Educational Research Review 28 (2019) 100281

Contents lists available at ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/edurev

Thematic Review

Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis

David C.D. van Alten*, Chris Phielix, Jeroen Janssen, Liesbeth Kester

Utrecht University, the Netherlands

ARTICLE INFO

Keywords:
Flipping the classroom
Inverted classroom
Learning outcomes
meta-Analysis
Satisfaction

ABSTRACT

In a flipped classroom, students study instructional material before class and apply this material during class. To provide a statistical synthesis of current research on effects of flipped classrooms, we conducted meta-analyses that included 114 studies which compared flipped and non-flipped classrooms in secondary and postsecondary education. We found a small positive effect on learning outcomes, but no effect was found on student satisfaction regarding the learning environment. In addition, we found considerable heterogeneity between studies. Moderator analyses showed that students in flipped classrooms achieve higher learning outcomes when the face-to-face class time was not reduced compared to non-flipped classrooms, or when quizzes were added in the flipped classrooms. We conclude that a flipping the classroom (FTC) approach is a promising pedagogical approach when appropriately designed. Our results provide insights into effective instructional FTC design characteristics that support an evidence-informed application of FTC.

1. Introduction

In a *flipping the classroom* (FTC) approach, students study instructional material before class (e.g., by watching online lectures) and apply the learning material during class. The popularity of this pedagogical approach, also known as *flipped learning* or the *inverted classroom*, has been growing rapidly during the last decade and has been applied and investigated in a wide variety of educational contexts (Bergmann & Sams, 2012). However, an overarching quantitative synthesis of existing empirical research into the effects of FTC, which would make it possible to draw general conclusions about FTC, is lacking. Therefore, we present a meta-analysis investigating the effects of FTC in secondary and postsecondary education on learning outcomes and student satisfaction, in comparison with a traditional (i.e., non-flipped) classroom in which students spend most classroom time on acquiring knowledge through lecture activities. We also provide valuable insights into possible differential effects of FTC by including several instructional design characteristics of FTC (such as group assignments, quizzes, and lecture activities) in a moderator analysis. The outcomes of this study is relevant to a wide variety of stakeholders such as teachers, educational policy makers and scholars who are interested in an evidence-informed application of FTC.

1.1. Previous research

In the current literature on FTC, a large variety of definitions of FTC can be found (Abeysekera & Dawson, 2015). We define a flipped classroom as students preparing instructional material before class (e.g., by watching a lecture video), and applying the

* Corresponding author. Department of Education, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, the Netherlands.
E-mail address: d.c.d.vanalten@uu.nl (D.C.D. van Alten).

https://doi.org/10.1016/j.edurev.2019.05.003
Received 9 November 2018; Received in revised form 15 March 2019; Accepted 29 May 2019
Available online 01 June 2019

instructional material during class (e.g., by working on problem solving assignments). An issue related to the definition of FTC seems to be whether the use of videos and/or computer technology before class defines FTC (see also Bernard, 2015). Some researchers define FTC as students watching online instructional videos before class (Cheng, Ritzhaupt, & Antonenko, 2018; Hew & Lo, 2018; Lo, Hew, & Chen, 2017; Scott, Green, & Etheridge, 2016; Strayer, 2012), or receiving computer-based individual instruction outside the classroom (Bishop & Verleger, 2013; Lo & Hew, 2017). Other researchers, however, adopt a more generic definition of FTC, in which neither videos nor technology necessarily define FTC (Abeysekera & Dawson, 2015; Akçayır & Akçayır, 2018; Kim, Kim, Khera, & Getman, 2014; Lage, Platt, & Treglia, 2000; Mullen & Sullivan Jr., 2015; O'Flaherty & Phillips, 2015; Peterson, 2015). In this broader definition, activities before class such as reading material are also considered to be a part of FTC. We chose to adhere to the last definition of FTC, because in our view the medium students use before class in itself does not affect learning in general (Clark, 1994; DeLozier & Rhodes, 2017). Moreover, our aim is to conduct a comprehensive review study of the effects of FTC without excluding certain design characteristics of FTC in advance.

Important benefits and challenges of FTC are provided by recent review studies (e.g., Akçayır & Akçayır, 2018; Lo et al., 2017). In addition, the application of FTC can also be evaluated by the following suppositions from an educational psychological perspective. First, conveying information and instruction through conventional lectures in the traditional classroom has its shortcomings, because the teacher delivers the instructional message in a one-size-fits-all manner and the possibilities for the students to interact with the teacher during the lecture (e.g., to ask or answer questions) are limited. This might lead to a lack of student engagement and it might hamper students to actively construct knowledge (Schmidt, Wagener, Smeets, Keemink, & Van Der Molen, 2015). However, this should not imply that teachers should stop ginving direct instruction. In an FTC approach, therefore, students receive information and instruction before class, for example, through video lectures. This has the advantage that students are able to control the viewing frequency and viewing pace of the instruction material before class (Abeysekera & Dawson, 2015). In this way, cognitive load could be reduced as learner-controlled video segments are better able to support students processing the learning material (i.e., segmentation effect, see Clark, Nguyen, & Sweller, 2005).

It is important to consider that a flipped classroom could induce an additional demand on students as it increases the appeal on the self-regulated learning capabilities of students. While some researchers presume that this may lead to positive effects in terms of learning outcomes and self-regulated learning capabilities of students (e.g., Lape et al., 2014), others hypothesize that student's lacking self-regulated learning capabilities could experience disadvantages in a flipped classroom (e.g., Lai & Hwang, 2016).

Second, students in traditional classrooms independently work on homework assignments to apply the information and instruction that was presented to them in lectures. From a cognitivist perspective, however, a lack of direct instructional guidance during application might result in cognitive overload and hinders students to store knowledge in their long-term memory (Kirschner, Sweller, & Clark, 2006). In a flipped classroom, however, students can devote classroom time to learning activities focused on applying the instructional material (such as working through problems and engaging in collaborative learning) with the guidance of the teacher (Roehl, Reddy, & Shannon, 2013). From an educational psychological perspective, this is beneficial for learning because learning happens when the learner is actively generating meaning, instead of passively receiving information (Wittrock, 1992). Our mind actively and dynamically constructs meaning by building relations with prior-knowledge (Wittrock, 1992). In the flipped classroom, this learning process is guided by the teachers. Students in a traditional classroom lack this guidance when they engage in similar activities during homework assignments after class.

Third, according to Chi and Wylie (2014), students can achieve deeper understanding of the learning material as they become more engaged. Chi and Wylie (2014) distinguish the following modes of engagement: *passive* (e.g., receiving knowledge, listening to a lecture and being able to recall information), *active* (e.g., manipulating knowledge, taking notes and being able to apply knowledge to similar contexts), *constructive* (e.g., generating knowledge, comparing and contrasting information and being able to transfer knowledge or procedures), and *interactive* (e.g., dialoguing, discussing with peers, being able to co-create knowledge) as modes of engagement. In an FTC approach, students are better able to achieve higher learning outcomes, as there is more classroom time available for learning activities that foster active, constructive, and interactive engagement modes. In contrast, in a traditional classroom most classroom time is devoted to activities such as lectures, with mostly a passive mode of engagement.

Fourth, as compared to traditional classrooms, FTC allows more classroom time for students to interact with their teacher(s) and peers, and to engage in learning activities which are differentiated based on their capabilities and needs (Bergmann & Sams, 2012). Classroom interaction enhances student relatedness (Abeysekera & Dawson, 2015), stimulates collaborative learning (DeLozier & Rhodes, 2017), and offers students the opportunity to receive support from their teacher or peers when applying knowledge during class. For instance, Van den Bergh, Ros, and Beijaard (2014) demonstrated that an essential component of a teacher's role in an active learning classroom (i.e., a classroom in which students engage in learning activities instead of passively listening) is to provide feedback to students to guide and facilitate students' learning processes. Thus, in a flipped classroom, there may be more room for students to receive effective feedback and differentiated instruction from their teacher.

Finally, FTC is often assumed to be a promising pedagogical approach that increases student satisfaction about the learning environment (O'Flaherty & Phillips, 2015; Seery, 2015). Lo and Hew (2017) and Betihavas, Bridgman, Kornhaber, and Cross (2016) however found mixed results about students' attitudes towards FTC; in general, students were positive about the learning environment, but a few studies showed opposite results. From a theoretical perspective, Abeysekera and Dawson (2015) hypothesize that a flipped classroom approach is able to satisfy students' needs for competence, autonomy and relatedness and, thus, entice greater levels of motivation (according to self-determination theory, see: Ryan & Deci, 2000).

Recent narrative reviews of FTC provided general hypotheses and views on the effects of FTC (e.g., Abeysekera & Dawson, 2015; Akçayır & Akçayır, 2018; Betihavas et al., 2016; Bishop & Verleger, 2013; DeLozier & Rhodes, 2017; Lo & Hew, 2017; Lundin, Bergviken Rensfeldt, Hillman, Lantz-Andersson, & Peterson, 2018; O'Flaherty & Phillips, 2015; Seery, 2015). However, these reviews

need to be interpreted with caution. Generalization of their conclusions is difficult because the number of studies considered in these reviews was small, and the included studies often lack the use of a control group (Lundin et al., 2018; O'Flaherty & Phillips, 2015). For these reasons, we conducted a meta-analysis, as a comprehensive synthesis of empirical studies that is able to draw more general conclusions about the effects of FTC on student learning outcomes and satisfaction.

Recently published meta-analyses already showed the potential of FTC in comparison with a more traditional approach in terms of student achievement. Lo et al. (2017) analyzed 21 studies in mathematics education and found a small effect ($g = 0.30$), and Hew and Lo (2018) analyzed 28 studies in health professions education and found a similar small effect ($g = 0.33$). Chen et al. (2018) compared 32 studies from health science and 14 non-health science studies and found an effect size of 0.47, while Cheng et al. (2018) included 55 studies from every domain and found an average effect of 0.19.

The current meta-analysis is able to add something to the scientific and practical relevance of these studies for the following reasons. First, we made no distinction in study domain or publication type, and were therefore able to include more studies ($k = 114$) to present a comprehensive meta-analysis on the effects of FTC. Second, our meta-analysis includes more outcome variables than student achievement, as we analyzed the effects of FTC on three types of outcome: assessed learning outcomes (determined by assessments such as exams), perceived learning outcomes (determined by students themselves in questionnaires), and student satisfaction. Third, we conducted extensive moderator analyses to examine if deeper lying instructional design characteristics, educational context characteristics, or study quality characteristics affect the effectiveness and attractiveness of FTC. We used meta-regression models to better explain the variation between studies that occurs if the heterogeneity between effect sizes is larger than can be expected by sampling error only (Borenstein, Hedges, Higgins, & Rothstein, 2009).

1.2. Rationale moderator variables

There is still little empirical evidence that provides insights into the instructional design characteristics of an effective flipped classroom approach, and therefore we examine if different implementations of a flipped classroom have different effects. We selected moderator variables based on previous research. First, we consulted FTC review studies determine which theoretical grounding is used to explain variance of effects. Second, we conducted a preliminary literature search to determine which variables are frequently reported in FTC intervention studies. We present our rationale for each included moderator variable in three categories: design characteristics, educational context, and study quality.

1.2.1. Design characteristics

The following moderator variables have been selected as FTC instructional design characteristics: *adding quizzes*, *adding small group assignments*, *face-to-face classroom time,* and *lecture activities*. *Quizzes* and clicker questions are frequently implemented in flipped classrooms, for example, to determine to what extent students understand the instructional material provided before class (DeLozier & Rhodes, 2017). Previous research has shown that quizzes have a positive effect on learning outcomes because of the testing effect (Bangert-Drowns, Kulik, & Kulik, 1991; Dirkx, Kester, & Kirschner, 2014; Roediger & Karpicke, 2006). In addition, Spanjers et al. (2015) found in their meta-analysis on blended learning that the inclusion of quizzes positively moderated the effects of blended learning on learning outcomes and satisfaction. As FTC can be regarded as a particular example of blended learning, defined as the combination between online and face-to-face education (Graham, 2006), we expected to find similar results. Hew and Lo (2018) and Lo et al. (2017) also found that instructors in a flipped classroom who employed quizzes at the start of their classes yielded higher learning outcomes as opposed to instructors who did not.

Another instructional design feature that is often applied in flipped classrooms is the addition of *small group assignments*. DeLozier and Rhodes (2017) showed that learning in small groups with learning activities such as pair-and-share, paired problem-solving assignments and group discussions in the FTC context enhances learning outcomes. Research has shown positive effects of co-operative learning on achievement and attitudes (Kyndt et al., 2013; Slavin, 1991). Lo et al. (2017) also suggest that small-group learning activities could enhance the effectivity of a flipped (mathematics) classroom.

Furthermore, we examined if the amount of *face-to-face classroom time* might explain variance of the effects between the studies. It is sometimes argued that face-to-face classroom time can be reduced in a flipped classroom because students spend more time on learning activities before class. For example, a study on reducing classroom time in higher education by two-thirds through implementing FTC showed that students in the flipped classroom achieved learning outcomes that were at least as good, and in one comparison even significantly better than the learning outcomes of students in the traditional classroom (Baepler, Walker, & Driessen, 2014). So, the authors concluded that active learning classrooms (blended or flipped) with less face-to-face time are at least as effective as traditional classrooms, and are thus more cost-efficient. In contrast, in a study that reduced face-to-face classroom time by making use of FTC in order to help solve the shortage of teachers in secondary education (Heyma et al., 2015), it appeared that students in the flipped classroom performed significantly worse in terms of learning outcomes. The authors suggested this could be due the students inability to deal with more responsibility to regulate their own learning (Heyma et al., 2015).

Finally, we analyzed the use of *lecture activities* during the flipped classroom as a potential moderator variable. Although it is the aim of a flipped classroom to reduce in class lectures and present instructional material before class, Walker, Cotner, Baepler, and Decker (2008) concluded that students still highly value the integration of in class microlectures in a flipped classroom. Moreover, McLaughlin et al. (2013) suggested that in class microlectures in a flipped classroom can be used to address misunderstandings or gaps in student knowledge. Still, large-group lectures have their shortcomings in terms of effectiveness and engagement (Schmidt et al., 2015).

1.2.2. Educational context

The following moderator variables with regard to the educational context were seen as valuable: *academic domain*, *educational level*, and *intervention duration*. We examined the possible influence of the *academic domain* in which FTC is applied (with humanities, social sciences, natural sciences and formal sciences as categories). In this way, we wanted to investigate whether FTC might be more effective in different academic domains. Only a small number of studies investigated possible differential effects of FTC in particular academic domains (Bernard, 2015; Hew & Lo, 2018; Lo et al., 2017). O'Flaherty and Phillips (2015) suggested that some academic domains may be better suited to apply FTC, as studies in different domains reported positive and negative student perceptions of FTC. In other research, academic domain sometimes appeared to significantly moderate the findings as well. For example, in a meta-analysis on the effects of technology use in postsecondary education (Schmid et al., 2014) and in a meta-analysis on the effects of cooperative learning (Kyndt et al., 2013).

Moreover, we included *educational level* (e.g., secondary and post-secondary education) in the moderator analyses to see whether FTC is as effective in higher education as in lower levels of education (Lo & Hew, 2017). FTC, as a more student-centered approach (Calimeris & Sauer, 2015), might work differently for students of different age groups because of differences in self-regulated learning abilities which tend to increase at higher ages (Wigfield, Klauda, & Cambria, 2011; Zimmerman & Martinez-Pons, 1990). We acknowledge that a flipped classroom could have different meanings in secondary and higher education and that the two settings can sometimes be very different in the way they are structured. It seems safe to assume, however, that the main working mechanism of FTC is similar in both secondary and higher education, as it has similar design characteristics. As we do not want to underestimate the difference in both educational contexts, we added educational type as a possible moderator variable to research if studies conducted in both contexts yielded differential effects.

Lastly, we considered *intervention duration* as moderator variable. Effects of FTC may be affected by a novelty effect, which means that just because FTC is a new approach, it affects the learning outcomes or satisfaction of students positively or negatively (Chandra & Lloyd, 2008). Furthermore, intervention duration might moderate the results if students need to get used for a longer period of time in order for a flipped classroom to work effectively.

1.2.3. Study quality

It is also possible that studies with a more rigorous experimental study design yielded significantly different results than studies with less quality (Cheung & Slavin, 2016). For example, in a meta-analysis by Lazonder and Harmsen (2016), it appeared that randomized experiments yielded significantly higher effects of guidance in inquiry-based learning on performance success than quasi-experiments. In previous narrative reviews on FTC it seemed that in the initial phase of research the methodological quality of research could be improved as, for instance, only a small proportion of the studies used a randomized design with a control group (Bishop & Verleger, 2013; DeLozier & Rhodes, 2017). Consequently, we included the following moderator variables concerning the study quality, to investigate if these moderator variables have an impact on the effects of FTC to estimate the magnitude of this problem. *Allocation type* was coded to distinguish studies that randomly allocated students to the flipped or control condition, randomly allocated pre-existing groups (e.g., there are eight classes in one school and four of them were randomly allocated to the flipped condition), and studies that were not able to apply randomization. *Control group type* constitutes the difference between studies that used a previous cohort as control group and studies that simultaneously studied a flipped and control condition (Cheung & Slavin, 2016). *Group equivalence test* means the manner in which studies assessed the comparability of students in the flipped and control conditions before the intervention in terms of prior knowledge (e.g., statistically tested prior knowledge in the form of a SAT math score or GPA, providing a descriptive statement, or not mentioned). Lastly, we examined *study type* as a possible moderator to investigate whether there is a difference in effect sizes reported in journals, dissertations, and conference papers. In this way, we could evaluate the presence of publication bias in our data, which means that published studies contain a significant higher effect size than non-published studies (Torgerson, 2006).

1.3. Research objectives

Our first aim was to conduct three meta-analyses to examine the average effect of a flipped classroom on assessed learning outcomes, perceived learning outcomes, and student satisfaction in comparison with a traditional classroom. We separately analyzed assessed and perceived learning outcomes as dependent variables, to make a distinction between learning outcomes determined by assessments (e.g., exams) and students themselves (e.g., questionnaire). As DeLozier and Rhodes (2017) have showed, initial (qualitative) research attempts on FTC were mainly focused on how students perceived their own learning in a flipped classroom, and they convincingly argued that students are often unable to correctly assess their own learning gains (see: Kruger & Dunning, 1999). Thus, there seems to be a fundamental difference between assessed learning outcomes determined by someone else than the student and perceived learning outcomes determined by the student and usually measured with questionnaires on how students perceive their own learning gains. Students' satisfaction of the learning environment, as third dependent variable in our study, was chosen to examine the question if participants of a flipped classroom give higher satisfaction scores than students in a non-flipped classroom. As Eichler and Peeples (2016) showed, student evaluations are often taken into account by policy makers in their evaluation of flipped classrooms. Student satisfaction is determined by the students and usually measured by self-reports on how satisfied they were with the learning environment.

Our second aim was to conduct moderator analyses to investigate if variety of effects between studies can be explained by instructional design characteristics, educational context, or study quality characteristics. We expected a significant amount of heterogeneity of the effects that cannot only be explained by sampling error, because FTC implementations vary from classroom to

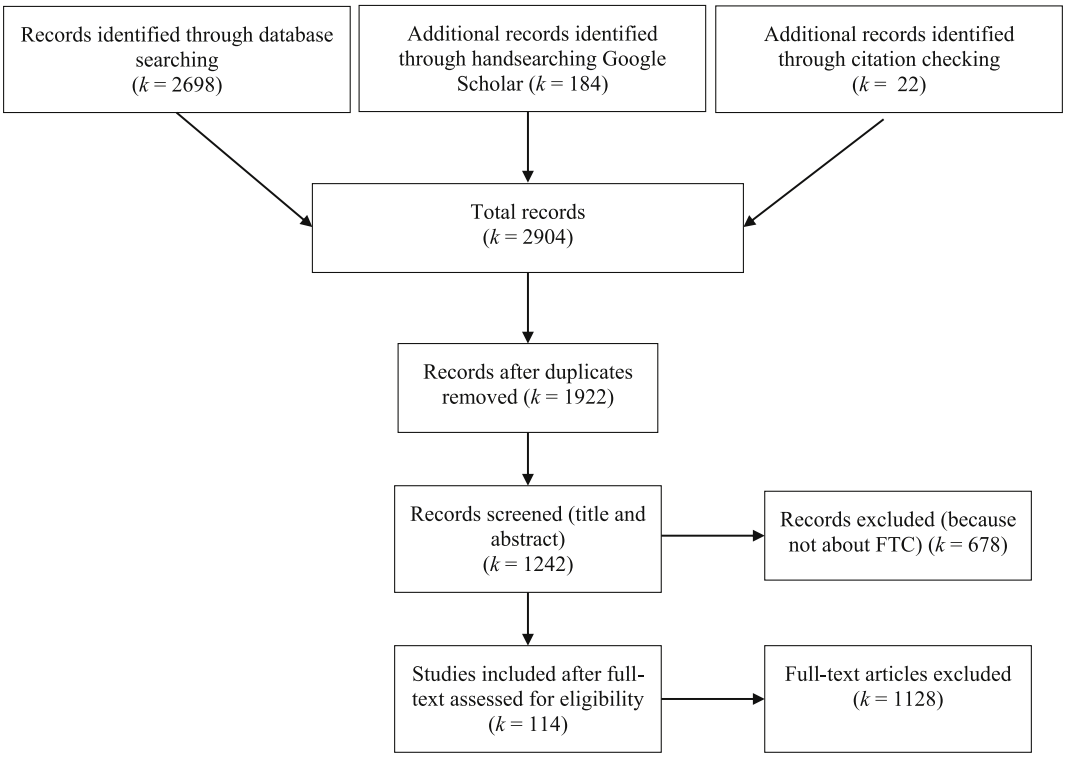

Fig. 1. Flow diagram of literature search and processing of records.

classroom and educational context to context. In short, the current study answers the following research questions: (a) what is the effectiveness of FTC on assessed and perceived learning outcomes and student satisfaction about the learning environment in comparison with a traditional classroom?; and (b) does the effect of FTC on assessed and perceived learning outcomes and student satisfaction depend on instructional design characteristics, educational context characteristics, or study quality characteristics?

2. Method

2.1. Literature search

The results of the literature search and the selection of articles can be consulted in Fig. 1, based on the PRISMA flow diagram (Moher, Liberati, Tetzlaff, Altman, & The PRISMA group, 2009). We decided to include unpublished studies (e.g., conference papers, and dissertations) and also take into account the possibility of publication bias, as recommended by Polanin, Tanner-Smith, and Hennessy (2016).

Our search process consisted of three parts. At the end of May 2016, we first consulted the databases OVID (MEDLINE® and PsycINFO), Web of Science, ERIC, and Scopus. As our goal was to retrieve all the available research on the phenomenon of flipping the classroom (including synonyms such as inverted classroom and flipped learning), the following search parameters were used in titles and abstracts: (Flip* OR invert*) AND (class* OR course* OR learning). A time period restriction was set to finding records after 1990, an ample margin since the concept of FTC appeared in the scientific literature at the end of that decade (i.e., Lage et al., 2000). Second, we manually screened Google Scholar with the same combination of search terms. Third, we manually checked the references of all the known (narrative) reviews of FTC. This yielded 22 additional records, of which three eventually were included. Two of these studies did not mention FTC, and one was an unpublished study that was not included in the consulted databases. We decided to include these three studies because their FTC interventions met our definition of FTC.

2.2. Inclusion of studies

Of the 2.904 retrieved records in total, the first author removed 982 duplicates (see Fig. 1). Then, during the first screening of titles and abstracts, 678 studies were excluded if a study's topic was clearly not about FTC (such as physics studies which mentioned *classes* of macromolecules in combination with *inverted* micellar behavior). To establish reliability of the screening process, every record was independently screened by at least two people (the first author screened every record, and the second and third author each screened half of the records). Disagreements between coders were resolved by discussion.

Next, the first three authors assessed the full texts of the remaining 1.242 records for eligibility by applying seven inclusion and

Table 1
Inclusion and exclusion criteria and numbers of studies excluded.

Criterion	Inclusion	Exclusion	Excluded studies
Language	English	Non-English language	14
Impact study	Studies investigating effects of FTC	All other non-impact studies	709
Control group	Studies making use of a traditional control group	Studies not making use of a traditional control group	290
Dependent variable	Studies investigating assessed learning outcomes, perceived learning outcomes or student satisfaction	All other studies investigating other dependent variables	34
Availability	The full-text of the study must be available to consult via University library systems or the internet	Studies of which the full-text was not available to consult	27
Statistical information	Studies reporting sufficient statistical information to calculate an effect size	Studies not reporting sufficient statistical information	54

Note. Articles can be excluded based on multiple criteria, but in this table for each study only the first encountered exclusion criterion is given.

exclusion criteria presented in Table 1. We excluded articles if the language was not English, if no quantitative effects of FTC were measured, and if studies did not make use of a between subjects design with a traditional classroom as control group (i.e., two-group RCT or quasi-experiments). In addition, studies which measured other dependent variables than learning outcomes or student satisfaction were excluded (e.g., they examined the opinions of involved teachers or directors, interaction in the classroom, self-regulated learning, collaborative learning, or critical thinking). In case the full-text of a record was not available via our own or other library systems, we consulted the websites of the scholars involved and sought contact with the first author to try to obtain the manuscript before we excluded a record.

We developed Table S1 (online only) with additional information of the 54 records we excluded, because they presented insufficient statistical information to calculate an effect size analysis (a complete list of references of these excluded studies is presented in appendix A, online only). In this table we provide an explanation of the insufficient data for each excluded study, together with parameters such as: the authors' conclusion with regard to assessed and perceived learning outcome and satisfaction, the publication type, and the total sample size.

Lastly, we checked our search database for cases in which a journal article was also retrieved as dissertation or conference paper (where the article was based on). This was also done for records that were excluded because insufficient statistical information was provided. This yielded one dissertation that was used to retrieve additional information in the coding process for that particular study. In total, 114 studies were included in our meta-analysis (a complete list of references of these included studies is presented in appendix B, online only).

2.3. Coding of study characteristics and moderator variables

The included studies were systematically analyzed by making use of two coding schemes: one scheme on the study level and one scheme on the effect size level (Lipsey & Wilson, 2001). In the first coding scheme, general study information (e.g., study type and publication year), sample characteristics (e.g., gender and age), educational context features (e.g., duration of the intervention, academic domain, and region), and methodological characteristics (e.g., type of comparison group, allocation, ratios face-to-face learning time between the two groups, and assessment of group equivalence) were coded for each study. In order to be able to compare possible effects of the various types of classroom activities, the following categories were used: lecture activities, individual assignments, group assignments, and quizzes (for similar categorizations of in class activities see DeLozier & Rhodes, 2017; Lo & Hew, 2017). For both the flipped and the control condition, it was coded whether each of these type of learning activities occurred during class, or outside class. We assumed that a particular classroom activity was not applied if it was not mentioned in the study (e.g., when a study mentions that group assignments were given to students in the flipped condition, but information about group work in the traditional condition was absent). Similarly, if studies did not explicitly state that face-to-face classroom time was reduced, sustained, or increased in the flipped condition, we assumed it was kept constant with the traditional condition. In the second coding scheme on effect size level, we gathered information about the type of dependent variable (e.g., assessed learning outcome) and the raw data that was used to calculate the particular effect size.

After three rounds of testing and improving the coding schemes, the first three authors and two trained research assistants independently double-coded a set of 22 randomly chosen articles (about 18% of the included studies). This resulted in 87% total agreement, and a substantial interrater-reliability with an average Cohen's κ of 0.77 (range 0.45–1, $p < .001$). In some cases an unrepresentatively low Cohen's κ could be explained because some categories were overrepresented while there was high agreement between the coders (e.g., the variable *region* in which the category North America was coded in 80% of the cases; see Feinstein & Cicchetti, 1990). Nevertheless, the first three authors independently double-coded all variables that had a Cohen's $\kappa < 0.80$ (i.e., educational level, gender, allocation, group equivalence test, face-to-face learning time, student experience, group assignments, and quizzes). Remaining disagreements were resolved by discussion amongst the first three authors.

2.4. Computation of effect sizes

We used the standardized mean difference as an effect size metric, calculated as the mean differences between the mean scores of

the flipped and control groups divided by the pooled standard deviation (Borenstein et al., 2009). A positive effect size in this study indicates that students in the flipped classroom (i.e., the experimental group) outperformed the students in the traditional classroom (i.e., the control group), students in the flipped classroom perceived they learned more than the students in the traditional classroom, or students in the flipped classroom were more satisfied with the learning environment. Hedges' g was used as effect size unit, which is a common adjustment of Cohen's d for studies with a small sample size (Hedges, 1981).

A total of 148 effect sizes were calculated on the basis of the following order of preference, because this hierarchical order maximizes the precision of the meta-analyses. The first preferred option was to calculate an effect size from adjusted posttest means that took potential pretreatment differences into account, either by regression analysis or covariance ($k = 8$). The second preferred method was to calculate effect sizes based on a pretest-posttest-control group design ($k = 18$). Morris (2008) demonstrated that an effect size based on the mean pre-post change in the intervention group minus the mean pre-post change in the control group divided by the pooled pretest standard deviation leads to the least biased estimate of treatment effects in pretest–posttest-control group designs. Formulas 8, 9 and 10 from Morris (2008) were used to calculate effect sizes, and formula 25 was used to calculate variances (formulas are provided in appendix C, online only). As only one study reported a correlation between the pretest and posttest ($r = 0.85$; Reza & Ijaz Baig, 2015) a sensitivity analysis was not possible (Borenstein et al., 2009). Therefore, we used a conservative estimation of $r = 0.70$, as recommended by Rosenthal (1991), for studies that did not report the correlation between pre- and posttest. In the case these above two options to calculate effect sizes were not possible, effect sizes were calculated on the basis of the raw means and standard deviations from the post-test ($k = 93$). When these were not reported, the least preferred method was the use of inferential statistics (such as the t-test, $k = 20$ and the F-test, $k = 9$). The online meta-analysis effect size calculator by Lipsey and Wilson (2001) was used in the case of adjusted post-test means and when studies reported raw means and standard errors. The metafor package (Viechtbauer, 2010) for the software program R (R Development Core Team, 2008) was used to calculate effect sizes from a pretest-posttest-control group design. In other cases, the software program Comprehensive Meta-Analysis (CMA, Version 3.3.070) was used.

2.5. Non-independence

Eight studies reported the data of multiple flipped or traditional classroom conditions. In these cases, the means and standard deviations were pooled for that particular condition (following the formulas provided in Spanjers et al., 2015). When multiple effect sizes of one dependent variable in one study could be calculated (e.g., from two midterm exams and one final exam), we chose to only include the measuring point at the end of the intervention to be able to make a fair comparison between all included studies. Ten studies presented multiple post-test results that matched one of our dependent variables (e.g., objective learning outcomes). An example would be a study that presented results for three different post-tests of objective learning outcomes: a recall, an application, and a comprehension test. In those cases individual fixed effect meta-analyses were carried out and the results were included as a composite effect size in the main meta-analysis (Borenstein et al., 2009). In addition, one study reported two different interventions (with different samples) and we treated this study as if it were two independent studies.

2.6. Data analysis

Three univariate meta-analyses were conducted for each dependent variable using the metafor package for R (Viechtbauer, 2010). We used the random-effects model, because we do not assume that there is one general effect size of FTC that is true for all studies (Borenstein et al., 2009). The study's variance consisted of a within study variance and between studies variance component, and studies were weighted according to the extent of variation (Borenstein et al., 2009). The effect of possible outliers in the data was checked by using forest and funnel plots and one-study-removed analysis in metafor for R.

We used the restricted maximum likelihood estimation method to fit all models. Besides standard error and 95% confidence intervals, we reported results of the Cochran Q-test for statistical heterogeneity. The I^2 statistic indicates that the proportion of the variation in effect sizes is due to true between-study heterogeneity rather than sampling error, and $> 75\%$ represents considerable heterogeneity. The τ^2 statistic indicates the extent of between-study variance. Using the Knapp and Hartung (2003) adjustment to the standard errors of the estimated coefficients, we accounted for the uncertainty in the estimate of residual heterogeneity. With this adjustment, individual coefficients and confidence intervals use the t-distribution with $k - p$ (p being the number of model coefficients) degrees of freedom.

We conducted moderator analyses with a mixed-effects meta-regression, allowing for between-study variation, using restricted maximum likelihood to estimate the amount of residual heterogeneity (τ^2_{res}). Moderator variables are added as (study level) covariates in the linear meta-regression model to examine systematic differences in the strength of the effect (Borenstein et al., 2009). For categorical moderator variables dummy variables were created and continuous variables were centered around the mean.

With the Knapp and Hartung (2003) adjustment, the omnibus test of the moderator analysis uses an F-distribution with m and $k - p$ degrees of freedom. I^2_{res} is given as a percentage of unaccounted variation between studies that is due to heterogeneity rather than sampling error and R^2 as the proportion of amount of heterogeneity accounted for by the moderator variable, often called the model predictive power (López-López, Marín-Martínez, Sánchez-Meca, Noortgate, & Viechtbauer, 2014).

Moderator analyses in a meta-analysis have certain statistical limitations (Hempel et al., 2013). Polanin and Pigott (2015) made clear that current practice in educational and psychological meta-analyses with regard to significance testing could be much improved. Moderator analyses are prone to Type I errors (incorrect rejection of a true null hypothesis leading to falsely identified moderator effects) if multiple (i.e. univariate) tests are conducted. Therefore, we used meta-regression models to account for multiple

testing (Polanin & Pigott, 2015). Type II errors are also common in moderator analyses, because of low power (Hempel et al., 2013). To aid the reader in understanding the magnitude of this problem, we provided retrospective power analysis where appropriate (Hedges & Pigott, 2004; Hempel et al., 2013). While it might be true that retrospective power analysis for non-significant results are uninformative if they only provide observed power (Valentine, Pigott, & Rothstein, 2010), we will use these (two-sided) power calculations to estimate the power to detect what we assume to be the smallest important effect size (i.e., $g = 0.20$) given the number of studies we found, and the average within study sample size in those studies (Hempel et al., 2013; Valentine et al., 2010).

2.7. Publication bias

Based on the recommendations by Banks, Kepes, and Banks (2012) for analyzing and reporting publication bias, we included unpublished studies to reduce the effect of publication bias on our data and added publication type as a moderator variable to explore the possibility that published journals yielded significantly higher effect sizes than unpublished papers. In addition, we performed a cumulative meta-analysis sorted on precision (i.e., increasing standard error, see Borenstein et al., 2009). Funnel plots were produced using the metafor package to inspect asymmetry and heterogeneity. We also computed a Pearson's correlation test between effect size and sample size, as a negative correlation could indicate publication bias (Levine, Asada, & Carpenter, 2009).

3. Results

3.1. Descriptive findings

The included studies fell into the time span from the first included article published in 2006, until the end of May 2016, which was the final date of the literature search. In Table S2 (online only), we present an overview of the variables publication type, educational level, academic domain, study design, outcome that was measured, the effect sizes, and sample sizes for each included study. In this Table S2 (online only), the design of each study is summarized in terms of the type of control group (simultaneous vs. previous cohort) and the type of allocation (completely randomized each individual student, randomized pre-existing classes, or no randomization to groups). Next, in Table 2 we provide the frequencies of a selection of variables we coded, for categories that contained at least one study. This can be a helpful way of getting an impression of the characteristics of our sample. In addition to the information in Table 2, we found that the years of experience of the teacher(s) with FTC was not reported in 50% of the studies, and in 41% of the studies it was their first experience with FTC. Age was reported in 20% of the studies ($M = 20.40$, $SD = 3.09$). It should also be noted that from the 12 studies that appeared not to have equal student populations in both research conditions after a group equivalence test (e.g., a *t*-test on GPA), as shown in Table 2, ten of these studies took this difference into account by means of, for example, an adjusted post-test means regression analysis.

In Fig. 2, a frequency distribution (in percentages) of the learning activities used in both the flipped and traditional classrooms is shown. About half of the flipped classrooms (53%) still contained some lecture activities inside the classroom, usually in the form of microlectures or just-in-time lectures (e.g., Eichler & Peeples, 2016). These microlectures are often tailored to the results of quizzes or other forms of formative assessment made by students. Furthermore, a clear distinction can be seen in the types and frequency of the different learning activity categories, where flipped classes clearly contained more active learning components than the traditional classes, such as quizzes (both before and during class) and group activities.

Regarding the possibility of publication bias affecting our data, funnel plots for each dependent variable were examined for asymmetry, as presented in Figs. S1, S2, S3 (online only). In addition, we conducted three cumulative meta-analyses by precision, presented in Figs. S4, S5, S6 (online only). They showed no positive drift towards the end as would be the case in the presence of a strong publication bias (Borenstein et al., 2009). Lastly, we found no strong evidence for a negative correlation between sample size and effect size in the data from assessed learning outcome, $r = -0.15$, $p = .494$, perceived learning outcome, $r = -0.03$, $p = .945$, and satisfaction, $r = -0.10$, $p = .674$, as it would be the case when there is a bias against nonsignificant findings (Levine et al., 2009).

3.2. Effects of flipping the classroom on assessed and perceived learning outcomes and student satisfaction

In Table 3, the results of the average weighted Hedges' *g* for assessed learning outcomes, perceived learning outcomes, and satisfaction are presented, including the 95% prediction intervals, the *Q*-test for heterogeneity, the between-study variance, and the percentage of variation between studies that is due to heterogeneity rather than sampling error. The average effect size *for assessed learning outcomes* ($g = 0.36$) was found to be significant. Given the number of studies and the average within study sample size in those studies, power to detect what we assume to be the smallest important effect ($g = 0.20$) was very high (0.99). The average effect on *perceived learning outcomes* was nearly identical ($g = 0.36$), but not significant ($p = .13$). However, the observed power was low (0.44), as was the power (0.17) to detect the smallest important effect ($g = 0.20$). For *student satisfaction*, a trivial and non-significant effect size ($g = 0.05$) was found. However, the power to detect an effect of $g = 0.20$ was low (0.40). The *Q*-test was significant for all three dependent variables, the distribution of effect sizes is considered heterogeneous with $I^2 > 75\%$ indicating that a large proportion of the variability appears to be true variance. Forest plots of the dependent variables assessed and perceived learning outcomes and student satisfaction sorted on highest effect size can be consulted in Figs. S7, S8, and S9 (online only).

The average effect size does not seem to be heavily affected by outliers, as was visible in the one-study-removed analyses in R. The average effect of assessed learning outcomes without the most influential study was $g = 0.35$, $p < .001$, 95% CI [0.27, 0.43]

Table 2
Frequencies of a selection of the coded variables and categories.

Variable	Category	Total studies (*k*)
Academic domain	Humanities	9
	Social sciences	15
	Natural sciences	51
	Formal sciences	40
Allocation	Randomized on individual level	11
	Randomized pre-existing groups	28
	No randomization	76
Educational level	Secondary education	11
	Higher education	104
Face-to-face time	Equal in both conditions	97
	Reduced in the flipped classroom	14
	Increased in the flipped classroom	4
Intervention and control group equivalence assessment	Statistically demonstrated	54
	Descriptive statement	18
	Only assumed to be equal	31
	Groups were not equal	12
Intervention duration	1–10 weeks	25
	> 10 weeks	90
Outcome measurement	Standardized test	22
	Non-standardized test (e.g., teacher made)	93
Region	Asia	13
	North America	91
	Europe	6
	Middle East	1
	Oceania	1
	The Caribbean	3
Study type	Journal publication	86
	Conference paper	23
	Doctoral dissertation	5
	Master thesis	1
Teacher	Same for both conditions	60
	Different for both conditions	21
Type of control group	Simultaneous group	57
	Previous cohort	58
Video-based instructional activities before class	Yes	110
	No	5

Note: In total 114 studies were included, but one study reported two independent interventions. If the categories for one variable do not add up to the total of 115 included interventions, it means that for this number of studies no information about this variable was reported.

I^2 = 86% (Viechtbauer & Cheung, 2010).

3.3. Moderator analyses

In Tables 4 and 5, the results of the meta-regression moderator analyses are presented for the effect on *assessed learning outcomes* and *satisfaction*. For *perceived learning outcomes*, moderator analyses were not regarded as meaningful, as the total number of studies was eight and thus most moderator categories could not be meaningfully compared. We conducted meta-regression analyses for each group of moderator variables separately (i.e., design characteristics, educational context characteristics, and study quality characteristics). First, this was necessary to account for multicollinearity, as we saw that some moderator variables correlated with each other. Thus, running a moderator analysis with all moderators in one model would have distorted the results. Second, we consider this approach as the best possible balance between the risk of type I errors (univariate moderator analysis) and type II errors (all variables in one model).

For *assessed learning outcome*, design characteristics moderated the findings as the omnibus test of all regression coefficients was significant (p = .016) and accounted for 10% of the amount of heterogeneity. It appeared that studies that shortened the classroom time of the flipped condition had a significantly lower (p = .027) average effect than studies in which the classroom time in both conditions was equal (with a difference of $g = -0.26$, while accounting for all the other variables in the model). In addition, adding quizzes in the flipped condition also showed a significant (p = .044) difference with studies where quizzes were not added or already applied in the traditional condition (with a difference of g = 0.19, while accounting for all the other variables in the model). The omnibus tests of all the regression coefficients in the educational context characteristics model (p = .058) and the study quality characteristics (p = .161) were not significant. We decided to drop type of control group as study quality moderator from the model, as it highly correlated with allocation type and therefore distorted the findings ($r = -.72$). The interpretation of both moderators is fairly similar, as, for instance, a study with a previous cohort design is not able to randomly allocate students to conditions. According to the meta-regression power analysis simulation study by Hempel et al. (2013), the power of our *assessed learning outcomes*

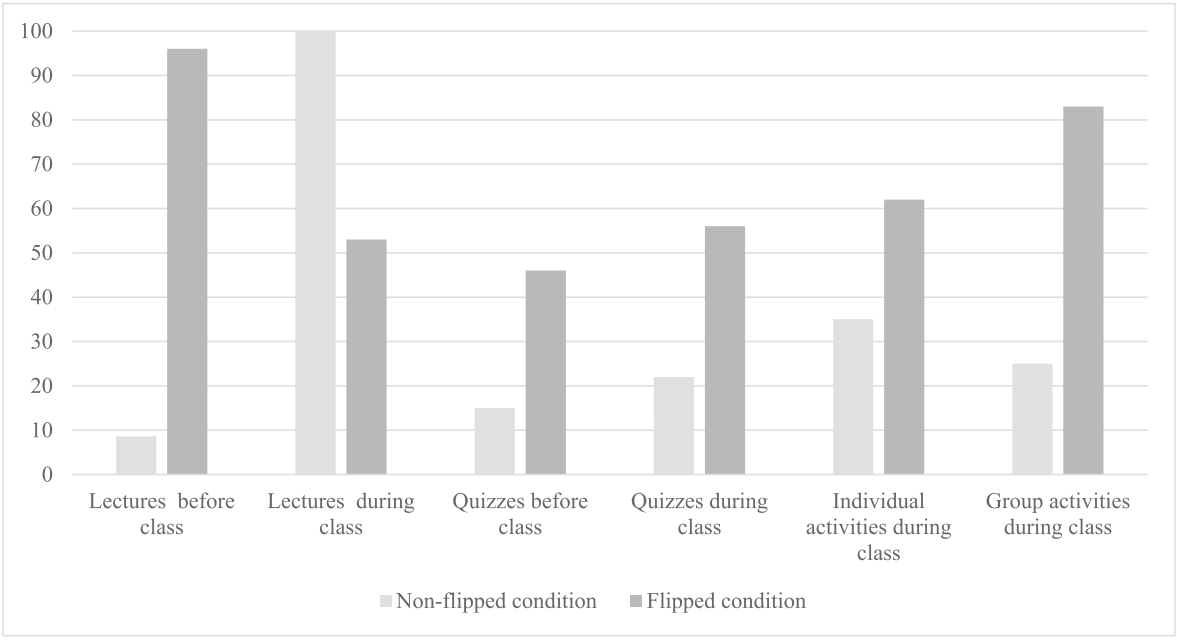

Fig. 2. Percentage frequency distribution of learning activities in both conditions.

Table 3
Results of the univariate random-effects meta-analyses.

Dependent variable	*k* (#students)	*g* (*p*)	*SE*	95% CI	*Q* (*p*)	df_q	τ^2 (*SE*)	I^2
Assessed learning outcomes	114 (20318)	0.36 (< .001)	0.04	[0.28, 0.44]	1221.86 (< .001)	113	0.14 (0.02)	88%
Perceived learning outcomes	8 (953)	0.36 (.13)	0.21	[-0.13, 0.85]	39.45 (< .001)	7	0.28 (0.18)	87%
Student satisfaction	22 (3501)	0.05 (.73)	0.13	[-0.23, 0.32]	181.99 (< .001)	21	0.33 (0.12)	92%

Note. *k* = number of studies; # students = total number of participants; *g* = mean weighted effect size in Hedges' *g*; *SE* = standard error; CI = confidence interval; *Q* = Cochran's heterogeneity test; df = degrees of freedom Q-test; τ^2 =between-study variance; I^2 = percentage of variation between studies that is due to heterogeneity rather than sampling error.

moderator models to detect a moderator effect of 0.2 was close to 80% (with 114 included studies, an average sample size of 180 students per study, and residual heterogeneity of 0.1).

For *satisfaction*, none of the moderator meta-regression models was significant. However, the power of these moderator models to detect a moderator effect of 0.2 was below 20% (with 22 included studies, an average sample size of 160 students per study, and residual heterogeneity of 0.3). Face-to-face ratio between conditions was dropped from the design characteristics model due to the unequal distribution of studies in the moderator categories. For the study quality model we also had to choose between the highly correlated variables allocation and type of control group. Here, we dropped allocation from the model, as it improved the R^2, and the distribution of studies in the categories was more evenly in the type of control group variable.

4. Discussion

4.1. Is FTC effective to enhance learning outcomes and student satisfaction?

The first research question we addressed in the present study was about the effects of FTC on assessed and perceived learning outcomes and student satisfaction. First, we found a significant small effect ($g > 0.2$; Cohen, 1988) of FTC on assessed learning outcomes. According to Hattie (2012), this effect is close to the hinge point (> 0.4) of average effects of educational interventions where teachers and researchers should strive for. However, while these two interpretative rules of thumb are often used to interpret effect sizes, what do they mean in terms of the underlying dependent variable assessed learning outcome? For instance, on a Math exam score with $M = 550$ and $SD = 100$, an effect size of 0.36 means that the average score of a student in the flipped classroom is 0.36 standard deviations above the average student in the traditional classroom (e.g., a score of 586 vs. 550). In other words, 64% of the students in the flipped classroom will be above the mean of the students in the traditional classroom. Thus, although the effect on assessed learning outcome may be regard as small, in the context of education it seems meaningful.

This is also true when we compare similar types of educational interventions in secondary education and other benchmarks. According to Hill, Bloom, Black, and Lipsey (2008) an effect of 0.36 is large if compared to the average annual gain in effect size

Table 4
Results of the moderator analyses of assessed learning outcomes.

Moderator Variable	*k*	Estimate (SE)	*p*	95% CI	*F (df1, df2)*[a]	*p*[b]	τ^2_{res} (*SE*)	I^2_{res}	R^2
Design characteristics					*F* (5, 108) = 2.93	.016	0.10 (0.02)	86%	.10
Intercept	114	0.302 (0.07)	< .001	[0.17, 0.44]					
Face-to-face time									
FC = TC *(ref)*	97								
FC < TC	14	−0.260 (0.12)	.027	[-0.49, −0.03]					
FC > TC	4	−0.237 (0.21)	.271	[-0.66, 0.18]					
Group assignments									
No addition in FC *(ref)*	72								
Addition in FC	42	0.088 (0.07)	.310	[-0.08, 0.26]					
Lectures during class									
No lectures in FC *(ref)*	53	0.028 (0.08)	.731	[-0.13, 0.19]					
Lectures in FC	61								
Quizzes									
No addition in FC *(ref)*	78	0.189 (0.09)	.044	[0.01, 0,37]					
Addition in FC	36								
Educational context characteristics					*F* (5, 108) = 2.22	.058	0.13 (0.02)	86%	.07
Intercept	114	0.309 (0.06)	< .001	[0.19, 0.42]					
Academic domain									
Natural sciences *(ref)*	51								
Humanities	9	0.162 (0.16)	.316	[-0.16, 0.48]					
Social sciences	15	0.283 (0.12)	.022	[0.04, 0.52]					
Formal sciences	39	−0.035 (0.09)	.698	[-0.22, 0.15]					
Educational level									
Higher education *(ref)*	103								
Secondary education	11	0.179 (0.14)	.215	[-0.11, 0.46]					
Intervention duration	114	−0.012 (0.01)	.142	[-0.03, 0.01]					
Study quality characteristics					*F* (8, 105) = 1.51	.161	0.13 (0.02)	86%	.05
Intercept	114	0.370 (0.07)	< .001	[0.23, 0.51]					
Allocation type									
Not random *(ref)*	75								
Pre-existing groups	28	0.198 (0.10)	.056	[-0.01, 0.40]					
Individual allocation	11	0.161 (0.14)	.258	[-0.12, 0.44]					
Group equivalence test									
Tested, equal *(ref)*	54								
Tested, not equal	12	−0.020 (0.13)	.884	[-0.29, 0.25]					
Not tested, descriptive statement	18	−0.225 (0.12)	.067	[-0.47, 0.02]					
Not tested, no descriptive statement	30	0.008 (0.13)	.938	[-0.20, 0.21]					
Study type									
Journal publication *(ref)*	86								
Conference	22	−0.179 (0.10)	.092	[-0.39, 0.03]					
Dissertation	5	−0.041 (0.21)	.849	[-0.46, 0.38]					
Thesis	1	0.388 (0.45)	.394	[-0.51, 1.29]					

Note. Ref = reference category; *k* = total number of studies, as well as effect sizes; *SE* = standard error; CI = confidence interval; τ^2_{res} = estimated amount of residual heterogeneity; I^2 = percentage of unaccounted variation between studies that is due to heterogeneity rather than sampling error; R^2 = proportion of amount of heterogeneity accounted for; FC = flipped classroom, TC = traditional classroom.

[a] Omnibus test of all regression coefficients of the moderators in the model.

[b] p-Value of the omnibus test.

($g < 0.24$) from nationally normed tests on reading and math in the higher grades of secondary school (i.e., grade > 9 where FTC is mostly applied), or if compared to the mean effect size from meta-analyses with comparable interventions in high school ($k = 28$, $d = 0.24$). For higher education, the meta-analyses by Schneider and Preckel (2017) provide an insight into comparable meta-analyses to gauge the relative size of effects. For example, on the basis of their ranking of 105 variables ordered by strength of their association with achievement, a 0.36 effect of flipping the classroom would be comparable to other interventions on the instruction variable technology such as Intelligent tutoring systems (0.35, rank 47) and blended learning (0.33, rank 52).

Our main results are comparable with previous (smaller scale) meta-analyses of FTC who found small average effect sizes ranged from 0.19 to 0.47 (Chen et al., 2018; Cheng et al., 2018; Hew & Lo, 2018; Lo et al., 2017). Together, these meta-analyses form a generalizable synthesis of the first generation of FTC research: the combined remarks and recommendations could positively affect future research. In a broader context, the results are also in line with a meta-analysis on blended learning by Spanjers et al. (2015), which found a comparable significant and small effect ($g = 0.34$) of blended learning on assessed learning outcomes. Even though FTC is a particular example of blended learning, our meta-analyses did not contain the same studies in their sample.

Second, we found a non-significant average effect of FTC on perceived learning outcomes. We chose to analyze assessed and perceived learning outcomes separately, as research has shown that learning outcomes determined by the subject itself are usually inaccurate (Kruger & Dunning, 1999). DeLozier and Rhodes (2017) showed that during initial FTC research mostly students

Table 5
Results of the moderator analyses of satisfaction.

Moderator Variable	*k*	Estimate (*SE*)	*p*	95% CI	*F (df1, df2)*[a]	*p*[b]	τ^2_{res} (*SE*)	I^2_{res}	R^2
Design characteristics					*F* (3, 18) = 1.76	.191	0.30 (0.12)	90%	.09
Intercept	22	−0.520 (0.30)	.099	[-1.15, 0.11]					
Group assignments									
No addition in FC (*ref*)	10								
Addition in FC	12	0.395 (0.31)	.221	[-0.26, 1.05]					
Lectures during class									
No lectures in FC (*ref*)	10								
Lectures in FC	12	0.554 (0.31)	.090	[-0.10, 1.20]					
Quizzes									
No addition in FC (*ref*)	15								
Addition in FC	7	0.156 (0.30)	.612	[-0.48, 0.79]					
Educational context characteristics					*F* (5, 16) = 0.36	.867	0.40 (0.15)	92%	.00
Intercept	22	−0.105 (0.25)	.681	[-0.64, 0.43]					
Academic domain									
Natural sciences *(ref)*	8								
Humanities	3	0.287 (0.47)	.548	[-0.70, 1.28]					
Social sciences	2	0.376 (0.56)	.515	[-0.82, 1.57]					
Formal sciences	9	0.086 (0.51)	.804	[-0.62, 0.78]					
Educational level									
Higher education *(ref)*	20								
Secondary education	2	0.086 (0.51)	.869	[-1.00, 1.17]					
Intervention duration	22	−0.039 (0.03)	.258	[-0.11, 0.03]					
Study quality characteristics					*F* (6, 15) = 1.45	.259	0.30 (0.13)	88%	.09
Intercept	22	0.552 (0.24)	.037	[0.04, 1.07]					
Control group type									
Previous cohort *(ref)*	12								
Simultaneous	10	−0.211 (0.28)	.458	[-0.80, 0.38]					
Group equivalence test									
Tested, equal *(ref)*	11								
Tested, not equal	2	−0.862 (0.45)	.075	[-1.82, 0.10]					
Not tested, descriptive statement	4	−0.539 (0.37)	.162	[-1.32, 0.24]					
Not tested, no descriptive statement	5	−0.770 (0.37)	.056	[-1.56, 0.02]					
Study type									
Journal publication *(ref)*	16								
Conference	4	−0.095 (0.40)	.816	[-0.95, 0.76]					
Dissertation	2	−0.475 (0.48)	.341	[-1.51, 0.56]					

Note. Ref = reference category; *k* = total number of studies, as well as effect sizes; *SE* = standard error; CI = confidence interval; τ^2_{res} = estimated amount of residual heterogeneity; I^2 = percentage of unaccounted variation between studies that is due to heterogeneity rather than sampling error; R^2 = proportion of amount of heterogeneity accounted for; FC = flipped classroom, TC = traditional classroom.

[a] Omnibus test of all regression coefficients of the moderators in the model.

[b] p-Value of the omnibus test.

perceptions of their own learning were measured, but we had to exclude many of these studies because they did not make use of a comparison group. It is possible that the effect is not significant due to the small sample size and therefore low power. However, it is notable that the average effect size is comparable to the effect on assessed learning outcome. More research should be conducted to be able to draw a more confident conclusion, although we regard assessed learning outcome as a better and more reliable measurement of learning outcome.

Third, we found a non-significant effect of FTC on student satisfaction, with an effect close to zero. We can conclude that on average, students are equally satisfied with flipped classrooms as traditional classrooms. These findings are in line with the meta-analysis of Spanjers et al. (2015) on blended learning, in which they found a non-significant trivial effect size ($g = 0.11$) for student satisfaction. They speculate that this might relate to the suggestion made by Sitzmann, Kraiger, Stewart, and Wisher (2006) that students perceive web-based instruction as more demanding than classroom instruction in terms of time commitment. Tune, Sturek, and Basile (2013) also report that an increase in workload was the most cited comment by students in their flipped classroom. Thus, in contrast with preliminary conclusions made in qualitative reviews on FTC, we conclude that there is no evidence that a flipped classroom leads to higher student satisfaction about the learning environment. As the variety between the studies was large and in some studies students in a flipped classroom were very positive and in other studies very negative in comparison with a traditional classroom, the design and educational context of a flipped classroom should be carefully planned.

4.2. Do instructional design, educational context, or study quality characteristics explain differential effects of FTC?

The second research question we addressed was about the possible moderator effects of instructional design, sample, and study quality characteristics on the effectiveness of FTC. For all meta-analyses, we found significant heterogeneity in effect sizes between studies, which is caused by more than random sampling alone. It means that in some studies, students in the flipped classroom

performed significantly worse than students in the traditional classroom (or did not perceive they learned better and vice versa). For student satisfaction, this means that in some studies students in the flipped classroom were more satisfied than students in the traditional classroom, and vice versa as the average is close to zero.

It should first be noted that we should be careful with the interpretation of the moderator analyses and not draw too major conclusions. The proportion of amount of heterogeneity accounted for in all the meta-regression models was not higher than 10%. In addition, while the meta-regression models on assessed learning outcomes were around 80%, the meta-regression models on student satisfaction had a very low power of 20%. Nevertheless, it is valuable to investigate patterns in the moderator analyses data which could contain indications of possible explanations for the large heterogeneity and provide helpful directions for further research.

We found that students in flipped classrooms in which the face-to-face time was not reduced performed significantly better compared to flipped classrooms that reduced classroom time. In most of the studies in which face-to-face classroom time in the flipped classroom was reduced, it was a deliberate decision (e.g., Baepler et al., 2014; Bossaer, Panus, Stewart, & Hagemeier, 2016; Gross, Pietri, Anderson, Moyano-Camihort, & Graham, 2015; Hibbard, Sung, & Wells, 2016; Koo et al., 2016). Our finding contradicts the conclusion of Baepler et al. (2014) that flipped classrooms with less face-to-face time are at least as effective as traditional classrooms. Our results show that sustaining face-to-face time is a critical feature for a successful implementation of a flipped classroom. This is especially important for policymakers and curriculum designers, as it is sometimes argued that a flipped classroom approach is implemented in order to reduce costs (i.e., reducing face-to-face time, see: Asef-Vaziri, 2015).

The addition of quizzes in a flipped classroom also seems to positively affect assessed learning outcomes. This is in line with research on the effectiveness of the testing effect (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007). It is also consistent with the study by Spanjers et al. (2015), which also found that quizzes were a significant moderator that affected the attractiveness and effectiveness of blended learning.

Other moderator effects were not found. Taking into account the low power for the moderator analysis (especially for student satisfaction), this should not be interpreted as strong evidence for the absence of moderator effects. For example, given the results of the omnibus tests in the educational context characteristics model ($p = .058$) model for assessed learning outcome, it might be justified to further explore patterns in the moderator analysis data such as the difference between studies from social sciences who had a higher average effect of 0.28 than studies from natural sciences ($p = .022$). This is in line with the FTC meta-analysis by Cheng et al. (2018), who found that subject area significantly moderated their results.

Still, some additional context can be helpful to interpret the non-significant results. Adding small group assignments was not a significant moderator, although we hypothesized that this could be the case. Research has shown that, for example, cooperative learning (i.e., an instructional technique in which students work together in small groups on a structured learning task, in which positive interdependence and individual accountability are important factors) has positive effects on achievement and attitudes (Kyndt et al., 2013; Slavin, 1991). An explanation for our finding might be that we were not able to verify if the group assignments included in our sample met the requirements of, for example, cooperative learning, due to the broad coding categories (e.g., quizzes, group assignments, lecture activities). Therefore, we recommend future research to describe their design of in classroom (group) activities in greater detail. However, group assignments did not negatively moderate the effect sizes, which means that including group assignments in an FTC implementation will not hamper learning outcomes either.

The same is true for lecture activities, as 53% of the flipped interventions still implemented some form of lecture activities in the classroom and our moderator analysis shows no negative result. In practice, adding microlectures can be used to address misunderstandings, for example on the basis of the outcomes of formative quizzes before class or at the beginning of a class (McLaughlin et al., 2013). Qualitative data from the study by Hagen and Fratta (2014) also showed that some students criticize the flipped classroom for a lack of lecture activities from their professor. This is in line with the conclusion by Walker et al. (2008) that students still value the integration of microlectures in a flipped classroom. The result of our moderator-analysis on student satisfaction seems to fit into this context. Although adding microlectures to the flipped classroom was not a significant moderator, we noticed a trend that student satisfaction was higher in flipped classrooms that included lecture activities, in comparison to flipped classrooms without lecture activities (with a difference of $g = 0.55$, $p = .090$). If we take in to account the low power of this particular meta-regression model, it is at least an interesting pattern which is worth to investigate in future research.

Lastly, the moderator variable educational level was unequally distributed in our data ($k = 11$ in secondary education and $k = 103$ higher education). It seems that, while FTC is also very widespread in secondary education (Bergmann & Sams, 2012), it is nececarry to conduct more research in this context with more rubust research designs (as may studies in secondary education were excluded because they did not make use of a control group). This is especially needed if one hypothesizes that FTC can work differently for students of different age groups (for example, because of differences in self-regulated learning abilities, see Wigfield et al., 2011; Zimmerman & Martinez-Pons, 1990). To evaluate if the applications of flipped classrooms in both contexts contained different design characteristics, we checked the distribution of the features we coded in both contexts. It appeared that quizzes, group assignments, and lecture activities in the flipped classroom evenly occurred.

4.3. Other potential sources of variation in effects of FTC

Although the moderator analyses are a valuable start to explore possible patterns in the heterogeneity of the included studies, there are other important potential sources of variation in effects of FTC to consider. Previous research has addressed the issue of a digital divide. This means that there are differences amongst subgroups such as gender, ethnicity, and socioeconomic background and how they have access to technology and how these differences might have an impact on a student's ability to benefit in terms of learning with technology (Jackson et al., 2008). In the context of flipped classrooms, where students are usually required to watch

online videos in their own time, it is necessary to evaluate the possible consequences of a digital divide in all its manifestations.

It seems that there is no direct evidence for strong differential effects due to the digital divide in our data, as most studies report that there are no differences in the effect of FTC on learning outcomes for gender (Adams, Garcia, & Traustadóttir, 2016; Scott et al., 2016; Wasserman, Quint, Norris, & Carr, 2015), no differences for gender and race (Lancellotti, Thomas, & Kohli, 2016; Olitsky & Cosgrove, 2016), and no differences for gender, race, and socioeconomic status (Calimeris & Sauer, 2015). There was one study in which more women reporting they watched instruction videos before class than men (Ossman & Warren, 2014).

Another potential source of variation in effects of FTC is a student's self-regulated learning (SRL) capability. We noticed that only a few of the included studies refer to SRL as an important variable in understanding the working principle of FTC and even fewer studies measure SRL as a variable. Lape et al. (2014) conclude that there were no significant metacognitive gains visible after the intervention for students both in the flipped and control condition. This question presumes that students working in a flipped classroom will improve their SRL capabilities more than students in a traditional classroom (mediation effect). Evidence for this is found in a meta-analytical FTC study within nursing education which included 29 Chinese studies, which found that self-learning abilities improved significantly better by students in flipped classrooms compared with traditional classrooms (Tan, Yue, & Fu, 2017). However, it can also be hypothesized that SRL is an important moderator for the success of a flipped classroom, as students with better SRL capabilities can benefit more from a flipped classroom, which would presumably lead to higher learning outcomes (Lim & Morris, 2009). Then, teachers and researchers should consider to what extent their flipped classroom should include SRL support for students, as there is evidence that flipped classrooms with SRL support achieve higher learning outcomes in comparison with flipped classrooms without SRL support (Moos & Bonde, 2016; Shyr & Chen, 2018).

The same is true for motivation, as the effect of FTC can be dependent on how well students are motivated to study in advance. For instance, students' motivation in a flipped classroom could be positively or negatively affected by their preference for this method (e.g., Asiksoy & Özdamli, 2016; Hibbard et al., 2016). If a student is not motivated to complete preparatory activities before class, this could lead to negative effects, as that student will probably experience difficulty applying learning materials during class. Therefore, it seems important to take students' motivation into account when designing a flipped classroom. For example, instructing students about the benefits of flipping the classroom could increase motivation to study before class, while a too precise in-class repetition of content from the instructional video could hamper the need for students to study before class. Unfortunately, we could not investigate motivation as moderator variable as few studies measured students' motivation.

4.4. Limitations

The following limitations in our review should be kept in mind when interpreting our results. We compared flipped classrooms with traditional classrooms. The frequency distribution of learning activities in Fig. 2 shows that flipped classrooms are comparable with each other, as well as traditional classrooms. While we were only able to compare these two main categories, we still acknowledge that flipped and traditional classrooms can differ in practice. We recommend future research to explicitly describe both research condition in more detail. In addition, there are a few variables which contain important information about variation in effects but are not systematically reported in current studies on FTC, such as age, gender, class size, and experience with a flipped classroom of both teachers and students. Some variables that we coded could not be used in the moderator analyses, because one category was heavily underrepresented in the data (e.g., 109 included studies did make use of online videos and only five studies did not). For example, the effect of the medium of instruction before class in the flipped condition could not be meaningfully compared, as 109 included studies did make use of online videos and only five studies did not. Lastly, most studies were conducted in the region North America (mainly the United States) and in the context of post-secondary education educational. This means that these two contexts are overrepresented in current research on FTC and in our data.

The information about costs of a flipped classroom implementation was also underrepresented in our data. Although we argue that a flipped classroom is not inherently dependent on the use of technology, we see that in most implementations several high- and low-cost technologies are used. For instance to record instruction video's, to design online learning environments where video's and additional learning assignments are distributed, and sometimes technologies to apply the learning material in class. Besides technology costs, implementing a flipped classroom also means an initial huge time investment (e.g., Talbert, 2015), even if one decide to find free to use online videos. One might wonder if the costs of implementing FTC is justified given the results we presented in this meta-analysis. We are, however, unable to answer this question, as this is heavily context dependent and the included studies often did not report details about costs. Nevertheless, we would like to point out some interesting studies that might be helpful for teachers and policy-makers. Touchton (2015) showed that initial implementation costs can be high, but that this investment can pay off on the long term. Peterson (2015), Hudson et al. (2015), and O'Flaherty and Phillips (2015) note that IT support can be critical for a successful implementation of FTC. Asef-Vaziri (2015) shows that some institutions are interested in FTC in order to reduce costs, when a significant portion of face-to-face classroom time is replaced by online learning. This last idea should be explored with caution, since our moderator analysis shows that sustaining face-to-face time leads to significant higher effects on assessed learning outcome. Still, some included studies report a successful cost reduction by implementing a flipped classroom while at the same time student learning outcomes were at least equal to the previous traditional classroom (Hudson et al., 2015; Olitsky & Cosgrove, 2016).

Besides, there are a few biases related to the average duration of the interventions. First, it usually was the first experience with FTC for teachers. Although our results show that these first implementations of FTC were on average enhancing assessed learning outcomes, we think that such big changes in a curriculum become even more effective after (minor) adjustments based on evaluation of the first implementation. For instance, the study by Hudson et al. (2015) showed that student satisfaction in their flipped classroom only increased after several semesters of redesigning the flipped course. In addition, the behavior of teachers during class changes

with a flipped approach, but research has shown that, in general, training and time is needed to achieve effective change in teachers' behavior (Niemi, 2002; Van den Bergh et al., 2014). Second, it seems that students in the flipped classroom seem to need an adaption time to get used to it that can take several weeks to a full semester (Hotle & Garrow, 2015). This illustrates a possible bias in our dataset, because we were mainly able to examine the effects of first-time flipped classroom interventions, which could become more effective in the long-term after refinements and redesign.

Another potential bias is a combination of effects that are almost inherent to educational research. Some students may have been aware they are participating in a research condition which may affect their expectations and behavior (Hawthorne-effect), and some students in the flipped classroom could have shown better performance than students in a traditional classroom from the expectation of teachers that this innovation would lead to better learning outcomes alone (Pygmalion-effect). It turned out that 60 studies in our sample had the same teacher for both conditions (in which a teacher's preference and experience could be a bias) and 21 studies had different teachers for both condition (which is a bias on itself). Therefore, we cannot exclude the influence of these effects nor control for them in our meta-analysis. However, there are reasons to assume that the impact of these underlying effects is not high. First, it is clear that not all students in the flipped classrooms were satisfied with their learning environment in comparison to students in a traditional classroom: the effect size for satisfaction was $g = 0.05$, and heterogeneity was large and significant. In addition, Fig. S9 (online only) shows that 12 out of 22 studies reported a negative or neutral effect size. Second, we separately analyzed assessed and perceived learning outcomes to establish a possible difference between what students think they have learned, and what was measured in achievement tests. Still, we recommend future research into the effect of FTC in comparison with non-flipped classrooms to better account for the possibility of a Hawthorne-effect, by, for example, use a between and within subjects research design that switches both methods (preferably also several times) between independent groups of students.

In addition, we had to exclude 54 studies because they did not report enough statistical information to calculate effect sizes. If we compare the results from the meta-analyses with the results of studies that were excluded because they lacked sufficient statistical information as qualitatively analyzed and shown in Table S1 (online only), it seems that the reason they contain non-significant results does not occur more often than that the authors report a positive finding. Because of the largely equal distribution of positive and neutral reported outcomes, we consider that excluding these studies did not have a great impact on our general findings. However, the statistical power of the meta-analyses as well as the moderator analyses could therefore have been reduced.

The influence of publication bias should be never neglected in any meta-analysis. However, results of the tests we performed to estimate the effect of publication bias seem to indicate that it did not have a major impact on our results (e.g., funnel plots, cumulative analyses, correlation test between effect size and sample size). In addition, we also included non-published articles, and we did not find a moderator effect for publication type.

4.5. Conclusion

The present study quantitatively synthesized the results of 114 studies about the effects of FTC on learning outcomes and student satisfaction. With respect to the issue of defining FTC, we found that 95% of the included flipped interventions used video based instruction activities. Therefore, we continue to follow Abeysekera and Dawson (2015) in their recommendation to adopt a high-level uniform and broad definition of FTC, provided that researchers are very specific about the operationalization of the definition.

In general, we can conclude that students in flipped classrooms achieve significantly higher assessed learning outcomes than students in traditional classrooms, and are equally satisfied with the learning environment. The main implication following our results is that flipped classrooms are worth implementing. 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. The significant heterogeneity in the effect sizes of the studies means that it matters how flipped classrooms are implemented. Our study provides a few general directions to effective design of a flipped classroom, as sustaining face-to-face time and adding quizzes are critical features for a successful implementation of a flipped classroom.

Acknowledgments

This work was supported by funding from the Dutch Ministry of Education, Culture and Science, (grant number OCW/PromoDoc/1065001). The authors would like to thank E. Papangeli, and J. Teunissen for their help with coding. Declarations of interest: none.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2019.05.003.

References

References marked with an asterisk indicate studies included in the meta-analyses. A complete reference list of studies included in the meta-analyses is available as appendix B (online only).

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. *Higher Education Research and Development, 34*(1), 1–14. https://doi.org/10.1080/07294360.2014.934336.

(*)Adams, A. E. M., Garcia, J., & Traustadóttir, T. (2016). A quasi experiment to determine the effectiveness of a "partially flipped" versus "fully flipped" undergraduate class in genetics and evolution. *CBE-life Sciences Education, 15*(2), 1–9. https://doi.org/10.1187/cbe.15-07-0157.

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education, 126*, 334–345. https://doi.org/10.1016/j.

compedu.2018.07.021.
(*)Asef-Vaziri, A. (2015). The flipped classroom of operations Management : A not-for-cost-reduction platform. *Decision Sciences Journal of Innovative Education, 13*(1), 71–89. https://doi.org/10.1111/dsji.12054.
(*)Asiksoy, G., & Özdamli, F. (2016). Flipped classroom adapted to the ARCS model of motivation and applied to a physics course. *Eurasia Journal of Mathematics, Science and Technology Education, 12*, 1589–1603. https://doi.org/10.12973/eurasia.2016.1251a.
(*)Baepler, P., Walker, J. D., & Driessen, M. (2014). It's not about seat time: Blending, flipping, and efficiency in active learning classrooms. *Computers & Education, 78*, 227–236. https://doi.org/10.1016/j.compedu.2014.06.006.
Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research, 85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818.
Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis, 34*, 259–277. https://doi.org/10.3102/0162373712446144.
Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day.* Alexandria, Va: International Society for Technology in Education.
Bernard, J. S. (2015). The flipped classroom: Fertile ground for nursing education research. *International Journal of Nursing Education Scholarship, 12*(1), 1–11. https://doi.org/10.1515/ijnes-2015-0005.
Betihavas, V., Bridgman, H., Kornhaber, R., & Cross, M. (2016). The evidence for "flipping out": A systematic review of the flipped classroom in nursing education. *Nurse Education Today, 38*, 15–21. https://doi.org/10.1016/j.nedt.2015.12.010.
Bishop, J. L., & Verleger, M. (2013). The flipped classroom: A survey of the research. *Proccedings of the Annual Conference of the American Society for Engineering Education, 30*(9), 1–18.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, UK: John Wiley & Sons, Ltd. http://doi.org/10.1002/9780470743386.
(*)Bossaer, J. B., Panus, P., Stewart, D. W., & Hagemeier, N. E. (2016). Student performance in a pharmacotherapy oncology module before and after flipping the classroom. *American Journal of Pharmaceutical Education, 80*(2), 3–8. https://doi.org/10.5688/ajpe80231.
(*)Calimeris, L., & Sauer, K. M. (2015). Flipping out about the flip: All hype or is there hope? *International Review of Economics Education, 20*, 13–28. https://doi.org/10.1016/j.iree.2015.08.001.
Chandra, V., & Lloyd, M. (2008). The methodological nettle: ICT and student achievement. *British Journal of Educational Technology, 39*(6), 1087–1098. https://doi.org/10.1111/j.1467-8535.2007.00790.x.
Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2018). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research & Development,* 1–32. https://doi.org/10.1007/s11423-018-9633-7.
Chen, K. S., Monrouxe, L., Lu, Y. H., Jenq, C. C., Chang, Y. J., Chang, Y. C., et al. (2018). Academic outcomes of flipped classroom learning: A meta-analysis. *Medical Education, 52*(9), 910–924. https://doi.org/10.1111/medu.13616.
Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292. https://doi.org/10.3102/0013189X16656615.
Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823.
Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research & Development, 42*(2), 21–29. https://doi.org/10.1007/BF02299088.
Clark, R. C., Nguyen, F., & Sweller, J. (2005). *Efficiency in learning: Evidence-based guidelines to manage cognitive load.* (San Francisco, CA: Pfeiffer).
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Routledge.
DeLozier, S. J., & Rhodes, M. G. (2017). Flipped classrooms: A review of key ideas and recommendations for practice. *Educational Psychology Review, 29*(1), 141–151. http://dx.doi.org/10.1007/s10648-015-9356-9.
Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research, 107*(5), 357–364. https://doi.org/10.1080/00220671.2013.823370.
(*)Eichler, J. F., & Peeples, J. (2016). Flipped classroom modules for large enrollment general chemistry courses: A low barrier approach to increase active learning and improve student grades. *Chemistry Education: Research and Practice, 17*, 197–208. https://doi.org/10.1039/C5RP00159E.
Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L.
Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions. In C. J. Bonk, & C. R. Graham (Eds.). *The handbook of blended learning: Global perspectives, local designs* (pp. 3–21). (Pfeiffer San Francisco, CA).
(*)Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., & Graham, M. J. (2015). Increased preclass preparation underlies student outcome improvement in the flipped classroom. *CBE-Life Sciences Education, 14*(4), 1–8. https://doi.org/10.1187/cbe.15-02-0040.
(*)Hagen, E. J., & Fratta, D. (2014). Hybrid learning in geological engineering: Why, how, and to what end? Preliminary results. *Geo-congress technical papers: Geo-characterization and modeling for sustainability, USA* (pp. 3920–3929). . https://doi.org/10.1061/9780784413272.380.
Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning.* London and New York: Routledge.
Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107.
Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445. http://dx.doi.org/10.1037/1082-989X.9.4.426.
Hempel, S., Miles, J. N., Booth, M. J., Wang, Z., Morton, S. C., & Shekelle, P. G. (2013). Risk of bias: A simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic Reviews, 2*(1), 107. https://doi.org/10.1186/2046-4053-2-107.
Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: A meta-analysis. *BMC Medical Education, 18*(1), 1–12. https://doi.org/10.1186/s12909-018-1144-z.
Heyma, A., Bisschop, P., van den Berg, E., Wartenbergh-Cras, F., Kurver, B., Muskens, M., et al. (2015). *Effectmeting InnovatieImpuls onderwijs: Eindrapport.* (SEO-rapport No. 2015-28). Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2015/06/29/effectmeting-innovatieimpuls-onderwijs.
(*)Hibbard, L., Sung, S., & Wells, B. (2016). Examining the effectiveness of a semi self paced format in a college general chemistry sequence. *Journal of Chemical Education, 93*(1), 24–30. https://doi.org/10.1021/acsjcherned.5b00592.
Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x.
(*)Hotle, S. L., & Garrow, L. A. (2015). Effects of the traditional and flipped classrooms on undergraduate student opinions and success. *Journal of Professional Issues in Engineering Education and Practice, 142*(1), 5015005. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000259.
(*)Hudson, D. L., Whisenhunt, B. L., Shoptaugh, C. F., Visio, M. E., Cathey, C., & Rost, A. D. (2015). Change takes time: Understanding and responding to culture change in course redesign. *Scholarship of Teaching and Learning in Psychology, 1*(4), 255–268. https://doi.org/10.1037/stl0000043.
Jackson, L. A., Zhao, Y., Kolenic, A., III, Fitzgerald, H. E., Harold, R., & Von Eye, A. (2008). Race, gender, and information technology use: The new digital divide. *CyberPsychology and Behavior, 11*(4), 437–442. https://doi.org/10.1089/cpb.2007.0157.
Kim, M. K., Kim, S. M., Khera, O., & Getman, J. (2014). The experience of three flipped classrooms in an urban university: An exploration of design principles. *Internet and Higher Education, 22*, 37–50. https://doi.org/10.1016/j.iheduc.2014.04.003.
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1.
Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*, 2693–2710. https://doi.org/10.1002/sim.1482.
(*) Koo, C. L., Demps, E. L., Farris, C., Bowman, J. D., Panahi, L., & Boyle, P. (2016). Impact of flipped classroom design on student performance and perceptions in a

pharmacotherapy course. *American Journal of Pharmaceutical Education, 80*(2), 1–9. (No. 33) https://doi.org/10.5688/ajpe80233.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.
Kyndt, E., Raes, E., Lismont, B., Timmers, F., Cascallar, E., & Dochy, F. (2013). A meta-analysis of the effects of face-to-face cooperative learning. Do recent studies falsify or verify earlier findings? *Educational Research Review, 10*, 133–149. https://doi.org/10.1016/j.edurev.2013.02.002.
Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education, 31*(1), 30–43. https://doi.org/10.1080/00220480009596759.
Lai, C.-L., & Hwang, G.-J. (2016). A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Computers & Education, 100*, 126–140. https://doi.org/10.1016/j.compedu.2016.05.006.
(*)Lancellotti, M., Thomas, S., & Kohli, C. (2016). Online video modules for improvement in student learning. *The Journal of Education for Business, 91*(1), 19–22. https://doi.org/10.1080/08832323.2015.1108281.
(*)Lape, N. K., Levy, R., Yong, D. H., Haushalter, K. A., Eddy, R., & Hankel, N. (2014). Probing the inverted classroom: A controlled study of teaching and learning outcomes in undergraduate engineering and mathematics. *ASEE Annual Conference and Exposition, Conference Proceedings, USA, 121*, 9475. https://www.asee.org/public/conferences/32/papers/9475/download.
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.
Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs, 76*(3), 286–302. https://doi.org/10.1080/03637750903074685.
Lim, D. H., & Morris, M. L. (2009). Learner and instructional factors influencing learning outcomes with online learning environment. *Educational Technology & Society, 12*(4), 282–293.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage Publications.
Lo, C. K., & Hew, K. F. (2017). A critical review of flipped classroom challenges in K-12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning, 12*(4), 1–22. https://doi.org/10.1186/s41039-016-0044-2.
Lo, C. K., Hew, K. F., & Chen, G. (2017). Toward a set of design principles for mathematics flipped classrooms: A synthesis of research in mathematics education. *Educational Research Review, 22*, 50–73. https://doi.org/10.1016/j.edurev.2017.08.002.
López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology, 67*(1), 30–48. https://doi.org/10.1111/bmsp.12002.
Lundin, M., Bergviken Rensfeldt, A., Hillman, T., Lantz-Andersson, A., & Peterson, L. (2018). Higher education dominance and siloed knowledge: A systematic review of flipped classroom research. *International Journal of Educational Technology in Higher Education, 15*(20), 1–30. https://doi.org/10.1186/s41239-018-0101-6.
McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154.
(*) McLaughlin, J. E., Griffin, L. M., Esserman, D. A., Davidson, C. A., Glatt, D. M., Roth, M. T., et al. (2013). Pharmacy student engagement, performance, and perception in a flipped satellite classroom. *American Journal of Pharmaceutical Education, 77*(9), 1–8. (No. 196) https://doi.org/10.5688/ajpe779196.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), 1–6. https://doi.org/10.1371/journal.pmed.1000097.
Moos, D. C., & Bonde, C. (2016). Flipping the classroom: Embedding self-regulated learning prompts in videos. *Technology, Knowledge and Learning, 21*(2), 225–242. https://doi.org/10.1007/s10758-015-9269-1.
Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059.
(*)Mullen, J. S., & Sullivan, J. M., Jr. (2015). Student-perceived effectiveness of online content delivery modes. *Proceedings of frontiers in education conference, USA,1906–1909*https://doi.org/10.1109/FIE.2015.7344335.
Niemi, H. (2002). Active learning—a cultural change needed in teacher education and schools. *Teaching and Teacher Education, 18*(7), 763–780. https://doi.org/10.1016/S0742-051X(02)00042-2.
(*)Olitsky, N. H., & Cosgrove, S. B. (2016). The better blend? Flipping the principles of microeconomics classroom. *International Review of Economics Education, 21*, 1–11. https://doi.org/10.1016/j.iree.2015.10.004.
(*) Ossman, K. A., & Warren, G. (2014). Effect of flipping the classroom on student performance in first year engi- neering courses effect of flipping the classroom on student performance in first-year engineering courses. *ASEE annual conference and exposition, conference proceedings, USA: Vol. 121*, (pp. 8754–). . Retrieved from https://peer.asee.org/20342.
O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education, 25*, 85–95. https://doi.org/10.1016/j.iheduc.2015.02.002.
(*)Peterson, D. J. (2015). The flipped classroom improves student achievement and course satisfaction in a statistics course: A quasi-experimental study. *Teaching of Psychology, 43*(1), 10–15. https://doi.org/10.1177/0098628315620063.
Polanin, J. R., & Pigott, T. D. (2015). The use of meta-analytic statistical significance testing. *Research Synthesis Methods, 6*(1), 63–73. https://doi.org/10.1002/jrsm.1124.
Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*(1), 207–236. https://doi.org/10.3102/0034654315582067.
R Development Core Team (2008). *R: A language and environment for statistical computing.* Vienna, Austria http://www.r-project.org.
(*)Reza, S., & Ijaz Baig, M. (2015). A study of inverted classroom pedagogy in computer science teaching. *International Journal of Research Studies in Educational Technology, 4*(2), 19–30. https://doi.org/10.5861/ijrset.2015.1091.
Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x.
Roehl, A., Reddy, S. L., & Shannon, G. J. (2013). The flipped classroom: An opportunity to engage millennial students through active learning strategies. *Journal of Family and Consumer Sciences, 105*(2), 44–49.
Rosenthal, R. (1991). *Meta-analytic procedures for social research, Vol. 6.* Thousand Oaks, CA: Sage Publications.
Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020.
Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., et al. (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. *Computers & Education, 72*, 271–291. https://doi.org/10.1016/j.compedu.2013.11.002.
Schmidt, H. G., Wagener, S. L., Smeets, G. A. C. M., Keemink, L. M., & Van Der Molen, H. T. (2015). On the use and misuse of lectures in higher education. *Health Professions Education, 1*(1), 12–18. https://doi.org/10.1016/j.hpe.2015.11.010.
Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565. https://doi.org/10.1037/bul0000098.
(*)Scott, C. E., Green, L. E., & Etheridge, D. L. (2016). A comparison between flipped and lecture-based instruction in the calculus classroom. *Journal of Applied Research in Higher Education, 8*(2), 252–264. https://doi.org/10.1108/JARHE-04-2015-0024.
Seery, M. K. (2015). Flipped learning in higher education chemistry: Emerging trends and potential directions. *Chemistry Education: Research and Practice, 16*(4), 758–768. https://doi.org/10.1039/C5RP00136F.
Shyr, W. J., & Chen, C. H. (2018). Designing a technology-enhanced flipped learning system to facilitate students' self-regulation and performance. *Journal of Computer Assisted Learning, 34*(1), 53–62. https://doi.org/10.1111/jcal.12213.
Sitzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction. *Personnel Psychology, 59*(3),

623–664. https://doi.org/10.1111/j.1744-6570.2006.00049.x.
Slavin, R. E. (1991). Synthesis of research of cooperative learning. *Educational Leadership, 48*(5), 71–82.
Spanjers, I. A. E., Könings, K. D., Leppink, J., Verstegen, D. M. L., de Jong, N., Czabanowska, K., et al. (2015). The promised land of blended learning: Quizzes as a moderator. *Educational Research Review, 15*, 59–74. https://doi.org/10.1016/j.edurev.2015.05.001.
Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 15*(2), 171–193. https://doi.org/10.1007/s10984-012-9108-4.
Talbert, R. (2015). Inverting the transition-to-proof classroom. *Primus, 25*(8), 614–626. https://doi.org/10.1080/10511970.2015.1050616.
Tan, C., Yue, W.-G., & Fu, Y. (2017). Effectiveness of flipped classrooms in nursing education: Systematic review and meta-analysis. *Chinese Nursing Research, 4*(4), 192–200. https://doi.org/10.1016/J.CNRE.2017.10.006.
Torgerson, C. J. (2006). Publication bias: The achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*(1), 89–102. https://doi.org/10.1111/j.1467-8527.2006.00332.x.
(*)Touchton, M. (2015). Flipping the classroom and student performance in advanced statistics: Evidence from a quasi-experiment. *Journal of Political Science Education, 11*(1), 28–44. https://doi.org/10.1080/15512169.2014.985105.
(*)Tune, J. D., Sturek, M., & Basile, D. P. (2013). Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. *Advances in Physiology Education, 37*(4), 316–320. https://doi.org/10.1152/advan.00091.2013.
Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis: Erratum. *Journal of Educational and Behavioral Statistics, 35*(3), 375. http://dx.doi.org/10.3102/1076998610376621.
Van den Bergh, L., Ros, A., & Beijaard, D. (2014). Improving teacher feedback during active learning: Effects of a professional development program. *American Educational Research Journal, 51*(4), 772–809. https://doi.org/10.3102/0002831214531322.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03.
Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11.
Walker, J. D., Cotner, S. H., Baepler, P. M., & Decker, M. D. (2008). A delicate balance: Integrating active learning into a large lecture course. *CBE-Life Sciences Education, 7*, 361–367. https://doi.org/10.1187/cbe.08-02-0004.
(*)Wasserman, N. H., Quint, C., Norris, S. A., & Carr, T. (2015). Exploring flipped classroom instruction in calculus III. *International Journal of Science and Mathematics Education, 15*, 545–568. https://doi.org/10.1007/s10763-015-9704-8.
Wigfield, A., Klauda, S. L., & Cambria, J. (2011). Influences on the development of academic self-regulatory processes. In D. H. S. Barry, & J. Zimmerman (Eds.). *Handbook of self-regulation of learning and performance* (pp. 33–48). New York: Routledge. https://doi.org/10.4324/9780203839010.ch3.
Wittrock, M. C. (1992). Generative learning processes of the brain. *Educational Psychologist, 27*, 531–541. https://doi.org/10.1207/s15326985ep2704_8.
Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*(1), 51–59. https://doi.org/10.1037/0022-0663.82.1.51.