.69, 0.08]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). Random effects are not reported. CI = credible interval.

**Figure A1**  
Distributions in the Control and Intervention Groups for the Hydrostatic Pressure and Buoyancy Force Pre- and Posttest Across Genders



*Note.* Violin shapes indicate densities of score distributions overlaid with boxplots. Points above and below distributions indicate outliers. Squared points represent mean values, and lines connecting squared points visualize changes between pre- and posttests. max. = maximum.

**Table A3**  
Results From Bayesian Repeated-Measures Multilevel Model Regression of Hydrostatic Pressure and Buoyancy Force Test Scores on Time, Condition, Gender (Female = 0, Male = 1), and Their Interactions

Parameter	Estimate	Error	90% CI
Intercept	8.77	.25	[8.36, 9.17]
Time	5.07	.39	[4.42, 5.71]
Condition	0.88	.38	[0.23, 1.50]
Gender	0.79	.26	[0.36, 1.22]
Time × Condition	0.37	.52	[-0.49, 1.21]
Time × Gender	-0.15	.42	[-0.82, 0.54]
Condition × Gender	-0.15	.48	[-0.94, 0.64]
Time × Condition × Gender	-0.23	.66	[-1.34, 0.89]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). Intercept represents the estimate in female learners in the control group at pretest. Random effects capturing the multilevel structure are available from the additional online materials which are available at the OSF page (<https://osf.io/94rxq/>). CI = credible interval; OSF = Open Science Framework.

Similar to the repeated-measures model presented in Table A2, the interaction effect included 0. This result confirms that the effect of condition was similar for male and female learners, although the

**Table A4**  
Results From Bayesian Multilevel Model Regressing Hydrostatic Pressure and Buoyancy Force Posttest on Pretest, Condition, and Gender

Parameter	Estimate	Error	90% CI
Intercept	10.00	.42	[9.30, 10.66]
Pretest	0.43	.05	[0.38, 0.48]
Condition	1.02	.42	[0.31, 1.68]
Gender	0.26	.33	[-0.27, 0.82]
Condition × Gender	-0.33	.51	[-1.13, 0.54]

*Note.* Error indicates standard deviation of the estimate (comparable to standard error). Intercept represents female students in the control condition. Condition coded as *control group* = 0, *intervention group* = 1. Gender coded as *female* = 0, *male* = 1. Random effects capturing the multilevel structure are available from the additional online materials which are available at the OSF page (<https://osf.io/94rxq/>). CI = credible interval; OSF = Open Science Framework.

point estimate of the interaction indicated a potentially stronger intervention effect for females.

Received January 20, 2023  
Revision received March 15, 2024  
Accepted April 25, 2024 ■