frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 February 2026
DOI 10.3389/feduc.2026.1741733

OPEN ACCESS

EDITED BY
I Kadek Suartama,
Ganesha University of Education, Indonesia

REVIEWED BY
Teguh Arie Sandy,
Ahli Media Consultant, Indonesia
I Gde Wawan Sudatha,
Ganesha University of Education, Indonesia

*CORRESPONDENCE
Patrick Wiesner
patrick.wiesner@mathematik.uni-regensburg.de

RECEIVED 07 November 2025
REVISED 29 December 2025
ACCEPTED 05 January 2026
PUBLISHED 19 February 2026

CITATION
Wiesner P, Krauss S, Stegmüller N and Binder K (2026) Is flipped classroom really superior?—Questioning the flip in K-12 teaching.
Front. Educ. 11:1741733.
doi: 10.3389/feduc.2026.1741733

Is flipped classroom really superior?—Questioning the flip in K-12 teaching

Patrick Wiesner[1]*, Stefan Krauss[1], Nathalie Stegmüller[1] and Karin Binder[2]

[1]Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany, [2]Mathematics Education, Faculty of Mathematics, University of Paderborn, Paderborn, Germany

There is a contradiction between seven meta-analyses, all of which indicate a substantial benefit of the flipped classroom (FC) method for K-12 teaching and some larger study that found no such benefit when compared to "traditional" teaching. In the theoretical part of the paper, we shed light on this contradiction by consulting general literature on meta-analyses. Ranking the 50 included FC studies by the number of classes per experimental condition, we found a negative correlation between the "size" of a study and the effect in favor of FC. In the empirical part, we present an FC study with three conditions concerning mathematical teaching, based on n = 950 students aged 11–13, in which many relevant covariates (e.g., quality of instruction) were addressed. One FC condition was based on students' knowledge acquisition through instructional videos at home (FCn: n = 12 classes). Considering that self-regulation support might play a crucial role especially for young students working at home, another FC condition (FCS: n = 12 classes) was implemented, in which students could learn additional math-free strategies concerning watching instructional videos. Both FC-conditions were experimentally compared with a control group of traditional teaching (TT: n = 13 classes). No significant effect on learning gains was found between FCn and TT, indicating that "flipping" alone may not be more effective per se. However, a significant difference was found between FCS and FCn. Thus, supporting students' self-regulation in addition may in indeed open the door to successful FC, even with very young students.

KEYWORDS

flipped classroom, inverted learning, K-12, instructional video, self-regulated learning, math education, flipped learning, FALKE research program

1 Introduction

Several meta-analyses document positive effects of the flipped classroom (FC) method in K-12 education (Cheng et al., 2019; Güler et al., 2022; Låg and Sæle, 2019; Strelan et al., 2020; van Alten et al., 2019; Wagner et al., 2020; Zhu, 2021). According to Table 1, the advantage of FC seems obvious from an empirical perspective. However, when attempting to demonstrate this effect in controlled "large-scale" designs while implementing multiple classes per experimental condition, the evidence astonishingly fails (Wagner and Urhahne, 2021). In the present article, we review the existing meta-analyses regarding K-12 teaching (i.e., primary and secondary education) and discuss possible alternative reasons for the repeated confirmation of positive effects in favor of

TABLE 1 Overview of meta-analyses comparing FC with traditional teaching considering learning gains in K-12.

Meta-analysis (year) Journal	K-12 level (# FC studies included)	Effect sizes in favor of FC (Cohen's *d* / Hedges *g*)*
Doğan et al. (2021)** Education & Information Technologies	Secondary level (11)	$d = 0.58$
	Primary level (4)	$d = 1.89$
Zhu (2021) Educational Technology Research and Development	Secondary & primary level together (27)	$d = 0.54$
Wagner et al. (2020) Zeitschrift für Pädagogische Psychologie	Secondary level (25)	$d = 0.42$
Strelan et al. (2020)** Educational Research Review	Secondary level (21)	$g = 0.64$
	Primary level (3)	$g = 0.47$
Låg and Sæle (2019)** AERA Open	Secondary level (16)	$g = 0.45$
	Primary level (12)	$g = 0.44$
van Alten et al. (2019)** Educational Research Review	Secondary level (11)	$g = 0.36$
Cheng et al. (2019)** Education Tech Research	Secondary level (12)	$g = 0.21$

*The effect sizes *d* and *g* only vary by a slightly different calculation of the pooled standard deviation: $|g \text{ and } d| = 0.2$ small effect, $|g \text{ and } d| = 0.5$ medium effect, $|g \text{ and } d| = 0.8$ large effect.
**Meta-analysis also includes studies regarding tertiary level (which are not listed in this table).

the FC method. Finally, we present the so far largest controlled FC study[1] in K-12 teaching, in which we implement about 12 classes per experimental condition and control multiple possible covariates that might alternatively affect learning gains. Since self-regulation support for watching instructional videos at home seems to be crucial especially for young students (see below), an additional FC condition was implemented in which such strategies—addressing "before," "while" and "after" watching the video—were explicitly provided.

1.1 Flipped classroom

The flipped classroom (FC; synonyms: inverted classroom, flipped learning, inverted learning) is a teaching method that has become increasingly popular in recent years with a growing focus on research that has all been conducted within the last 15 years (Cevikbas and Kaiser, 2020; Hwang et al., 2019; Kapur et al., 2022). Zhang et al. (2024) for example provide an overview of current research trends on FC across all age groups and professions. Although there is no single definition of FC, almost all characterizations include the reversed order of traditional teaching as a "minimum requirement." In contrast to traditional teaching—i.e., the knowledge acquisition happens in school and most of the practice has to be done as homework—the concept of FC means that new knowledge is acquired independently by students before the actual lesson to allow more time for reflection, elaboration, and practice in class (Bishop and Verleger, 2013; Lage et al., 2000). In terms of acquiring knowledge at home, students are often required to watch instructional videos and are sometimes given short additional quizzes to test their understanding of the content (e.g., Hew et al., 2021; Wagner et al., 2020). Some authors specify options for the subsequent plenary phase, for example group work, class surveys, cooperative learning, and student presentations or discussions (Bergmann and Sams, 2012; Bishop and Verleger, 2013; Lo et al., 2017). Meanwhile, there are a lot of individual studies investigating the effect of FC in K-12 teaching (see Supplementary Table 1). Taken together, these studies provide important insights into ideas, conditions, and consequences on teaching with the FC method.

1.2 The effect of flipped classroom in K-12 education: meta-analytical evidence

Like any method, FC has potential opportunities as well as potential pitfalls (Cevikbas and Kaiser, 2023). We summarize seven meta-analyses investigating the effect of FC regarding K-12 teaching that all were conducted within the last 7 years (in Table 1). We selected analyses that examined the effects in the K-12 sector separately and explicitly listed the included studies. All seven meta-analyses generally conclude that FC is more effective than traditional teaching in K-12 with effect sizes ranging from $g = 0.21$ to $d = 1.89$. The effect of these meta-analyses is in line with a recent review of meta-analyses from Li et al. (2024) with an overall positive effect of $g = 0.53$ in favor of FC in K-12 teaching. There are even two more reviews on meta-analyses (Hew et al., 2021; Kapur et al., 2022), yet, they deal predominantly with the tertiary level (for the restriction to K-12 teaching in the present article see 1.3).

A common finding of the meta-analyses is that most of the included studies compare FC based on instructional videos that have to be watched at home by students with traditional teaching without videos (e.g., see Zhu, 2021). Furthermore, the positive effects of FC are often attributed to short quizzes (e.g., see van Alten et al., 2019), at least when students receive feedback (Wagner et al., 2020).

However, the evidence is mixed in terms of the subjects considered. While Wagner et al. (2020) reported a larger effect in STEM subjects than in the humanities, Strelan et al. (2020) found the opposite.

In Supplementary Table 1, we provide an extensive overview of the about 50 individual studies comparing learning gains from FC to traditional teaching that were included in at least one of the seven meta-analyses in Table 1. The first letters of the names (bold) in Table 1 are used in the left column in Supplementary Table 1 to indicate which study was included in which meta-analysis. Since most meta-analyses critically remark the generally small sizes of the FC studies conducted so far (e.g., Cheng et al., 2019; Wagner et al., 2020), we sorted the studies in Supplementary Table 1, first, according to the average number of classes per experimental condition and, second, by the total number of participating students. As statistical multilevel modeling, which diminishes potential class effects, is only possible if more classes per condition are implemented, we chose the number of

1 The study was part of FALKE-d (Frei et al., 2020).

classes per condition as the core principle to judge the "largeness" of an FC study. Notably, in Supplementary Table 1 the bivariate Pearson correlation coefficient between the number of classes per condition and the effect sizes reported in Supplementary Table 1 yields $r = -0.3$ i.e. the larger a study, the smaller the observed effect in favor of FC.

More specifically, from Supplementary Table 1, which contains 51 individual FC studies, it becomes clear that 40 studies implemented only one class per experimental condition. So far, there seems to be only one 'large' FC study concerning K-12 teaching (Wagner and Urhahne, 2021) with an average of 10 classes per condition, which was too new to be included in the meta-analyses of Table 1. In this study, however, the most effective method was not FC, but watching the instructional video in class followed by student-centered instruction. Since the learning materials were basically identical in all conditions, the study by Wagner and Urhahne (2021) has a high internal validity, but at the same time, there is a trade-off regarding external validity because instructional videos are typically not shown in class. Thus, there currently seems to be a major contradiction between the results of seven meta-analyses and an actual, large and internally valid FC study.

In experimental psychology, it is not uncommon for effects found in small studies to fail to be replicated in large and controlled experiments. Beyond the well-known problems demonstrated in the replication crisis (e.g., Maxwell et al., 2015; Shrout and Rodgers, 2018), Bartoš et al. (2023), Kvarven et al. (2020), and Sotola (2022) examined meta-analyses in particular. The latter one investigated meta-analyses that were published in *Psychology Bulletin* and demonstrated that the risk of a meta-analysis confirming a positive effect when, in fact, there is a null effect is particularly high when the number of participants in the included studies is very small. He claimed in such cases that for every included study, there must be another study that was not published, meaning that these estimated unpublished 50 percent of all conducted studies might show no or a negative effect. While Kvarven et al. (2020) assume that effect sizes in meta-analyses are overestimated almost by a factor of three, Bartoš et al. (2023) stress that such effect sizes might be overestimated even more. The general conclusion of all three reviews is that meta-analyses, of course, are not unreliable per se, but should probably be interpreted with caution, and all agree that the reported effect sizes might be overestimated.

In the absence of contrary arguments or specialties that discriminate the meta-analyses of Table 1 from those reviewed by Bartoš et al. (2023), Kvarven et al. (2020) and Sotola (2022), one might speculate that the theoretical analyses of the three papers may also—at least partly—underlie the seemingly overwhelming evidence in favor of FC instruction. This hypothesis is consistent with Kapur et al.'s (2022) review of meta-analyses exclusively on FC, which also suggests that effect sizes might be overestimated due to small sample sizes of the included studies and publication bias (see also Hew et al., 2021). Yet, neither Kapur et al. (2022) nor Hew et al. (2021) focus on K-12 students (instead 90% of the studies considered by Kapur et al. were from the tertiary level). Of course, all authors of the meta-analyses in Table 1 are aware that their analyses depend on the individual studies included (e.g., Strelan et al., 2020; Låg and Sæle, 2019), and discuss a possible publication bias (Zhu, 2021), small study bias (Doğan et al., 2021), or the novelty effect, meaning that there is an initial boost in performance or enthusiasm that occurs when a new tool or method is introduced (Wagner et al., 2020).

1.3 The role of self-regulation for K-12 students

To provide the desired benefit of extended learning time for FC, students must prepare effectively for the lesson at home, since, otherwise, the success of the FC method is threatened (Cheng et al., 2019; Gillette et al., 2018; van Alten et al., 2019). The issue of self-regulated learning ("SRL") and the need to support self-regulation, especially for young students, is the reason why we focus on K-12 teaching in this paper. While studying at university requires strong self-regulation skills, students at school usually receive more help and explicit support (Vosniadou, 2020). In the K-12 classroom, teachers implementing the FC method cannot automatically assume that learning at home, especially for very young students, will be competent and successful (or be done at all). Indeed, the review of meta-analyses by Kapur et al. (2022) indirectly supports the claim that the relevance of SRL support decreases as students get older: At the tertiary level, the average effect size for FC (without specific SRL support and compared to traditional teaching) is the highest ($g = 0.93$), while the younger the students, the lower the effect size (high schools: $g = 0.63$, elementary schools: $g = 0.40$).

However, in experimental FC studies, elaborated support regarding the appropriate use of instructional videos is found only rarely. Table 2 summarizes three studies that focus specifically on SRL in K-12 teaching by experimentally comparing two FC conditions with and without explicit SRL support (all three without implementing an additional control group regarding traditional teaching).

TABLE 2 Overview of studies comparing FC with and without SRL support in K-12 education.

Authors (year) Journal	# Students, (# classes)	Subject (grade)	Conditions: # students, (# classes)		Results (relating to learning gains)
			FC with SRL support	FC without SRL support	
van Alten et al. (2020a) Computers in Human Behavior	154, (6)	History (8th grade)	74, **(6)**	80, **(6)**	No significant effect
van Alten et al. (2020b) Computers & Education	115, (5)	History (8th grade)	50, **(2°)**	65, **(3°)**	Significant positive effect in favor of FC with SRL
Lai and Hwang (2016) Computers & Education	44, (2)	Mathematics (4th grade)	20, **(1)**	24, **(1)**	Significant positive effect in favor of FC with SRL

°Number of classes estimated by ourselves. Numbers in bold indicate the number of classes per condition.

The SRL support in the studies in Table 2 consists, for example, of short prompts that are implemented in the videos. However, general research on SRL over the last few decades has shown that it is not easy to guide students to actually use SRL strategies by simply transferring knowledge, but rather practical exercises are needed in addition (Dignath and Büttner, 2008; Zeidner and Stoeger, 2019). In principle, it would be possible to address SRL even more dedicatedly—based on general research on SRL (e.g., see Pintrich, 1999)—and to develop strategies explicitly designed for watching instructional videos, for instance, by following the ICAP framework (Chi and Wylie, 2014), which especially deals with digital media.

1.4 Rationale for the present study

In sum, despite many existing studies on the FC method in K-12 (Supplementary Table 1), it is still not certain whether this method is generally more effective than traditional teaching. Wagner et al. (2020, p. 14) state that "in particular, more randomized controlled trials with larger sample sizes and objective quantitative measures are needed." With respect to experimental control, for instance, Låg and Sæle (2019) and Cheng et al. (2019) mention that most of the existing FC studies do not apply precise outcome measures and are therefore "noisy" (Strelan et al., 2020). For example, often researchers themselves were the teachers in the studies, which violates the double-blind principle (Låg and Sæle, 2019), or there was no pretest to control for differences (Cheng et al., 2019).

In the present paper, we report a controlled FC study regarding mathematics teaching in grades 6 and 7. With the implementation of $n = 12$ classes per condition (a total of $N = 950$ students were examined), this is the largest study in this area so far. Specifically, (a) a control group of "traditional teaching" was investigated and, therefore, we contribute to Supplementary Table 1. In addition, we (b) implemented two FC-conditions, one with and one without explicit SRL-support for watching the instructional videos at home. In doing so, we contribute to Table 2. Furthermore, we tried to manage all the requirements mentioned in Låg and Sæle (2019) such as clearly reporting the study design with a detailed description of the intervention, double-blind assessments, and psychometrically sound outcome measures.

2 Research questions

The main objectives of the present study are to investigate the effectiveness of an FC setting based on instructional videos in a robust, ecologically valid, and well-documented design when teaching mathematics to 11–13 year old students. Note that the following research questions are pedagogical in nature; the subject of mathematics serves rather as an exemplary discipline.

Research question 1 (RQ1): Do the mathematical learning gains of (young) students taught in an FC setting based on instructional videos differ from those taught in a traditional setting?

Research question 2 (RQ2): What is the effect of explicit support of SRL for watching instructional videos at home in an FC setting?

Since Wagner and Urhahne (2021) could not replicate the positive effects in favor of FC with a large sample, we do not have a clear hypothesis on RQ1. Regarding RQ2, we speculate that SRL support might be the crucial factor for the effectiveness of FC, especially for students approximately of age 12 years.

3 Method

3.1 Study design

The intervention study with a pre-post design was conducted in 6th and 7th grade mathematics classes in Bavarian schools (Figure 1). The participating classes were randomly distributed into three groups. The students were taught by their respective math teachers who received standardized training and materials prior to the intervention (see 3.3.4). Within each group, the teachers were following the identical manual (see 3.2). This uniform design of the lessons is intended to eliminate differences due to varying materials and to create comparability within, but also between experimental conditions. Basically, only the order of the used materials changed between the groups; therefore, all students received the same tasks and explanations except for the instructional videos (see 3.3.1) that were only obligatory in the FC conditions (Figure 1). Students' performance was measured before and after the intervention using a pretest and a posttest (see 3.3.5). The lessons are described in more detail in 3.2.

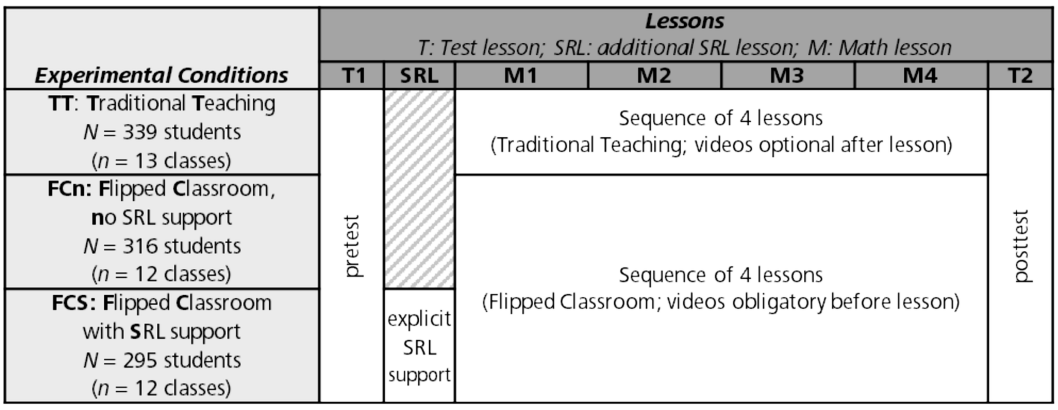

FIGURE 1
Overview of the study design.

To answer RQ1, we compare classes in the traditional teaching (TT) condition with classes in the (typical) FC setting, in which students are provided with instructional videos but receive no explicit SRL support for watching them at home (FCn). A second FC condition (FCS) was implemented to answer RQ2: this group received an additional lesson on strategies (not mathematics-specific) about how to use instructional videos (see column "SRL" in Figure 1). It is important to note that no mathematical content was conveyed during this lesson to guarantee that no group had an advantage in this regard (3.2.2).

3.2 Intervention

3.2.1 Four-lesson sequence in mathematics: M1–M4

3.2.1.1 Structure of each lesson

Since the lessons of both FC groups were identical except for the additional SRL lesson (Figure 1), we will report them together in the following (Table 3, right). Basically, also the TT group and the two FC conditions shared the same mathematical components (for an overview of the content of M1 – M4 see Figure 2).

In the traditional setting (left in Table 3), after the revision of homework—not in the first lesson—teachers were of course instructed to follow the explanations that were also given in the videos from the FC conditions. The new content then had to be consolidated with worksheets, and teachers provided a short quiz to check their students' understanding. The remaining time was spent doing exercises (part A). Finally, students were assigned homework (exercises, part B) to follow up on the lesson. Students were informed that there were also instructional videos that could be voluntarily $n = 12$ classes watched after each lesson to review the topic.

In both FC settings, as a preparation for each lesson, the students received an instructional video, followed by a quiz (the same as in TT) with automatic feedback on a Moodle-based learning platform. At the beginning of the class, the quiz questions were reviewed, giving the teacher the opportunity to specifically address poorly-answered questions. Students were also encouraged to ask questions related to the video. The content of the video was then consolidated using the same worksheets as the TT group. Finally, the teacher reminded the students of their upcoming homework, namely, to watch the next instructional video and to answer the quiz questions.

3.2.1.2 Content of the four mathematics lessons

The sequence of the four lessons dealt with *numerical formats of relative frequencies* and was identical with respect to the content covered for all groups (Figure 2). Students usually learn at school that relative frequencies can be written as *common fractions* (e.g., 1/4), *decimal fractions* (e.g., 0.25), or *percentages* (e.g., 25%). However, in media and everyday language one often finds alternative formats such as *natural frequencies* (e.g., "1 out of 4"), the notation *every* n^{th} (e.g., "every fourth"), or *odds* (e.g., "1 to 3"). Since common and decimal fractions and percentages are standard in teaching mathematics, we describe below the three other formats that were implemented in our study.

Natural frequencies consist of two absolute frequencies a and b (i.e., "a out of b" with $a, b \in N$, $a \leq b$ and $b \neq 0$). In applied research, they are particularly famous because of their beneficial effect for understanding conditional probabilities (in so-called Bayesian situations; Gigerenzer and Hoffrage, 1995; McDowell and Jacobs, 2017). In the notation *every* n^{th}, e.g., 1/4 becomes "every fourth," and in general 1/n becomes "every n^{th}." *Odds* describe the 'confrontation' of the number of cases "a" with a particular characteristic to the number of cases "b" without that characteristic. Thus, in terms of odds, the probability of 50% can be expressed as "1: 1" (read: "1 to 1") and the probability of 25% can be expressed as "1: 3" (read: "1 to 3"). Note that the odds of 1 to 3, therefore, correspond to the probability of 1/4.

Since empirical studies have documented students' problems especially with the mutual conversions of such formats (Wiesner et al., 2023), the first lesson began with an introduction to *natural frequencies* (M1) and their equivalence to *common fractions*. This was followed by a repetition of the known conversions between *common fractions*, *decimal fractions,* and *percentages* (M2). Finally, the notations *every* n^{th} (M3) and *odds* (M4) were introduced, each time related to *natural frequencies*, which is the most intuitive format (Binder et al., 2015; Gigerenzer and Hoffrage, 1995).

3.2.2 Additional lesson for SRL (only given to the FCS group)

After the pretest (see Figure 1) and before the mathematics sequence, an additional lesson ("SRL") was taught exclusively in the FCS condition. The goal of this lesson was to help students work successfully with the instructional video at home by using specific strategies. This lesson was designed "math-free" to ensure that the students did not have an advantage in the following mathematics

TABLE 3 Typical structure of a lesson in TT and in both FC settings.

Location	Traditional Teaching	Time	Location	Flipped Classroom	Time
School	- Revision of homework - Explanation of new content by teacher (following the structure of the instructional video)	≈ 5 min ≈ 10 min	Home	- Instructional video - Quiz (the same as in TT)	≈ 15 min
	- Consolidation (worksheet)	≈ 5 min	School	- Revision of the quiz & opportunity for questions	≈ 5 min
	- Quiz (same as in FC)	≈ 5 min			
	- Exercises (part A)	≈ 20 min		- Consolidation (worksheet)	≈ 5 min
Home	- Homework exercises (part B) - Opportunity to watch the instructional video	≈ 15 min		- Exercises (part A and B)	≈ 35 min

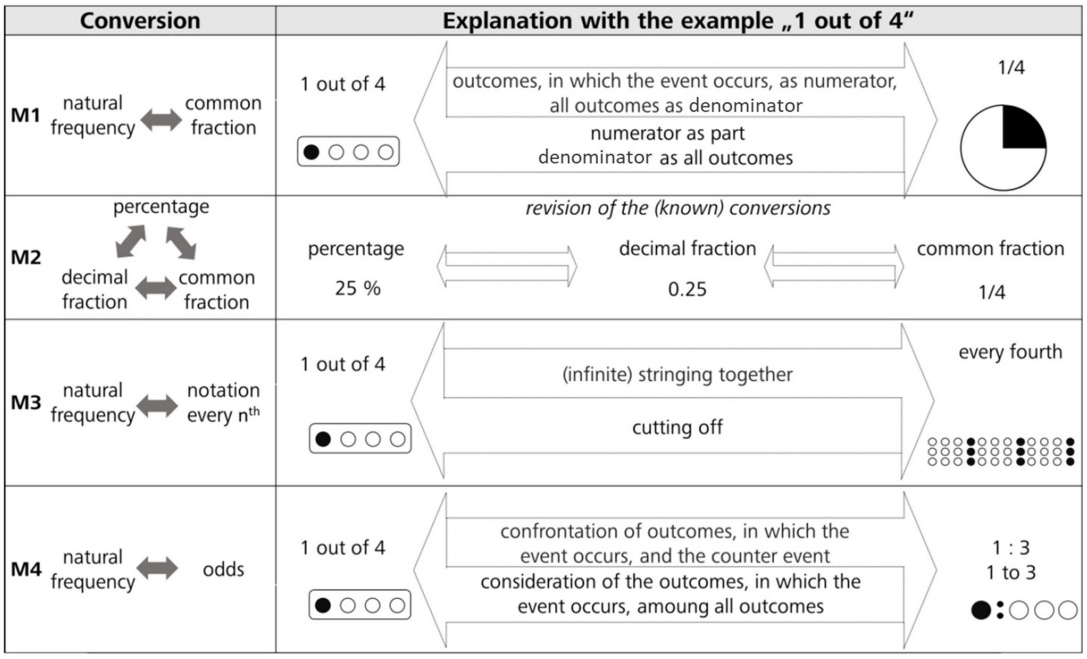

FIGURE 2
Overview of the content of the four lessons M1 to M4.

sequence. Following the SRL model by Pintrich (1999), four cognitive strategies were developed, specifically focusing on instructional videos:

- *repetition strategy* (ensuring awareness before watching): watching the video several times, the first viewing serves as an overview; from the second viewing onwards the following strategies are added
- *before and after strategy* (before *and* after watching): noting down one's own knowledge before and after watching the video
- *note taking strategy* (while watching): stopping the video at important points (or statements that are not understood) to take notes in bullet points (or just to note a single question mark and the timestamp of the issue)
- *ordering strategy* (after watching): sorting notes from the note-taking tratégy to structure the content of the video, adding more notes, or removing redundant ones

All strategies are in line with the ICAP framework on digital learning (Interactive, Constructive, Active, and Passive), which describes in detail how students' learning performance increases from passive to interactive as they become more engaged with digital learning tools such as instructional videos (Chi and Wylie, 2014). The strategy lesson itself was structured as follows: First, students were presented the "monkey business illusion" video by Simons (2010) in order to illustrate that absorbing all the information by watching a video only once can be difficult and to motivate the need for strategies, especially when watching instructional videos. For the development of the strategies in class, the teachers of the FCS group followed the respective manual.

Afterwards, the strategies were consolidated with a so-called "Strategy Prism" (see 3.3.3), which is supposed to help the students remember the strategies and support metacognition. Finally, and most importantly, according to pertinent SRL research (Dignath and Büttner, 2008; Stoeger et al., 2015), the strategies were put into practice during a class exercise on a non-subject-specific video about democracy with the teachers' guidance. In the FCS group, before each of the four instructional videos, students were reminded about the strategies by a one-minute video in which the Strategy Prism was recalled.

3.3 Materials and administration

All teachers received a standardized teacher training in which the content and the standardized implementation of the lessons were explained by the test administrator (the first author). They received background information on the administration of the study and teaching material (e.g., manuals for each lesson, worksheets, quizzes, solutions, etc.). The instructional videos and the quizzes were provided online.

3.3.1 Instructional videos

The instructional videos were created using the lightboard technique (Lubrick et al., 2019), because this technique allows one to implement design recommendations from various authors, such as the Gaze Guidance Principle or the Dynamic Drawing Principle (Brame, 2016; Mayer et al., 2020). For each video, the same structure was used: first, the new type of numerical format to be learned was introduced by an example. Then, the link to an already known notation was shown by using and connecting their 'basic ideas' (*Grundvorstellungen*; see pictorials in Figure 2). Finally, the new format was placed into an overall graphic overview, in which all numerical formats were integrated and that was completed step by step in the four-lesson sequence.

The length of an instructional video ranged from 6 to 9 min. The videos were, in principle, online available throughout the course to all students in the three groups, yet obligatory as preparation only in the FC conditions. Furthermore, to control whether students had actually watched the videos before the lesson (manipulation check), the teachers and researchers were able to verify this via tracking data. This showed that the classes in the FC conditions actually did the flip and that the students in these two conditions used the videos significantly more than the students in the TT conditions.

3.3.2 Quizzes

Directly after watching the videos at home, the students of the FC groups were asked to complete a short quiz of four tasks, which was also done in class in TT. The questions were either closed (e.g., Which answers are correct? "1 out of 2" is "1/2," "50%," "0.2," or "20%") or semi-open (e.g., Reduce the natural frequency as much as possible: "12 out of 24" is "___ out of ___"). For the two FC groups, the quizzes were also available online. As with the instructional videos, teachers and researchers could check to see if students had completed the quizzes.

3.3.3 The strategy prism (only given to FCS group)

The Strategy Prism, which was distributed to by teachers in the FCS condition, serves as a reminder that summarizes all four strategies in a nutshell to support metacognitive skills. It was designed as a flyer that can be folded into a prism, thereby creating three surface areas. The three faces represent the three time periods "before watching," "while watching," and "after watching" and displayed relevant aspects briefly in bullet points. The prism can be folded for easy transportation and quickly set up when needed. Therefore, whenever students want to watch a video, they can easily build the prism and turn the relevant face towards them according to the phase that they are in.

3.3.4 Standardized guidelines for teachers

The mathematical content of the training was identical for teachers in the three conditions. Teachers in the FC conditions received additional information on how to implement the FC method, and teachers in the FCS group also received a plan for the additional strategy lesson.

Table 4 summarizes an evaluation of the guidelines and teaching materials by the participating teachers after completion of the study. Especially the result of the last item demonstrates a high level of experimental control.

3.3.5 Pre- and posttest for measuring the learning gains

The pre- and posttest each consisted of 17 items; both tests shared the same structure, meaning that the items were formulated identically with only the numerical values differing between pre- and posttest. The first four (closed) items were similar to the following example: "Every fifth means 20%" a) "true," b) "false." The next eight (semi-opened) items required active conversions between two formats (e.g., "1/6" is "__ out of 18"). The following three items required students to decide whether a given inequality is true or false. In the remaining two items, students had to identify the largest and the smallest out of four provided ones in different formats. Each item of the pre- and posttest yielded one point and the scales of both tests were formed by adding the corresponding 17 scores.

TABLE 4 Teachers' evaluation of the content, training, and materials of the mathematics sequence.

Item	Mean (SD)
The content of the mathematics lessons was relevant for my students.	3.75 (0.55)
The explications by the test administrator were clear and helpful.	3.79 (0.54)
The teaching materials were clearly structured.	3.75 (0.44)
During the study, I knew exactly what to do at any time.	3.95 (0.22)

N = 34; Likert scale: 1 = no agreement; 2 = rather no agreement; 3 = rather agreement; 4 = total agreement.

The implemented videos, quizzes, all teaching materials for the four math lessons as well as the pre- and posttest can be found via the this link.

3.3.6 Control variables

To be able to statistically control the influence of potentially relevant covariates, the following control variables were additionally collected in the pre- and posttest lessons. For students: *gender*, *native language*, *own computer or tablet*, *instructional quality* (*global*), *instructional quality (of the four mathematics lessons), previous use of digital media in spare time/ for school, reading habits*, *previous use of instructional videos at home*, *strategies for using instructional videos before intervention study* and *strategies for using instructional videos during intervention study*; for teachers: *gender*, *work experience* (i.e., *years of being a teacher*), *instructional quality* (*global), instructional quality (concerning the four mathematics lessons*), *positive attitude towards teaching (in general), attitude towards digital media*, *attitude towards instructional videos*, and *previous use of instructional videos*. In Supplementary Table 2, an example item is presented and the corresponding scales are summarized with indication of internal consistency for each scale.

3.4 Participants

Data were gathered from $N = 950$ students of grade 6 and 7 of all school types across various locations in the state of Bavaria (Germany) in the school year 2022/23. In Germany basically there are an academic track ("Gymnasium"), a medium track ("Realschule") and a vocational track ("Mittelschule"). The classes were chosen so that all students had the same prior knowledge: All classes knew fractions (common and decimal) and percentages, but none of them had explicit experience with the other three formats before. The students were taught in $n = 37$ classes; these classes were distributed among $n = 20$ schools. Three teachers taught two different classes (two of the TT group and one of the FCn group); thus the students were taught by $n = 34$ teachers. Table 5 shows the distribution of students, classes, and teachers separated by condition. An overview of class size and gender distribution is provided in Supplementary Table 2, Table B.4.

Of all teachers who volunteered to participate in the study, 16 were female and 18 were male, with an average work experience of 14.2 years. In all three groups, the average work experience, gender distribution of teachers and school type were comparable. Prior to data collection, the Bavarian Ministry of Education, school principals, teachers, parents, and the students themselves consented to the committed participation in the study.

3.5 Statistical analysis

3.5.1 Coding of pre- and posttest

Concerning the closed items, the correct choice was scored with one point. In semi-open items (with a gap), all possible correct answers were scored as correct. For example, regarding the item 3b (pretest) concerning the conversion of 30% into a common fraction, several answers (e.g., 3/10 or also 30/100) received one point. For implementation in the statistical regression model (3.5.2), manifest scales were formed for the pre- and the posttest, respectively. In sum, a maximum score of 17 could be achieved on both the pretest and the posttest.

3.5.2 Statistical model

From a statistical point of view, the two research questions RQ1 and RQ2 can be modeled simultaneously by setting the FCn condition as reference group. Due to the longitudinal and nested structure of the data, a linear mixed model (LMM) was used to predict students' performance in the posttest while controlling for students' performance in the pretest (Hilbert et al., 2019). Since the FCn group served as reference group, the factors "TT" (0: no Traditional Teaching; 1: Traditional Teaching) and "FCS" (0: no SRL support; 1: SRL support) were included via dummy coding as well as the measurement point "time" (0: pretest, 1: posttest). In addition, since the effectiveness of the training from pretest to posttest was expected to vary between the different training groups, two interaction terms TT × time and FCS × time were modeled. The model equation was (Table 6 for the meaning of the various β coefficients).

$$y = \beta_{\text{int}} + \beta_{time} \cdot time + \beta_{TT} \cdot TT + \beta_{FCS} \cdot FCS + \beta_{TT\times time} \cdot (TT \times time) + \beta_{FCS\times time} \cdot (FCS \times time) + u_{0i_{school}} + u_{0i_{class}}$$

To account for the multilevel structure in the school context, the nesting of "school" and "class" was added using random effects in the LMM, which allows the individual parameters to be adjusted in order to obtain a more realistic estimate.

TABLE 5 Number of participating students (schools) and classes/teachers separated by experimental condition.

Condition	Students (schools)	Classes/teachers
TT	339 (6)	13/11
FCn	316 (5)	12/11
FCS	295 (9)	12/12

The applied LMM has the following three interpretative advantages compared to a mixed ANOVA (Hilbert et al., 2019): a) If there were significant differences in the pretest score, the model would not only indicate whether they exist, but also where they are. In addition, significant differences between the interactions could be directly located. Both aspects eliminate the need for post-hoc analyses, which would result in a loss of statistical power; b) an LMM can handle missing values; thereby statistical power is increased; c) the LMM has less stringent model requirements (homoscedasticity; i.e., it is not required that the variances of the error terms are equal).

4 Results

4.1 Descriptive results of pre- and posttest

Table 7 provides an overview of the descriptive results from the pre- and posttest. Both scales yielded satisfactory internal consistency aggregated across all groups as indicated by Cronbach's α ($\alpha_{\text{pretest}} = 0.79$; $\alpha_{\text{posttest}} = 0.84$). Across all three groups the average score in the pretest was 8.51 (SD = 2.96) and in the posttest 11.38 (SD = 3.30) with a Cohen's d of 0.95 (The male participants achieved an average score of 8.67 (SD = 3.06) in the pretest and 11.75 (SD = 3.25) in the posttest, corresponding to a Cohen's d of 0.98. For the female participants, the score was 8.34 (SD = 2.83) in the pretest and 10.94 (SD = 3.22) in the posttest, with a Cohen's d of 0.86.).

Figure 3 displays the resulting line diagram regarding students' pre- and post-performances separated by the three experimental conditions.

4.2 Results of the linear mixed model

According to the research questions, in the LMM, the FCn condition was chosen as reference group (Table 8). There was no significant difference in the pretest scores between the TT group and the reference group FCn ($\beta_{TT} = -1.61$, $p = 0.10$). Similarly, the pretest scores of the FCS group did not significantly differ from those of FCn group ($\beta_S = -1.39$, $p = 0.07$). Consequently, the remaining difference cannot be significant (obviously, the pretest scores of TT and FCS are closest to each other, Figure 3). The general comparability of the three groups with respect to the pretest contributes to the internal validity of conclusions from the model, since possible interaction effects are less likely biased by varying a priori competencies of the students.

Note that the intercept of 9.33 (i.e., the pretest score of the reference group FCn) in Table 8 slightly varies from Figure 3 (9.12), which indicates the correction due to the implemented multi-level modeling. The significance only indicates its difference from 0. Moving on to the posttest results, there was a significant improvement of the reference group (FCn) due to the teaching sequence ($\beta_{time} = 2.48, p < 0.001$). In terms of inferential statistics, RQ1 and RQ2 can be answered by comparing the slopes in Figure 3. The difference in the slopes is analyzed by the interactions. Regarding RQ1, the interaction effect between time and TT did not yield significant differences in the improvement ($\beta_{TT\times time} = 0.45$, $p = 0.07$). Regarding RQ2, the interaction effect between time and FCS was significant ($\beta_{FCS\times time} = 0.76$, $p = 0.003$), indicating

TABLE 6 Interpretation of the β coefficients in the LMM.

β coefficient	Interpretation
β_{Int}	Intercept: β_{Int} is the pre-value of the FCn group. Equation: $y_{FCn/pre} = \beta_{Int}$
β_{TT}	Main effect "TT" (Traditional Teaching): β_{TT} indicates the difference attributable solely to the traditional teaching factor (= difference between FCn and TT) at the pretest score. Equation: $y_{TT/pre} = \beta_{Int} + \beta_{TT}$
β_{FCS}	Main effect "FCS" (SRL support): β_{FCS} indicates the difference attributable solely to the 'strategies factor' (= difference between FCn and FCS) at the pretest score. Equation: $y_{FCS/pre} = \beta_{Int} + \beta_{FCS}$
β_{time}	Main effect time: β_{time} indicates difference attributable solely to the factor of time (= difference between post and pre) for the FCn group. Equation: $y_{FCn/post} = \beta_{Int} + \beta_{time}$
$\beta_{TT\times time}$	Interaction effect TT × time (Traditional Teaching × time): $\beta_{TT \times time}$ indicates the difference between FCn and TT concerning learning gains. Equation: $y_{TT/post} = \beta_{Int} + \beta_{time} + \beta_{TT} + \beta_{TT\times time}$
$\beta_{FCS\times time}$	Interaction effect FCS × time (SRL support × time): $\beta_{FCS\times time}$ indicates the difference between FCn and FCS concerning learning gains. Equation: $y_{S/post} = \beta_{Int} + \beta_{time} + \beta_{FCS} + \beta_{FCS\times time}$

TABLE 7 Scales of pretest and posttest: Descriptive results.

Condition	Pretest	Posttest	Effect size
	M (*SD*)	M (*SD*)	Cohens *d*
Traditional Teaching (TT)	8.14 *(2.69)*	11.05 *(3.34)*	1.06
FC with no SRL support (FCn)	9.12 *(3.02)*	11.59 *(3.47)*	0.80
FC with SRL support (FCS)	8.25 *(3.09)*	11.54 *(3.03)*	1.04

According to Cohen: $|d|$ = 0.2 small effect, $|d|$ = 0.5 medium effect, $|d|$ = 0.8 large effect.

different learning gains between the participants of the FCn condition and the FCS condition. Therefore, supporting students by explicitly providing them with strategies for watching instructional videos indeed seems to have a crucial effect. Since the model displayed in Table 8 uses FCn as reference group, it allows no inferential statement on the difference between TT and FCS. However, when, e.g., alternatively selecting TT as the reference group, no significant difference between the TT and the FCS condition occurs ($\beta_{FCS\times time}$ = 0.30, p = 0.22).

The Intraclass Correlation Coefficient (69%) was computed to assess the reliability of the model and the proportion of variance explained by the random effects. Furthermore, the marginal R^2 indicated that the fixed factors accounted for approximately 19% of the variance, while the conditional R^2 suggested that fixed and random factors together explained approximately 64% variance.

Including all control variables (see 3.3.6 and Supplementary Table 2, respectively) individually as possible predictors in the LMM, only the interaction of time with the student scale *strategies for using instructional videos* (after intervention) was significant (aggregated across all groups), providing further evidence from another (self-reported) perspective for the importance of teaching such strategies.

Regarding RQ1, the learning gains between FCn and TT did not differ significantly (without implementation of control variables), yet with a p-value close to significance (p = 0.07). Here, six student variables turned this interaction into being significant when individually implemented in the LMM (but no teacher variable). However, one must note that the p-value varied only moderately even in these cases, meaning that individually implementing student variables leads to corresponding p-values ranging from 0.02 to 0.13. Thus, the (nearly or indeed) significant difference between both groups was also relatively stable with respect to the control variables implemented (see Supplementary Table 2).

Concerning RQ2, only one of the teacher variables, namely the teachers' self-perceived instructional quality during the intervention, changes the significance of the differential improvements between FCS and FCn into non-significance (RQ2). Furthermore, none of these variables affected the non-significance of the differential improvements of the TT and the FCS group in the alternative model (see above). Taken together, this indicates a robust difference between the two FC conditions, on the one hand, and, on the other hand, a stable comparability between a professionally prepared TT and an FC teaching in which students are not left alone with instructional videos. Note that the superiority of FCS (significant with respect to FCn and only at a descriptive level compared to TT) is remarkable especially, when considering group differences in the control variables. Concerning the four variables in which an ANOVA revealed a significant difference (2 x *instructional quality, previous use of digital media/videos*; see Supplementary Table 2), the prerequisites in the FCS conditions were even *worse*.

4.3 Results at class level

Figure 4 provides a comprehensive visual summary of class-level interactions, illustrating the overall diversity observed in the study by highlighting the substantial variations between classes.

Note that a large range of both pretest and posttest scores was observed across the sample in general but also within each group (Figure 4). One intriguing class-related result is, for example, that class no. 36 in the FS n group that achieved the highest posttest score of 14.9 also reached the second highest pretest score of 10.6. Interestingly, this class also exhibited the second-largest improvement. Class no. 7, which was part of the FCS condition, had an average pretest score of 4.5 and demonstrated the greatest improvement with a gain of 4.8 points. Notably, in all conditions, there were single classes with small and with large improvements. In sum, Figure 4 illustrates that a) neither floor nor ceiling effects

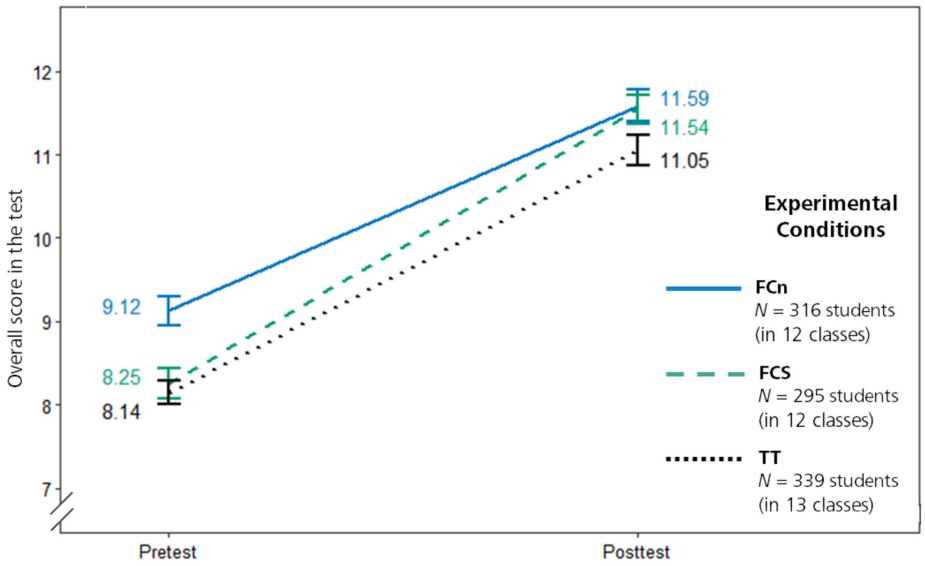

FIGURE 3
Students' performances regarding time (pre vs. post) and experimental conditions (TT vs. FCn vs. FCS), with bars indicating the standard error (SE). For easier identification of the interaction effects, the vertical axis was truncated at score = 6.

TABLE 8 Parameters of estimate, standard error, *t*-value and *p*-value (rounded to two decimals) of the LMM.

Marginal R^2 = 0.19	Estimate	*SE*	*t*	*p*
(Intercept)	**9.33**	0.66	14.21	**<0.001*****
Time	**2.48**	0.18	14.10	**<0.001*****
TT	−1.61	0.93	−1.72	0.10
FCS	−1.39	0.75	−1.85	0.07
TT × time	0.45	0.25	1.83	0.07
FCS × time	**0.76**	0.26	2.98	**0.003****

Estimate, Estimated Coefficients; *SE*, Standard Error; *t*, *t*-values; *p*, *p*-value; R^2, coefficient of determination; ICC, intraclass correlation. Values in bold are significant.

were present in the study, and b) no single class reduced their competence level, which indicates the effectiveness of the sequence in general. However, it also becomes clear that c) when randomly choosing just one or two classes per condition, based on our data, almost any interaction effect had been possible.

5 Discussion

The starting point of the present paper was an observed contradiction between a 'large' FC study (Wagner and Urhahne, 2021) and seven meta-analyses concerning K-12 teaching that had not yet included this recent study. While Wagner and Urhahne (2021) found no benefit of FC teaching in a highly controlled setting with several classes per experimental condition, all seven meta-analyses, in contrast, claim a beneficial effect of the FC method. Consulting general literature on meta-analyses, including critiques, revealed that it is no exception in empirical psychological and pedagogical research that many relatively small studies find effects that cannot be replicated by large experiments. Sotola (2022), for example, cautions that effect sizes in meta-analyses might be overestimated, especially in in-depth studies with small samples.

We attempted to consider these thoughts regarding FC studies in the K-12 classroom. Sorting the K-12 FC studies from the seven meta-analyses by the criterion "number of classes per experimental condition" (Supplementary Table 1) made apparent that most of the individual FC studies implemented only one or two classes per condition. This fact is not only in line with many theoretical deliberations of Sotola (2022), Kvarven et al. (2020), and Bartoš et al. (2023); it also underlines the call of the authors of the FC meta-analyses for large and controlled empirical studies investigating the FC method, particularly for school students (Låg and Sæle, 2019; van Alten et al., 2019; Wagner and Urhahne, 2021).

In the present paper, we specifically focus on K-12 teaching, in which self-regulation support for students might be more necessary than for university students, since the latter ones should have a higher responsibility for their learning progress than school students (van Alten et al., 2020a). In the empirical part, we implemented two experimental conditions (FCn vs. TT) to meet open demands and address the above contradiction (RQ1) through a further large and highly controlled study (and thus adding to Supplementary Table 1). In addition, a second FC condition was created (FCS) in which the students received an additional math-free SRL lesson to explicitly support working with instructional videos at home to answer RQ2 (contributing to Table 2).

Regarding RQ1, by applying a multi-level LMM, no significant difference was found between the learning gains of the students in the FCn and TT conditions. Yet, when implementing certain student control variables, the students of the TT group had a significantly higher learning increase. Regarding RQ2, the learning gains differed significantly in favor of the FCS group over the FCn group. All interaction effects were relatively stable when pertinent control scales entered the LMM as individual covariates.

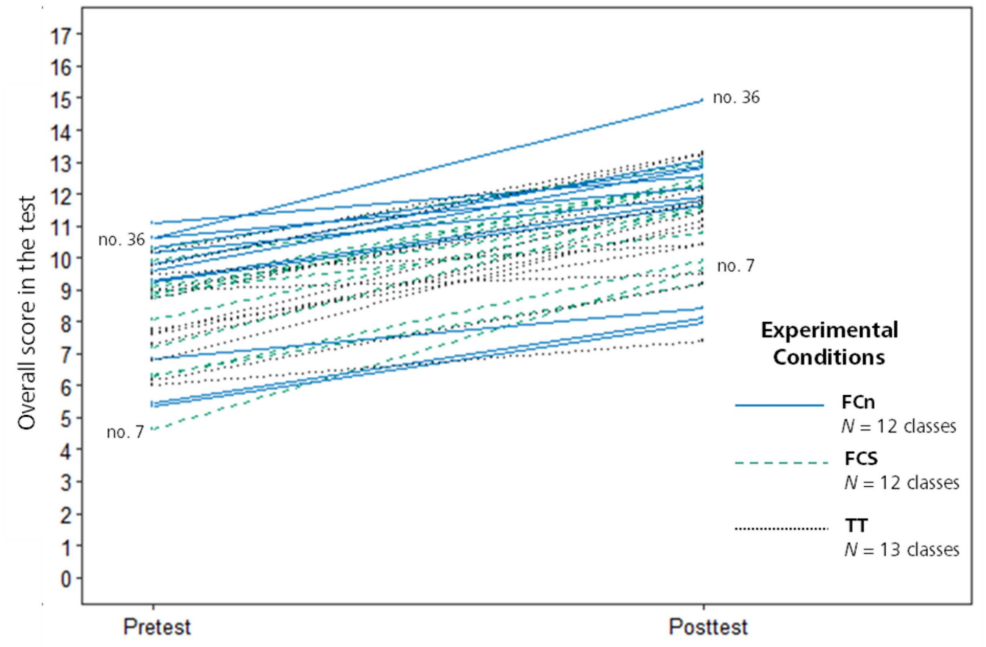

FIGURE 4
Learning gains of the 37 single classes.

Our study suggests that by providing school students with support for their SRL—i.e., by teaching specific strategies for working with instructional videos—FC might, indeed, become an effective method to enhance learning outcomes. One possible explanation for the significantly better outcomes observed in the FCS group can be provided by the ICAP framework (Chi and Wylie, 2014). When students passively watch the video without being aware that they need to *actively* engage in the content, their knowledge acquisition remains passive. However, when students are made aware that simply watching videos may lead to missed opportunities for learning and that certain strategies can enhance their learning experience, their interaction with the learning content surpasses mere passivity. In fact, Chi and Wylie (2014) suggest that, for instance, organizing and sorting notes can be viewed as an interactive process. Therefore, it might be crucial to allocate sufficient time to prepare students comprehensively to ensure a successful implementation of FC.

The present study highlights the importance of including several classes in educational interventions and using statistical multilevel modeling in general. Particularly, Figure 4 emphasizes the need of addressing class effects to avoid misleading conclusions. Therefore, a comprehensive approach incorporating diverse classes within each condition is crucial for as internally valid assessments of intervention effects as possible.

6 Limitations and future research

One limitation of our study is the short intervention duration of 2 weeks. This short time frame perhaps did not allow the students to fully adapt to the FC method, and novelty effects cannot be entirely ruled out, even if many students may have already encountered this approach during the COVID-19 pandemic. In future research, it would be possible, for instance, to first "train" the FC method prior to actual data collection, allowing students to become more accustomed to the approach.

In addition, it is important to recognize that the results may not be generalizable to the entire K-12 education system because of the different characteristics of students in different grades or concerning different subjects. Conducting large-scale studies with multiple classes per condition would be desirable with other grades and subjects in order to obtain a clearer picture of the effectiveness of the FC approach in primary and secondary teaching. Currently, in the FALKE-d project, which includes this study, analyses are being conducted based on parallel designs and comparable sample sizes in five other school disciplines (chemistry, physics, music, german and politcs in elementary school). Furthermore and for the same reasons, it is not possible to infer conclusions on the tertiary level from the present study. Expanding ideas from our theoretical part to university teaching would be intriguing, for example, to see whether comparable contradictions exist there as well.

Furthermore, with the breakthrough of AI, knowledge transfer could be more individualized for students (e.g., via chatbots rather than via videos). Investigating the impact of switching to such forms of instruction is still in its infancy (Lo and Hew, 2023). To the best of our knowledge, empirical findings are currently available only for the higher education sector (e.g., Tang et al., 2025).

7 Conclusion

Since the completion of our study, now, two large FC studies in K-12 education contradict the findings of seven meta-analyses. The years of publication of the meta-analyses (and of the individual FC studies) show that research on FC is still in the early stages. We are in line with van Alten et al. (2019, p. 15) that "Careful attention

should be paid, however, to the design of the flipped classroom as simply flipping before and during classroom activities might be not enough." Our data suggest that in addition to "flipping" the classroom in K-12 teaching practice, students must be supported to watch instructional videos at home or this format will not be beneficial to them.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material; further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval of an ethical council was not required for the study involving human samples in accordance with the local legislation and institutional requirements because approval has been obtained from the state government. Written informed consent for participation in this study was providedby the participants' legal guardians/next of kin.

Author contributions

PW: Software, Conceptualization, Writing – review & editing, Investigation, Resources, Writing – original draft, Supervision, Data curation, Project administration, Validation, Methodology, Formal analysis, Visualization. SK: Methodology, Conceptualization, Validation, Investigation, Supervision, Resources, Writing – review & editing, Funding acquisition, Project administration, Writing – original draft. NS: Writing – review & editing, Writing – original draft, Data curation, Investigation. KB: Validation, Writing – review & editing, Writing – original draft, Formal analysis, Software.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This project is part of the "Qualitätsoffensive Lehrerbildung," a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication. (grant number: 1JA2010). Open Access publishing supported by the University of Regensburg.

Acknowledgments

The authors would like to thank all the members of the FALKE-d project, of which this study was a part. The authors would also like to thank the participating teachers and students.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2026.1741733/full#supplementary-material

References

Bartoš, F., Maier, M., Shanks, D. R., Stanley, T. D., Sladekova, M., and Wagenmakers, E.-J. (2023). Meta-analyses in psychology often overestimate evidence for and size of effects. *R. Soc. Open Sci.* 10:230224. doi: 10.1098/rsos.230224

Bergmann, J., and Sams, A. (2012). Flip your classroom: Reach every student in every class every day. Eugene, Oregon: International Society for Technology in Education.

Binder, K., Krauss, S., and Bruckmaier, G. (2015). Effects of visualizing statistical information - an empirical study on tree diagrams and 2 × 2 tables. *Front. Psychol.* 6:1186. doi: 10.3389/fpsyg.2015.01186

Bishop, J., and Verleger, M. (2013). The flipped classroom: a survey of the research, in 2013 ASEE Annual Conference & 6/23/2013, 23.1200.1–23.1200.18.

Brame, C. J. (2016). Effective educational videos: principles and guidelines for maximizing student learning from video content. *CBE Life Sci. Educ.* 15:125. doi: 10.1187/cbe.16-03-0125

Cevikbas, M., and Kaiser, G. (2020). Flipped classroom as a reform-oriented approach to teaching mathematics. *ZDM* 52, 1291–1305. doi: 10.1007/s11858-020-01191-5

Cevikbas, M., and Kaiser, G. (2023). Can flipped classroom pedagogy offer promising perspectives for mathematics education on pandemic-related issues? A systematic literature review. *ZDM* 55, 177–191. doi: 10.1007/s11858-022-01388-w

Cheng, L., Ritzhaupt, A. D., and Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: a meta-analysis. *Educ. Technol. Res. Dev.* 67, 793–824. doi: 10.1007/s11423-018-9633-7

Chi, M. T. H., and Wylie, R. (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* 49, 219–243. doi: 10.1080/00461520.2014.965823

Dignath, C., and Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacogn. Learn.* 3, 231–264. doi: 10.1007/s11409-008-9029-x

Doğan, Y., Batdı, V., and Yaşar, M. D. (2021). Effectiveness of flipped classroom practices in teaching of science: a mixed research synthesis. *Res. Sci. Technol. Educ.* 41, 393–421. doi: 10.1080/02635143.2021.1909553

Frei, M., Asen-Molz, K., Hilbert, S., Schilcher, A., and Krauss, S. (2020). Die Wirksamkeit von Erklärvideos im Rahmen der Methode Flipped Classroom. Bildung, Schule, Digitalisierung. 34, 284–290.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295x.102.4.684

Gillette, C., Rudolph, M., Kimble, C., Rockich-Winston, N., Smith, L., and Broedel-Zaugg, K. (2018). A Meta-analysis of outcomes comparing flipped classroom and lecture. *Am. J. Pharm. Educ.* 82:6898. doi: 10.5688/ajpe6898

Güler, M., Kokoç, M., and Önder Bütüner, S. (2022). Does a flipped classroom model work in mathematics education? A meta-analysis. *Educ. Inf. Technol.* 28, 57–79. doi: 10.1007/s10639-022-11143-z

Hew, K. F., Bai, S., Dawson, P., and Lo, C. K. (2021). Meta-analyses of flipped classroom studies: a review of methodology. *Educ. Res. Rev.* 33:100393. doi: 10.1016/j.edurev.2021.100393

Hilbert, S., Stadler, M., Lindl, A., Naumann, F., and Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *Test. Psychomet. Methodol. Appl. Psychol.* 26, 101–119. doi: 10.4473/TPM26.1.6

Hwang, G.-J., Yin, C., and Chu, H.-C. (2019). The era of flipped learning: promoting active learning and higher order thinking with innovative flipped learning strategies and supporting systems. *Interact. Learn. Environ.* 27, 991–994. doi: 10.1080/10494820.2019.1667150

Kapur, M., Hattie, J., Grossman, I., and Sinha, T. (2022). Fail, flip, fix, and feed – rethinking flipped learning: a review of meta-analyses and a subsequent meta-analysis. *Front. Educ.* 7:956416. doi: 10.3389/feduc.2022.956416

Kvarven, A., Strømland, E., and Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nat. Hum. Behav.* 4, 423–434. doi: 10.1038/s41562-019-0787-z

Låg, T., and Sæle, R. G. (2019). Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis. *AERA Open* 5:233285841987048. doi: 10.1177/2332858419870489

Lage, M. J., Platt, G. J., and Treglia, M. (2000). Inverting the classroom: a gateway to creating an inclusive learning environment. *J. Econ. Educ.* 31, 30–43. doi: 10.1080/00220480009596759

Lai, C.-L., and Hwang, G.-J. (2016). A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Comput. Educ.* 100, 126–140. doi: 10.1016/j.compedu.2016.05.006

Li, S., Fu, W., Liu, X., and Hwang, G.-J. (2024). Effectiveness of flipped classrooms for K–12 students: evidence from a three-level meta-analysis. *Rev. Educ. Res.* 95:261732. doi: 10.3102/00346543241261732

Lo, C. K., and Hew, K. F. (2023). A review of integrating AI-based chatbots into flipped learning: new possibilities and challenges. *Front. Educ.* 8:1175715. doi: 10.3389/feduc.2023.1175715

Lo, C. K., Hew, K. F., and Chen, G. (2017). Toward a set of design principles for mathematics flipped classrooms: a synthesis of research in mathematics education. *Educ. Res. Rev.* 22, 50–73. doi: 10.1016/j.edurev.2017.08.002

Lubrick, M., Zhou, G., and Zhang, J. (2019). Is the future bright? The potential of lightboard videos for student achievement and engagement in learning. *Eurasia J. Math. Sci. Technol. Educ.* 15:108437. doi: 10.29333/ejmste/108437

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70, 487–498. doi: 10.1037/a0039400

Mayer, R. E., Fiorella, L., and Stull, A. (2020). Five ways to increase the effectiveness of instructional video. *Educ. Technol. Res. Dev.* 68, 837–852. doi: 10.1007/s11423-020-09749-6

McDowell, M., and Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol. Bull.* 143, 1273–1312. doi: 10.1037/bul0000126

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *Int. J. Educ. Res.* 31, 459–470. doi: 10.1016/S0883-0355(99)00015-4

Shrout, P. E., and Rodgers, J. L. (2018). Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* 69, 487–510. doi: 10.1146/annurev-psych-122216-011845

Simons, D. J. (2010). Monkeying around with the gorillas in our midst: familiarity with an inattentional-blindness task does not improve the detection of unexpected events. *i-Perception* 1, 3–6.

Sotola, L. K. (2022). Garbage in, garbage out? Evaluating the evidentiary value of published Meta-analyses using Z-curve analysis. *Collabra Psychol.* 8:32571. doi: 10.1525/collabra.32571

Stoeger, H., Fleischmann, S., and Obergriesser, S. (2015). Self-regulated learning (SRL) and the gifted learner in primary school: the theoretical basis and empirical findings on a research program dedicated to ensuring that all students learn to regulate their own learning. *Asia Pac. Educ. Rev.* 16, 257–267. doi: 10.1007/s12564-015-9376-7

Strelan, P., Osborn, A., and Palmer, E. (2020). The flipped classroom: a meta-analysis of effects on student performance across disciplines and education levels. *Educ. Res. Rev.* 30:100314. doi: 10.1016/j.edurev.2020.100314

Tang, V., Painho, M., and Vorobeva, D. (2025). Integrating tailored generative AI into the flipped classroom: a pilot implementation in higher education. *Innov. Educ. Teach. Int.*, 62:1–19. doi: 10.1080/14703297.2025.2523898

van Alten, D. C., Phielix, C., Janssen, J., and Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: a meta-analysis. *Educ. Res. Rev.* 28:100281. doi: 10.1016/j.edurev.2019.05.003

van Alten, D. C., Phielix, C., Janssen, J., and Kester, L. (2020a). Effects of self-regulated learning prompts in a flipped history classroom. *Comput. Hum. Behav.* 108:106318. doi: 10.1016/j.chb.2020.106318

van Alten, D. C., Phielix, C., Janssen, J., and Kester, L. (2020b). Self-regulated learning support in flipped learning videos enhances learning outcomes. *Comput. Educ.* 158:104000. doi: 10.1016/j.compedu.2020.104000

Vosniadou, S. (2020). Bridging secondary and higher education. The importance of self-regulated learning. *Eur. Rev.* 28, S94–S103. doi: 10.1017/s1062798720000939

Wagner, M., Gegenfurtner, A., and Urhahne, D. (2020). Effectiveness of the flipped classroom on student achievement in secondary education: a meta-analysis. *Z. Padagog. Psychol.* 35, 11–31. doi: 10.1024/1010-0652/a000274

Wagner, M., and Urhahne, D. (2021). Disentangling the effects of flipped classroom instruction in EFL secondary education: when is it effective and for whom? *Learn. Instr.* 75:101490. doi: 10.1016/j.learninstruc.2021.101490

Wiesner, P., Binder, K., Krauss, S., Steib, N., and Leusch, C. (2023). Sechs verschiedene Darstellungsarten für "25%" - und wie man sie ineinander umrechnen kann. *Stoch. Sch.* 43, 2–12.

Zeidner, M., and Stoeger, H. (2019) 'Self-regulated learning (SRL): a guide for the perplexed', High Abil. Stud., vol. 30, 1–2, pp. 9–51.

Zhang, F., Wang, H., Zhang, H., and Sun, Q. (2024). The landscape of flipped classroom research: a bibliometrics analysis. *Front. Educ.* 9:1165547. doi: 10.3389/feduc.2024.1165547

Zhu, G. (2021). Is flipping effective? A meta-analysis of the effect of flipped instruction on K-12 students' academic achievement. *Educ. Technol. Res. Dev.* 69, 733–761. doi: 10.1007/s11423-021-09983-6