Check for updates

AERA Open
July-September 2019, Vol. 5, No. 3, pp. 1–17
DOI: 10.1177/2332858419870489
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2019. http://journals.sagepub.com/home/ero

Does the Flipped Classroom Improve Student Learning and Satisfaction? A Systematic Review and Meta-Analysis

Torstein Låg
Rannveig Grøm Sæle
UiT The Arctic University of Norway

We searched and meta-analyzed studies comparing flipped classroom teaching with traditional, lecture-based teaching to evaluate the evidence for the flipped classroom's influence on continuous-learning measures, pass/fail rates, and student evaluations of teaching. Eight electronic reference databases were searched to retrieve relevant studies. Our results indicate a small effect in favor of the flipped classroom on learning (Hedges' g = 0.35, 95% confidence interval [CI] [0.31, 0.40], k = 272). However, analyses restricted to studies with sufficient power resulted in an estimate of 0.24 (95% CI [0.18, 0.31], k = 90). Effects on pass rates (odds ratio = 1.55, 95% CI [1.34, 1.78], k = 45) and student satisfaction (Hedges' g = 0.16, 95% CI [0.06, 0.26], k = 69) were small and also likely influenced by publication bias. There is some support for the notion that the positive impact on learning may increase slightly if testing student preparation is part of the implementation.

Keywords: *flipped classroom, active learning, achievement, classroom research, learning environments, systematic review, meta-analysis, publication bias*

The flipped classroom teaching model is colloquially defined as one in which the activities traditionally done by students outside class (e.g., practicing problem solving) are moved into the classroom session, whereas what is traditionally done in class (e.g., expository, information transmission teaching) is done outside and prior to class. Is the flipped teaching model more effective than traditional classroom teaching approaches? Is its popularity warranted? A number of authors have noted the lack of rigorous evaluations of the flipped classroom model (e.g., Abeysekera & Dawson, 2015; Bishop & Verleger, 2013; Goodwin & Miller, 2013; Hung, 2015; Love, Hodge, Grandgenett, & Swift, 2014; O'Flaherty & Phillips, 2015). And although there have been recent efforts to summarize the extant evidence, they have typically been somewhat limited in scope (K. S. Chen et al., 2018; Cheng, Ritzhaupt, & Antonenko, 2018; Gillette et al., 2018; Hew & Lo, 2018; Hu et al., 2018; Tan, Yue, & Fu, 2017) and/or without meta-analyses (Betihavas, Bridgman, Kornhaber, & Cross, 2016; Evans, Vanden Bosch, Harrington, Schoofs, & Coviak, 2019; Karabulut-Ilgu, Jaramillo Cherrez, & Jahren, 2018; Lundin, Bergviken Rensfeldt, Hillman, Lantz-Andersson, & Peterson, 2018; O'Flaherty & Phillips, 2015).

For this review, we performed a thorough, systematic search of the literature on the flipped classroom model, spanning all levels of education, all disciplines, and all types of research reports, and systematized and meta-analyzed the findings. The results should be of interest to teachers and educators who wish to base their decisions regarding choice of teaching interventions on accumulated evidence, as well as for education researchers.

Defining the Flipped Classroom

Definitions of the flipped classroom in the literature vary. Some emphasize the utilization of digital technologies (e.g., Bishop, 2014), some the social or interactive nature of the in-class activities (e.g., Abeysekera & Dawson, 2015), and some the importance of using a particular pedagogical approach, such as mastery learning (e.g., Bergmann & Sams, 2012) or collaborative learning (e.g., Foldnes, 2016). Despite the substantial variation, it is possible to distil some common, core features that are present in virtually all the definitions. For the purposes of this review, and based on our perusal of the definitions in the literature, we propose the following working definition:

The flipped classroom is a teaching model that moves most of the teacher-centered instruction out of the classroom to free up time in the classroom for more student-centered learning activities.

This definition is based on no particular pedagogical approach or ideology, apart from the flipping itself, and it prescribes no particular kinds of instruction or classroom activities. As such, it is even more neutral than the one proposed by Abeysekera and Dawson (2015), who additionally require that classroom activities are social and that mandatory pre- or postclass activities are included. Although

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

mandatory preparatory activities, for instance a quiz, may be a good idea to ensure that students are in fact prepared for classroom activities, and although there is evidence suggesting that social learning activities are effective (Burgess, McGregor, & Mellis, 2014; Johnson, Johnson, & Smith, 2014; Pai, Sears, & Maeda, 2015; Tomcho & Foels, 2012), there are several examples of teaching setups that do not fulfil these two criteria but nevertheless are considered instances of flipping by their originators, and that would satisfy our more general definition. And whereas Abeysekera and Dawson's (2015) mission is partly to prescribe, and hence influence the course of future research on the flipped classroom model, our own aim is primarily descriptive; we hope to provide a summary of the effect of the flipped classroom model in general, as compared with traditional teaching models.

As opposed to the active learning–based flipped classroom, traditional teaching often implies a more passive role for the student (Prince, 2004). For the purposes of this meta-analysis, we define traditional teaching as predominantly teacher-centered, lecture-based information transfer.

Why Flipping Should Work

Proponents of the flipped classroom model typically argue that it has a number of advantages over traditional teaching models. These include more personalized teaching and learning (Bergmann & Sams, 2012; O'Flaherty & Phillips, 2015), better use of class time and flexible technology (Herreid, Schiller, Herreid, & Wright, 2014), and allowing students to take more responsibility for their own learning (O'Flaherty & Phillips, 2015). Although these certainly seem like advantages, there is not yet conclusive evidence linking them to any effects of flipping the classroom on student learning or satisfaction. However, there appears to be sound logic behind the idea of making time for more learning *activity*, because evidence is accumulating that teaching for active learning leads to better student performance and lower failure rates than lecture-based teaching (Freeman, Eddy, McDonough, et al., 2014; Michael, 2006; Prince, 2004). Hence, if flipping the classroom leads students to "do meaningful learning activities and to think about what they are doing" (Prince, 2004, p. 223) to a greater extent than they otherwise would, current evidence should allow us to predict a positive impact of such interventions.

Both the evidence in favor of as well as the movement toward teaching for active learning may be a consequence of a gradual shift from a paradigm of teaching to one of learning, as described by Barr and Tagg (1995). Under the instruction paradigm, teaching was the end point in higher education; that is, the institutional aim and goal was to teach. Under the learning paradigm, on the other hand, student learning is the end point, whereas teaching is the means, the method of producing learning (Fear et al., 2003). Active learning methods tend to be more student and learning oriented than traditional teaching methods, in that they require students to not only listen but also to read, write, discuss, and be engaged in problem-solving activities (Freeman, Eddy, McDonough, et al., 2014; Prince, 2004). At their best, these methods may help students grow as self-regulated learners—that is, to think about, participate in, and regulate their own learning process (Zimmerman, 1986; Zimmerman & Schunk, 2011).

In his massive synthesis of meta-analyses on teaching and learning strategies, Hattie (2009, 2011, 2015) found that promotion of metacognitive and self-regulatory strategies stimulating active learning is an important factor for academic performance. Hattie's main conclusion is that learning should be *visible*. Moving problem-solving and discussion activities into the classroom seems likely to make students' learning processes—not only outcomes revealed on tests and exams—more visible to both student and teacher. The flipped classroom may accommodate more interaction between teachers and students. Indeed, if such interaction leads to closer relationships between teachers and students, it may improve students' academic achievement and persistence (Robinson, Scott, & Gottfried, 2019).

Another contribution to the positive impact of active learning teaching methods may be that they tend to further student-to-student social interaction. A number of the most popular pedagogical approaches under the active learning umbrella emphasize the social nature of the prescribed activities. This is evident already in the labels chosen by their proponents, such as "cooperative learning" (Johnson & Johnson, 1999), "collaborative learning" (Bruffee, 1987), "team-based learning" (Michaelsen, Knight, & Fink, 2002), and so forth. Reviews of studies investigating the effects of such approaches on student achievement and other outcomes seem to indicate they do, indeed, contribute positively (e.g., Burgess et al., 2014; Johnson et al., 2014; Liu & Beaujean, 2017; Pai et al., 2015). Intriguingly, Foldnes (2016) found a stronger effect on learning when the flipped classroom intervention involved cooperative learning than when it did not. Thus, it seems possible that the extent to which learning activities are social in nature could influence the effect of flipped classroom interventions.

Many proponents of the flipped classroom advocate the use of some mechanism to ensure student preparation, for instance, in the form of a quiz before or at the beginning of class (Talbert, 2017). Such mechanisms may serve as motivators for studying. Furthermore, an extensive literature documents the positive impact of practice testing on learning outcomes (Dunlosky & Rawson, 2015; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger, 2011). Hew and Lo (2018) found in their meta-analysis of flipped classroom studies in medicine and health education that quizzes at the beginning of class were a significant moderator of the summary effect size estimate. Thus, it seems appropriate to

investigate if this association holds in a larger sample of studies and across disciplines.

There are also some potential weaknesses of the flipped classroom model. First, the elements we have mentioned above (preparation, tests, activity), which supposedly should enhance learning, are also likely to entail a larger workload on students. This could affect outcomes, in particular student satisfaction (Centra, 2003). Second, although the theory neutrality of the flipped classroom, reflected in our definition, may be seen as an advantage, it may also be that it leads to a lack of pedagogical rationale (Bishop & Verleger, 2013; Lundin et al., 2018) on which to design the specifics of a flipped classroom intervention, thus leaving teachers and students to more of a trial-and-error approach.

Student Perceptions of Flipped Classrooms

Student evaluations of teaching are widely relied on to inform decisions at course, program, and university levels, and they constitute an important source of data in quality assurance systems. Thus, the possible impact on student ratings is likely to, and arguably should, be taken into account when considering changes to teaching. Moreover, even though academic achievement or objective measures of learning may be considered more important, no evaluation of a teaching model or method should rely on only one type of data (Benton & Cashin, 2012). This is underscored by the controversies surrounding the uses of student ratings. Benton and Cashin (2012) concluded that the reliability, stability, and generalizability of student ratings of professors is satisfactory, and one might intuitively expect a positive association between student evaluations of teaching effectiveness and learning outcomes. However, a recent meta-analysis found that these correlations were minimal (Uttl, White, & Gonzalez, 2017). Others have voiced their concern that student evaluations may be systematically biased (Stroebe, 2016) or that they may not actually reflect teaching effectiveness (Boring, Ottoboni, & Stark, 2016). Furthermore, there are reports of flipped classroom interventions being associated with improved learning outcomes along with mediocre or reduced student satisfaction in the same student sample (e.g., Della Ratta, 2015; Missildine, Fountain, Summers, & Gosselin, 2013). Given the unobvious relationship between learning and student satisfaction, we cannot assume that an effect on learning is accompanied by a corresponding effect on satisfaction, and the possible effect of the flipped classroom on both outcomes should therefore be investigated.

Previous Reviews

A number of reviews have already provided qualitative summaries of the flipped classroom literature. These leave a somewhat mixed impression with regard to the model's benefits. Some have found that studies of the flipped classroom tend to report increased satisfaction and improvement in examination results or course grades (Akçayır & Akçayır, 2018; Brewer & Movahedazarhouligh, 2018; Karabulut-Ilgu et al., 2018), but the evidence is at least partly anecdotal (Bishop & Verleger, 2013) and often mixed (Betihavas et al., 2016; F. Chen, Lui, & Martinelli, 2017), especially with regard to student perceptions (Akçayır & Akçayır, 2018; Brewer & Movahedazarhouligh, 2018). Some reviewers claim that there is little "compelling evidence for the effectiveness of the method" (Evans et al., 2019, p. 74) or even "very little evidence of effectiveness" (Abeysekera & Dawson, 2015, p. 1). An early review noted that academic improvement was sometimes accompanied by negative student attitudes (O'Flaherty & Phillips, 2015).

Meta-analyses could sharpen the somewhat blurry picture left by the qualitative summaries in narrative and systematic reviews. However, to date, most meta-analyses of flipped classroom studies are limited to the medical and health professions disciplines. Although all of these report positive effects, summary estimates range from statistically nonsignificant (Gillette et al., 2018) or relatively modest ($Z = 0.33$; Hew & Lo, 2018), through moderate (0.47; K. S. Chen et al., 2018), to very large (1.06–1.68; Hu et al., 2018; Tan et al., 2017). An analysis by Cheng, Ritzhaupt, and Antonenko (2018) is the only one that, like the present one, spans all disciplines, though with only 55 studies included in their final analysis. It reveals a modest overall effect size estimate (0.19) in favor of the flipped classroom but also evidence that the effect is moderated by subject area, with the arts and humanities showing the largest mean difference.

In sum, these analyses span a broad range of overall effect size estimates. Only one of them (K. S. Chen et al., 2018) used meta-regression to examine the influence of moderators, none of them examined the potentially moderating effects of the social nature of learning activities, and only Hew and Lo (2018) examined the influence of testing student preparations. None of the previous meta-analyses estimated the overall effect on student satisfaction or pass rates. Furthermore, in view of the large number of studies comparing the flipped classroom with traditional controls, previous analyses included relatively few studies in total (with Cheng et al., 2018, being the largest), thus possibly limiting their potential to detect moderating factors.

Aims of This Study

We designed this study to extend the insights of previous reviews and meta-analyses and to overcome some of their weaknesses. The main purpose was to investigate the effects of flipped classroom interventions on student learning outcomes and satisfaction. Furthermore, we wanted to investigate whether specific characteristics of the implementation moderate these effects. To address these issues, we conducted an extensive search of the literature and synthesized evidence

from comparisons of flipped classroom interventions with traditional teaching control conditions in the largest meta-analysis on this topic to date. The following research questions guided this effort:

1. Do flipped classroom interventions affect student learning (i.e., exam scores or grades and pass/fail rates) positively across different disciplines?
2. Is there an effect of flipped classroom interventions on student satisfaction?
3. Are any effects of flipped classroom interventions moderated by specific characteristics of the implementation (i.e., the social nature of learning activities, tests of student preparation)?

Methods

Information Sources and Search Strategy

A keyword search was performed in eight different reference databases: Education Research Complete (Ebsco), ERIC (Ebsco), Medline (Ovid), Embase (Ovid), PsycINFO (Ovid), Web of Science (Clarivate Analytics), Scopus (Elsevier), and BASE (Universität Bielefeld)—covering education research from a wide range of disciplines. Together, these databases provide extensive coverage of both peer-reviewed journal articles and gray literature (doctoral dissertations, master's theses, and conference contributions). All the databases were searched up to May 2017, with no lower date limit. The search strategies are provided in Table 1. In addition, we scanned reference lists from selected reviews to identify studies not captured in our database search.

Eligibility Criteria

We included randomized or quasi-experimental intervention studies comparing a flipped classroom intervention with traditional teaching, concurrent or nonconcurrent.

Participants. Studies with students at any level of education and from any academic discipline were eligible for inclusion.

Intervention and Comparison. Studies were eligible for inclusion if they compared flipped classroom teaching with traditional teaching—specifically, if (a) most of the information transfer in the intervention condition was moved outside of and prior to the classroom session, (b) most of the in-session teaching in the intervention condition involved active learning, and (c) the control condition was predominantly teacher centered and lecture based. In some studies, the control condition would contain elements of active learning. These studies were nevertheless included if the flipped classroom intervention increased the session time allocated to active learning and decreased the session time allocated to information transfer, relative to the control condition. Studies older than 2010 were not included as, until then, interventions would not have been explicitly labeled as a flipped classroom. Such studies would therefore be impossible to systematically identify.

Outcomes. The studies had to include at least one of the following outcome measures: (a) continuous-learning outcomes, (b) some measure of pass/failure rates, or (c) some measure of student perception of teaching/course quality as defined above. By continuous-learning outcomes, we mean grades or scores on exams, tests, or concept inventories. Pass/failure rates were retrieved from reported pass/fail rates, DFW rates (percentage of students who receive grades of D or F, or withdraw from class), or grade distributions.

Reasons for Exclusion. Studies were excluded if they were (a) not reported in English, German, or one of the Scandinavian languages, (b) if the statistics necessary to calculate effect sizes were neither reported nor obtainable on request, (c) if the comparison was pre/post only, or (d) if the full text was not obtainable.

Study Selection

A preliminary screening was performed by the first author alone, separating obviously ineligible studies (nonempirical reports, studies without a comparison group, or studies satisfying any of the other exclusion criteria above) from possibly eligible ones. Both authors discussed and determined inclusion from among the possibly eligible search results. If an otherwise eligible study report did not include the statistics required to calculate effects sizes, we emailed the authors and requested that they provide them. To avoid including the same study or sample more than once, we matched the author names of eligible study reports against those of already included studies. In the case of any matches, we carefully compared the reports.

Coding of Moderators and Risk of Bias

We developed a schema for extracting data and for coding quality indicators. We sought information about the publication, the study, the participants, the interventions, and the outcome measures. Notably, we coded whether or not the intervention activities were social in nature and whether student preparation was tested, to allow us to investigate Research Question 3 (if these characteristics influence summary effect sizes). Finally, we coded other moderators of possible interest. For instance, we coded discipline because most previous meta-analyses focused on health disciplines alone, whereas this study includes all disciplines. In addition, Cheng et al. (2018), found that studies in the arts and humanities yielded a larger summary effect size estimate than those in other disciplines. Discipline was coded as

TABLE 1
Search Strategies

Databases (providers)	Search strategies
Medline, Embase, PsycINFO (Ovid)	1. ((flip* OR invert*) adj3 (classroom* OR learning OR instruction*).mp 2. invertebrate*.mp 3. 1 not 2
Education Research Complete, ERIC (Ebsco)	(flip* OR invert*) N3 (classroom* OR learning OR instruction*)
Web of Science Core Collection (Clarivate Analytics)	1. TS=((flip* OR invert*) NEAR/3 (classroom* OR learning OR instruction*)) 2. TS=invertebrate* 3. #1 NOT #2
SCOPUS (Elsevier)	(TITLE-ABS-KEY ((flip* OR invert*) W/3 (classroom* OR learning OR instruction*))) AND NOT (TITLE-ABS-KEY (invertebrate*))
BASE (Universität Bielefeld)	(flip* invert*) classroom*

STEM (science, technology, engineering, and mathematics), M/H (medicine and health sciences), SS (social sciences), or HUM (humanities).

Risk of Bias in Individual Studies. Our quality indicators were loosely based on Goldstein, Lackey, and Schneider (2014) and on the risk of bias assessment guidelines in Higgins and Green (2011). We pilot tested the system on five studies. After minor revisions, both authors performed the coding of the remaining studies independently. Quality indicator ratings were assigned for design features (group equivalence, attrition, and implementation fidelity) and measurement features (quality of learning outcome assessment and student satisfaction assessment). Attrition and implementation fidelity proved impossible to code reliably as most study reports did not provide the pertinent details. These variables were therefore not included in any analyses.

We coded group equivalence into three categories: *unequal* (the groups were different at baseline; they were reported equivalent only on measures not related to learning, e.g., sex or age; or there was no mention of group equivalence in the study report), *equal* (students in both conditions were similar at baseline on measures related to learning, e.g., on a knowledge pretest, on grade point average, or on SAT scores), and *random* (participants were randomly allocated to conditions).

The quality of learning outcome assessment was coded into three categories: *different* (assessment was likely based on similar criteria in both conditions or on different criteria), *identical* (assessment was identical in both conditions), and *blind* (assessment was identical in both conditions and at least partly blind-coded).

Satisfaction assessment was coded into two categories: *simple* (a single item or a few items) and *comprehensive* (a measure based on a number of items regarding the student's satisfaction and perception of teaching or course quality).

Cohen's kappas for coding from the two authors were .56 for satisfaction assessment, .72 for social activities, and .74 for test of preparation. Weighted kappas (weights based on squared distances; Fleiss & Cohen, 1973) were .70 for group equivalence and .76 for learning outcome assessment. We resolved disagreements in coding by discussion.

Planned Methods of Analysis

The meta-analyses were done according to the method of Hedges and colleagues (Hedges & Olkin, 1985; Hedges & Vevea, 1998), which calculates weights after converting effect sizes into Fisher's z. All analyses were performed using Comprehensive Meta-Analysis Version 3 software (Biostat, Inc., Englewood Cliffs, NJ, USA). For the learning and satisfaction outcomes, we used Hedges' g as the effect size measure. For pass rates, we used odds ratio (OR).

Given the probable variability in effects sizes in our sample of studies, we chose the random effects model to estimate the mean effect size. This model arguably provides a better fit with real-world data and tends to reduce the probability of Type I errors in significance tests (Field, 2005; Hunter & Schmidt, 2000). With the random effects model, it is assumed that true effects sizes vary across studies, and therefore estimates of both within-studies and between-studies variance are included in the calculation of mean effect size (Field, 2001).

To investigate the potential influence of characteristics of the interventions, we performed four subgroup analyses, using discipline, education level, social learning activities, and test of preparation as moderators. In addition, to assess the potential influences of other study characteristics on the summary effects and on the influence of the moderators above, we performed a meta-regression. A meta-regression allows estimating the effects of both continuous and categorical predictors on the dependent estimates. Most important, like regular multiple regression, it estimates a moderator's influence while taking into account the influence of the other predictors in the model. In addition to the factors above, the meta-regression model also included publication year, group equivalence,

quality of learning/satisfaction measures, and whether or not teachers were the same in the intervention and control conditions. These latter factors were included to assess the robustness of the main estimates, as recommended by Seifert, Bowman, Wolniak, Rockenbach, and Mayhew (2017).

Heterogeneity Tests. We tested heterogeneity in the effect sizes by using the Q statistic. This test indicates whether the variability among study effect sizes is due to true variability in the population of studies or merely a result of sampling error. To quantify the amount of variability in study effect sizes that can be ascribed to true variance, we used the I^2 statistic (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Cumulation of Findings Within Studies. If a study reported more than one continuous measure of learning, we based the effect size on the final, cumulative exam, if it was reported. If the several measures were equally important (e.g., several exams or one concept inventory and one exam), we computed a single composite effect size from an average of effect sizes based on those measures. In doing this, we assumed a correlation of 1 between the measures, slightly overestimating the variance of the study effect size (Borenstein et al., 2009). In cases where a course grade was based on different grading schemes in the two conditions, we preferred less comprehensive but identical measures.

From studies reporting comparisons between one intervention group and several control groups, or between one control group and several intervention groups, we pooled means, standard deviations, and samples from the subgroups, if both satisfied our definitions of flipped or traditional classrooms.

Publication and Small-Study Bias. Although we have taken steps to avoid a biased selection of studies by searching several databases indexing conference papers, doctoral dissertations, master's theses, as well as a wide selection of institutional archives through BASE, we should nevertheless check for bias as statistically significant findings are more likely to be reported than null results (Dickersin, 2005; Greenwald, 1975; Pigott, Valentine, Polanin, Williams, & Canada, 2013; Rosenthal, 1979). Furthermore, there is a risk that small, underpowered studies overestimate any effects reported, partly because statistical tests in such studies will reach significance only if the sample mean difference happens to be very large (Button et al., 2013; Ioannidis, 2008; Sterne, Gavaghan, & Egger, 2000).

We assessed our data for the presence of publication and small-study bias for each of our three outcome measures by visually inspecting funnel plots (Light & Pillemer, 1984), calculating rank correlations (Kendall's Tau; Begg & Mazumdar, 1994) and Egger's regression intercept (Egger, Smith, Schneider, & Minder, 1997), and running cumulative meta-analyses. The impact of a potential publication bias was assessed using fail-safe N and trim-and-fill analyses (Borenstein et al., 2009; Duval & Tweedie, 2000a, 2000b; Sutton, 2009). More detail on these methods can be found in the online Appendix.

Results

Data used in the analyses in this article are available from https://doi.org/10.18710/QA8WBZ (Låg & Sæle, 2019).

Study Selection and Study Characteristics

Figure 1 depicts the selection process.

A total of 272 samples were included in the analyses on the continuous-learning measures. Pass rates were reported for 45 samples and student satisfaction for 69 samples. Table 2 tabulates the number of samples for each outcome measure for the categories defined by our categorical moderators, as well as descriptives for total sample size. Figure 2 illustrates the number of included samples by report publication year and indicates a marked growth in the number of studies from 2015. Because of the large number of included studies, full bibliographic references are in a separate file available from https://doi.org/10.18710/QA8WBZ.

Main Results

Effect of the Flipped Classroom on Student Learning. In our initial analysis, the overall mean effect size for the continuous-learning outcomes was a standardized mean difference (Hedges' g) of 0.35 (95% confidence interval [CI] [0.31, 0.40], $Z = 14.68$, $p < .001$, $k = 272$), indicating that, on average, student learning as measured by exam scores, test scores, or grades is somewhat more than one third of a standard deviation higher under flipped classroom conditions than in traditional classrooms. (See Figure 3 for a breakdown of summary effect size estimates and 95% CIs by discipline.) This is, however, probably an overestimate because of the influence of the smaller studies, as indicated by our sensitivity analyses reported below. An analysis based only on those studies with 80% power to detect a standardized mean difference of 0.30 (studies with a total sample size larger than 174) yields an estimated summary Hedges' g of 0.24 (95% CI [0.18, 0.31], $Z = 7.36$, $p < .001$, $k = 90$), which is probably closer to the true mean effect size (see the subsection Publication and Small-Study Bias below).

Effects of Flipped Classroom Interventions on Student Satisfaction and Pass Rates. The overall mean effect size for student satisfaction was Hedges $g = 0.16$ (95% CI [0.06, 0.26], $Z = 3.1$, $p < .01$), indicating that flipping the classroom has a small positive effect on student satisfaction. The overall mean effect size for pass rates was an odds ratio of 1.55 (95% CI [1.34, 1.78], $Z = 6.06$, $p < .001$). This corresponds to a risk ratio of 1.08, indicating that students in flipped classrooms are on average 1.08 times more likely to pass

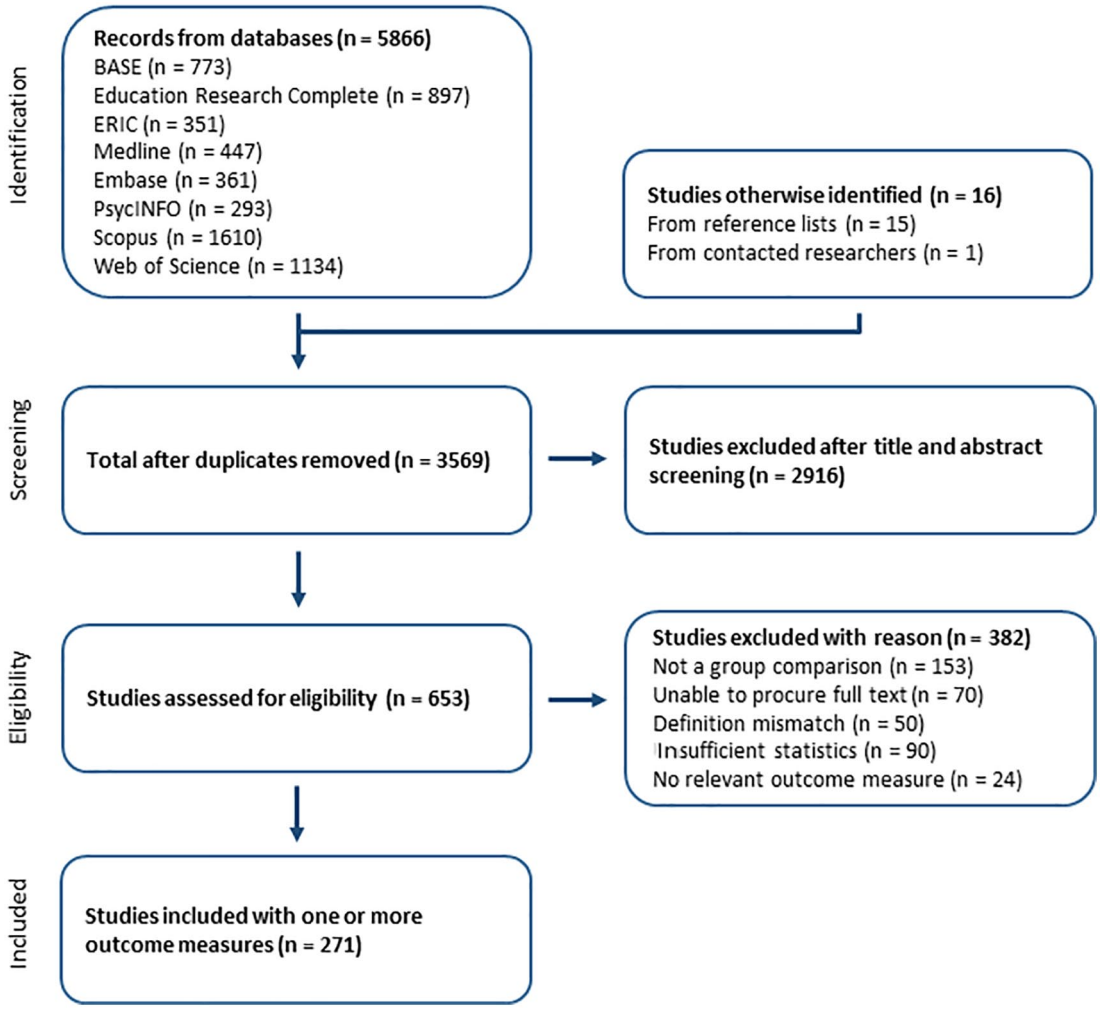

FIGURE 1. *Flow of studies.*

than students in traditional classrooms. Mean pass rates were 77.63 % under flipped classroom conditions and 75.91 % under traditional classroom conditions. Thus, the effect on pass rates, though statistically significant, is rather small. Again, both these estimates may be influenced by publication or small-study bias (see the relevant subsection below).

Subgroup Analyses. Test for heterogeneity indicated the presence of true study population variability for continuous-learning measures (Q = 1,591.42, degrees of freedom [df] = 271, $p < .001$, $I^2 = 82.97$), student satisfaction measures (Q = 374.41, $df = 68$, $p < .001$; $I^2 = 81.84$), and pass rates (Q = 91.26, $df = 44$, $p < .001$; $I^2 = 51.79$). All of the I^2 values are high, indicating substantial true variability in the study population. We next report the results of analyses that examined potential sources of this variability (see Table 3 for subgroup analyses and Table 4 for meta-regression).

The effect of discipline was statistically significant (Q = 8.06, $p < .05$) because of the difference in effect size between the humanities (Hedges' $g = 0.54$) and the STEM disciplines (Hedges' $g = 0.32$). Average mean differences between the flipped and control conditions were somewhat higher in studies on primary and secondary (Hedges' g = 0.44 and 0.45, respectively) than in studies on tertiary (Hedges' $g = 0.34$) education, but the difference was not statistically significant ($Q = 1.81$, $p = .404$). A comparison of interventions that included a test of preparation and those that did not indicated that whereas the summary effect size was higher (Hedges' $g = 0.40$) in interventions that tested preparation compared with interventions that did not (Hedges' $g = 0.31$), this difference was only marginally statistically significant ($Q = 3.8$, $p = .051$). There was no detectable difference between studies that described social and group-based activity in the intervention and studies that did not ($Q = 0.13$, $p = .72$).

We performed the same subgroup analyses for pass rates and satisfaction outcomes, with no statistically significant differences found for any of the moderators (all ps > .10).

TABLE 2
Number of Studies in Each of the Subcategories

Variable	*k*		
	Learning	Pass/fail	Satisfaction
Total	272	45	69
Discipline			
STEM	159	37	34
M/H	52	5	21
SS	33	2	11
HUM	28	1	3
Test of preparation			
No	132	18	32
Yes	140	27	37
Social activity			
No	77	9	13
Yes	195	36	56
Group equivalence			
Unequal	134	33	29
Equal	111	11	34
Randomized	27	1	6
Teachers			
Different across conditions	110	17	24
Same across conditions	162	28	45
Quality of learning assessment			
Different	77	25	
Identical	122	13	
Blind	73	7	
Quality of satisfaction assessment			
Simple			36
Comprehensive			33
Level			
Primary education	12		1
Secondary education	20		2
Higher education	239	45	65
Further education	1		1
Average study sample size, *M* (*SD*)	187.9 (212.97)	466.87 (594.28)	151.32 (128.44)

Note. *k* = number of studies; STEM = science, technology, engineering, and mathematics; M/H = medicine and health sciences; SS = social sciences; HUM = humanities; *SD* = standard deviation.

Meta-Regression Analysis. When all the moderators are included in the same analysis, the effect of discipline falls below the level of significance. Test of preparation, which was marginally significant in the subgroup analysis, is now significant ($p = .01$). The group equivalence moderator is a statistically significant predictor, indicating that studies with random allocation to groups have higher effect sizes than studies with unequal groups. Learning outcome assessment is also a significant predictor, with studies with identical assessment having slightly higher effect sizes than studies with identical and blind assessment. Importantly, study sample size seems to exert a fairly strong influence on effect sizes, with smaller studies having larger effect sizes.

Publication and Small-Study Bias

Funnel plots for all three outcome measures are provided in Figure 4. Details on the conduction and the results of the publication bias assessments are in the online Appendix. In this section, we present the overall conclusions.

For the continuous-learning measures outcome, there are clear indications of a small-study or publication bias. When

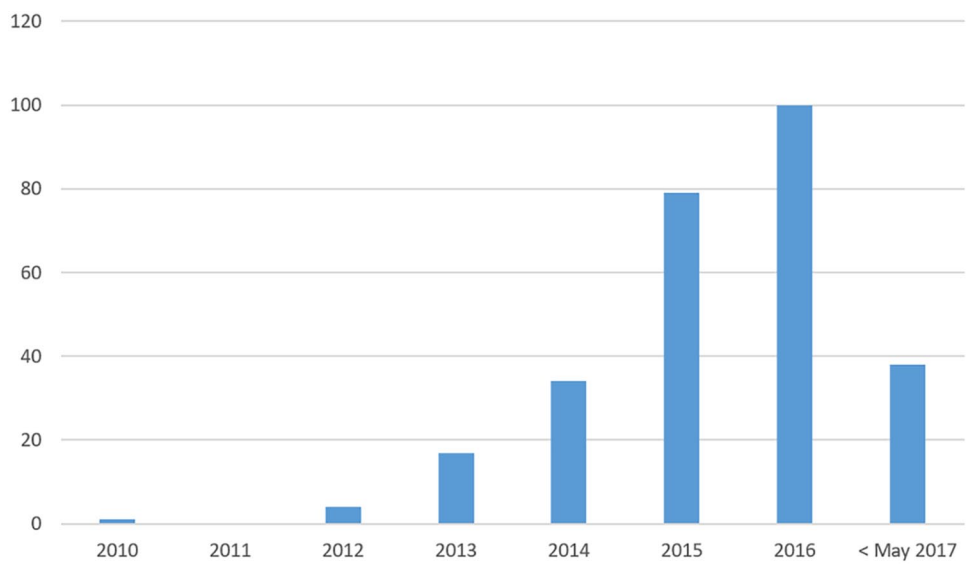

FIGURE 2. *Number of included studies by publication year.*

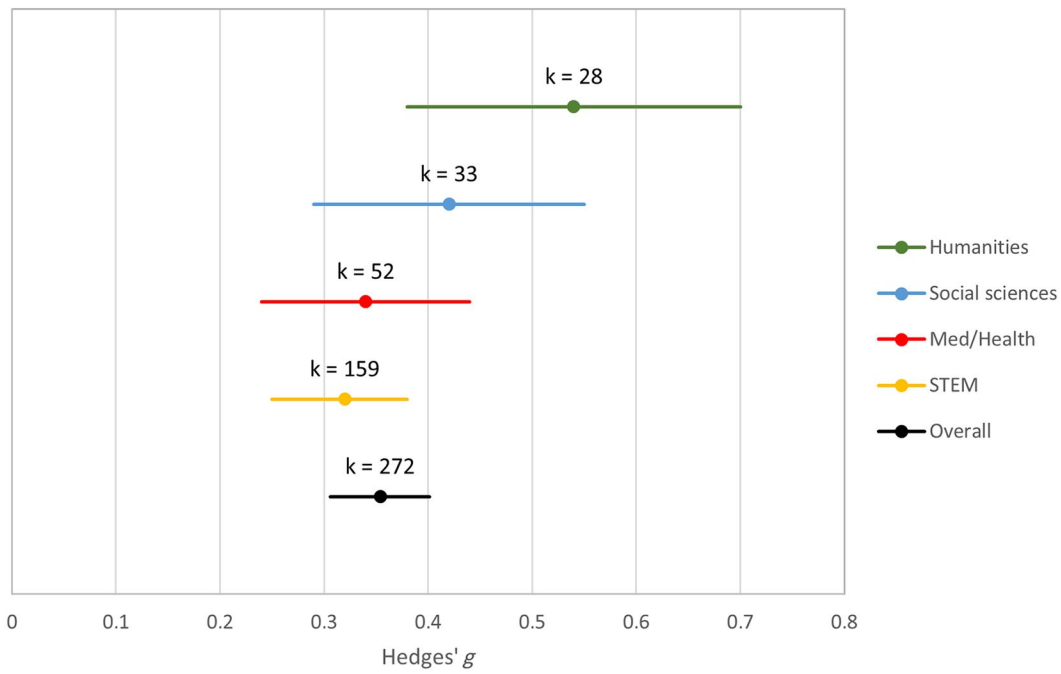

FIGURE 3. *Summary effect sizes and 95% confidence intervals by discipline (continuous-learning measures).*

restricting the meta-analysis to those studies with 80% power to detect a standardized mean difference of 0.30 (studies with a total sample size larger than 174), the summary estimate is reduced to Hedges' $g = 0.24$ (95% CI [0.18, 0.31], $Z = 7.36$, $p < .001$, $k = 90$). A trim-and-fill adjustment yields an estimate of 0.17 [0.12, 0.22]. Together, these adjustments suggest that the true weighted mean effect is somewhere around one fifth of a standard deviation.

For pass rates and student satisfaction outcomes, the results of publication bias assessments are less clear-cut, but they do seem to indicate the influence of bias. After applying trim and fill to the satisfaction outcome analysis, the summary estimate changes from 0.16 to −0.03 (95% CI [−0.14, 0.08]), indicating that, on average, flipped classroom interventions may have little influence on student satisfaction. For pass rates, the estimate changes from an OR of 1.55 to

TABLE 3
Subgroup Analyses of the Effect of the Flipped Classroom on Continuous-Learning Outcomes

Subgroup	*k*	Hedges' *g* [95% CI]	*Q* (*df*)	*p*
Baseline	272	0.35 [0.31, 0.40]	1591.42 (271)	<.001
Discipline			8.06 (3)	.045
STEM	159	0.32 [0.25, 0.38]		
M/H	52	0.34 [0.24, 0.44]		
SS	33	0.42 [0.29, 0.55]		
HUM	28	0.54 [0.38, 0.70]		
Education level			1.81 (2)	.404
Primary	12	0.44 [0.20, 0.67]		
Secondary	20	0.45 [0.27, 0.64]		
Tertiary	240	0.34 [0.29, 0.39]		
Test of preparation			3.8 (1)	.051
No	132	0.31 [0.24, 0.37]		
Yes	140	0.40 [0.33, 0.47]		
Social activity			0.13 (1)	.72
No	77	0.34 [0.25, 0.43]		
Yes	195	0.36 [0.30, 0.42]		

Note. k = number of studies; CI = confidence interval; *df* = degrees of freedom; STEM = science, technology, engineering, and mathematics; M/H = medicine and health sciences; SS = social sciences; HUM = humanities.

TABLE 4
Meta-Regression of Predictors on Learning Outcomes

	Q (*df*)	Coefficient [95% CI]	*Z*	*p*
Intercept		−67.89 [−150.76, 14.99]	−1.61	.11
Discipline[a]	4.57 (3)			.21
M/H		0.01 [−0.12, 0.13]	0.13	.90
SS		0.11 [0.04, 0.26]	1.44	.15
HUM		0.15 [−0.02, 0.32]	1.71	.09
Test of preparation		0.12 [0.02, 0.22]	2.44	.01
Social activity		−0.03 [−0.13, 0.08]	−0.48	.63
Group equivalence[b]	14.83 (2)			<.001
Equal		0.01 [−0.12, 0.10]	−0.19	.85
Randomized		0.32 [0.14, 0.49]	3.57	<.001
Learning outcome Assessment[b]	10.48 (2)			.005
Identical		0.11 [−0.01, 0.23]	1.87	.06
Blind		−0.08 [−0.21, 0.06]	−1.14	.25
Same teachers across groups		−0.04 [−0.13, 0.06]	−0.76	.44
Study sample size (range: 13–1,554)		−0.00 [−0.00, 0.00]	−2.46	.01
Publication year (range: 2010–2018)		0.03 [−0.01, 0.08)	1.61	.11

Note. CI = confidence interval; *df* = degrees of freedom; STEM = science, technology, engineering, and mathematics; M/H = medicine and health sciences; SS = social sciences; HUM = humanities. R^2 analog =.08.
[a]Reference category: STEM.
[b]Reference category: Different.

1.51 (95% CI [1.30, 1.74]), reducing the corresponding risk ratio from 1.08 to 1.03.

Note, though, that trim-and-fill adjustments should be interpreted with caution; they may be taken to be approximate indications of the severity of publication bias but should not be considered precise corrections (see, e.g., Peters, Sutton, Jones, Abrams, & Rushton, 2007).

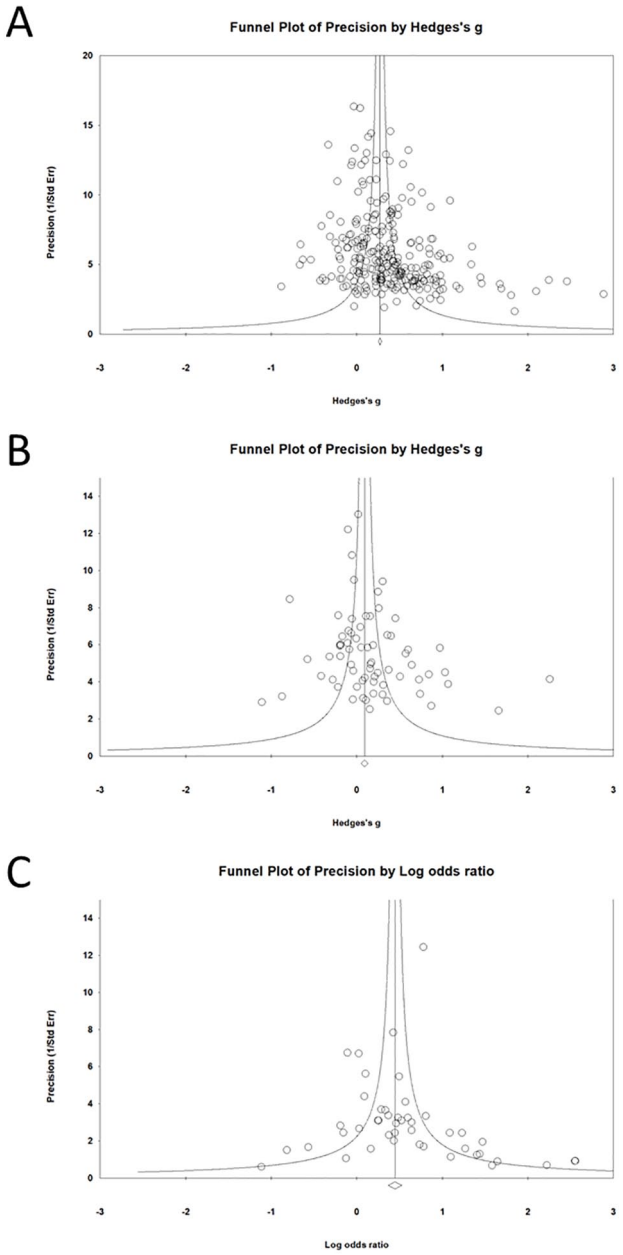

FIGURE 4. *Funnel plots for effect sizes. Panel A: continuous learning. Panel B: student satisfaction. Panel C: pass rates.*

Discussion

In this systematic review, we meta-analyzed studies comparing the flipped classroom with traditional teaching, to examine the effect of the flipped classroom pedagogy on learning outcomes and student satisfaction and to investigate whether specific characteristics of the implementation moderate such effects. In the following, we discuss each of the research questions in turn, before considering limitations and implications.

Do Flipped Classroom Interventions Positively Affect Student Learning Across Disciplines?

An initial, uncorrected estimate suggested that students in flipped classrooms outperformed students in traditional classrooms by 0.35 standard deviations on exams, tests, and grades. This estimate was, however, partly driven by the larger effects in underpowered studies. A reanalysis using only high-powered studies yielded a lower estimate of 0.24, and an adjustment using trim and fill (Duval & Tweedie, 2000a, 2000b) suggests an estimate of 0.17. Although estimates provided by trim-and-fill analyses are probably not very accurate (Peters et al., 2007), both adjustments together suggest that the summary effect, when accounting for small-study bias, is around one fifth of a standard deviation. This result is very similar to that obtained by Cheng et al. (2018). Like Cheng and colleagues, we found the effect to hold across disciplines, although the estimate for studies from the humanities were somewhat higher than for those from other disciplines, STEM in particular.

As the only meta-analysis to date, we investigated whether the flipped classroom students had higher pass rates than their traditional classroom counterparts. The meta-analytic estimates, both initial and adjusted for publication bias, might be considered small or even negligible. Still, they are statistically significant and in favor of the flipped classroom.

Thus, contrary to the conclusions drawn by some earlier reviewers (Abeysekera & Dawson, 2015; Evans et al., 2019), there is some evidence that the flipped classroom is effective. We have argued that the pedagogical possibilities within the flipped classroom setting, for example, for more learning activities, enhanced self-regulative abilities among students, and transparent learning processes, might be reasons to expect the model to be effective. Of course, our data can neither confirm nor disconfirm these ideas, but they do establish that the flipped classroom is likely to improve academic outcomes and that possible mechanisms should be explored.

Importantly, given the broad scope of our review, this result goes beyond those of most previous meta-analyses, which were confined to education in medicine and the health professions (K. S. Chen et al., 2018; Gillette et al., 2018; Hu et al., 2018; Tan et al., 2017). The summary estimate for continuous-learning measures is, however, both in our analyses and in those of Cheng and colleagues (2018), considerably lower than those obtained earlier. This is most likely consequence of including more studies, and studies from all disciplines, as few of the largest studies included by Cheng and colleagues and by us were included in previous meta-analyses. Furthermore, the sheer number of studies included in our analyses allowed us to uncover a publication bias—not uncovered in earlier meta-analyses—and adjust our estimates accordingly.

Another possible explanation for the relatively modest summary estimates is that the flipped classroom intervention's wide popularity is rather new. Hence, most of the studies included in this meta-analysis are first attempts at using the model. First-time implementations of new teaching methods may be more prone to unexpected obstacles and other teething problems, simply because of teacher and student inexperience. Thus, it might be argued that comparing an established teaching practice, incrementally improved through years of experience, with a first-time implementation of a new model does not provide for a fair test. It would be reasonable to expect that improvements may be made for future trials, perhaps also raising the effect.

We tested for differences between disciplines using both categorical moderator analyses and meta-regression. Although there is a trend toward a difference, such that STEM studies tend to have smaller effect sizes and studies on the humanities higher ones (reminiscent of the result in the meta-analysis by Cheng et al., 2018), most of the CIs around the discipline estimates overlap, indicating that variation between disciplines does not exceed that within disciplines. Hence, this difference should probably not be given too much weight. However, if we were to speculate on possible causes, we could point to the fact that a notable characteristic of the group of studies classified as belonging to the humanities is that a majority of them (23 of 33, or 70%) are on language-learning courses. Possibly, learning a new language or improving one's mastery of one includes learning processes particularly suited to the active, practice-oriented approach that is encouraged in a typical flipped classroom. Still, we reiterate that the most important finding is that summary estimates in favor of the flipped classroom are significant for all disciplines.

Is There an Effect of Flipped Classroom Interventions on Student Satisfaction?

The initial summary effect size estimate for the student satisfaction outcomes was Hedges' $g = 0.16$, but again, this is likely an overestimate due to small-study bias. Trim and fill yielded an adjusted estimate of −0.03, indicating that, on average, flipped classroom interventions may have little influence on student satisfaction.

The small summary estimate is due in part to there being a considerable number of effect sizes that are in favor of the control condition. Thus, some students are less satisfied with their experience of the flipped classroom than with traditional teaching methods. This is an important result because it illustrates a point made by several commentators, that student satisfaction is related to a number of factors that may be irrelevant to student learning (Boring et al., 2016; Spooren, Brockx, & Mortelmans, 2013; Uttl et al., 2017). Thus, students may dislike an educational intervention even though it improves learning. This can arise when there is a discrepancy between what students perceive to be an appropriate workload and the real workload entailed by the most effective teaching methods (Centra, 2003; Greenwald & Gillmore, 1997), or when workload is perceived to be not germane to instructional objectives (Marsh, 2001).

Still, the estimate should be interpreted with caution because of the considerable variation in how student satisfaction was measured in the included studies. This variation may have contributed to the equivocal result on this outcome.

Are Any Effects of Flipped Classroom Interventions Moderated by Specific Characteristics of the Implementation?

Meta-regression analysis on our full set of studies measuring continuous-learning outcomes indicated that implementations including a test of student preparation tended to yield slightly higher effect sizes. This interesting result is in line with that reported in the meta-analysis by Hew and Lo (2018), who found that quizzing students at the beginning of a flipped classroom session increased learning gains. This phenomenon may partly be a consequence of the indirect, motivational influence of testing, affecting students' willingness to engage in preparatory learning activities. It also seems likely that it is caused by the direct effects of retrieval practice (Dunlosky et al., 2013; Karpicke, 2012; Roediger, Agarwal, McDaniel, & McDermott, 2011), which such testing usually entails.

Moderator analyses did not provide evidence that the social nature of learning activities influenced learning gains from flipped classroom interventions. This is somewhat unexpected given the seemingly convincing evidence that collaborative, active teaching approaches generally are effective (Burgess et al., 2014; Foldnes, 2016; Johnson et al., 2014; Liu & Beaujean, 2017; Wiggins, Eddy, Grunspan, & Crowe, 2017). How the social activities are designed may influence the social interaction and its effect on learning outcomes. In this meta-analysis, interventions were characterized as social if they had any social element, for instance, letting students work in groups. Wiggins et al. (2017) showed that activities that were interactive in nature and depended on cooperation were beneficial in terms of both more student interaction and increased learning outcomes, compared with activities that were constructive in nature, even if students indeed worked in groups in both conditions. Another possible substantive explanation is that a number of factors (e.g., group duration and size, participant interdependence, and formal assessment) can influence the effectiveness of social activities (Tomcho & Foels, 2012). Recording such details may turn out to be difficult given the challenges with reliable coding of implementation characteristics in many flipped classroom study reports.

Indeed, our moderators account for very little of the large between-study heterogeneity, most likely because of somewhat less than ideal coding reliability. This would tend to obscure any associations. Another, potentially more interesting explanation is that we have not identified the moderators that systematically contribute to this variance. The considerable variability of the interventions and controls make it difficult to know what the effective components of flipped classroom interventions may be. And although the various theoretical approaches to designing learning activities may determine the effectiveness of the flipped classroom (Bishop & Verleger, 2013), the main goal of this analysis was to investigate if flipped classroom interventions, broadly conceived, are likely to affect learning and student satisfaction. Our definition of a flipped classroom intervention does not include any specified pedagogy or type of activities, nor did we exclude studies with elements of active learning in the traditional classroom. Hence, the variation in interventions and control conditions may indeed have deflated the effect size estimates. If the conditions were more homogeneous, a larger summary effect size estimate may have been expected (c.f. Freeman, Eddy, Jordt, Smith, & Wenderoth, 2014).

Limitations

An important limitation of this review is that descriptions of both the interventions themselves and aspects of the study design varied in quality and completeness across the included studies, making them difficult to code accurately. Our measures of coding reliability reflect this. For instance, attrition is rarely mentioned in the reports. In some studies, especially small ones, it may be held as implicit that students followed the whole course, but we do not know this. In other studies, only completing students were included, and no information regarding students entering a course and dropping out before exams were available. Still, if attrition differs between intervention groups and control groups, this may influence the results. Students likely to drop out in one condition could have reached a passing grade in another, hence the distribution of grades would be distorted, as discussed, for example, by Missildine et al. (2013).

Another limitation is that most of the included studies have a high risk of selection and/or detection bias. Few studies used random allocation (none, naturally, with concealment), many do not report baseline group equivalence, and few used blinded outcome assessment. Stronger study designs would allow for more confidence that flipped classroom interventions are likely to be effective. Even so, the meta-regression analysis on the present data revealed no clear indications that the poorly controlled studies systematically overestimated the advantages of flipping.

Because of the often sparse descriptions, we may have included studies of interventions that are not perfect matches to our definition of the flipped classroom. Conversely, there are likely to be studies investigating flipped classroom–like implementations that do not use the words *flipped* or *inverted* to label their intervention. These would not be included in our search results as it have would been impossible to identify them systematically. Indeed, and probably reflecting this, none of the previous meta-analyses on the flipped classroom included studies prior to 2010, which was the lower limit for inclusion in this meta-analysis.

Furthermore, a number of studies did not report enough data to calculate effect sizes, nor were they supplied by the authors on request, and could not be included. Although there is no reason to suspect a systematic bias from these missing studies, we do not know how they would have influenced the estimates.

Implications for Educators and Stakeholders

A modest summary estimate of about one fifth of a standard deviation on continuous learning outcomes may seem somewhat disappointing to proponents of the model. Judged against the widely used cutoffs proposed by Cohen (1988), it would be regarded as a "small" effect size. Effects of this magnitude may, however, still have considerable practical significance (Lipsey et al., 2012). Consider a hypothetical course with 250 students, assessed on a 100-point scale, with a mean final exam score of 75, a standard deviation of 15, a pass/fail cutoff at 60 points, and a B to A limit at 90 points. Under these circumstances, about 40 students will fail, and about 40 students will get an A. Shifting this distribution 0.2 standard deviations to the right would reduce the number of failing students to 29 and increase the number of As to 53. Many would consider this a substantial gain. Thus, for teachers or study program managers looking for ways to introduce more active learning into classroom sessions, the flipped classroom seems to provide a viable route that is likely to positively affect student learning.

Nevertheless, we caution against interpreting the summary estimates from this study as the most likely consequence of implementing flipped classroom teaching regardless of circumstances. First, because of the extensive heterogeneity in our material, we know that the flipped classroom works really well in some cases but not at all in other cases. Furthermore, because our coded moderators can account for very little of this variability, we do not at present know under what circumstances the model is most likely to yield strong positive effects. An important exception is testing student preparation, which, given our results, is likely to increase learning gains under the flipped classroom model.

Another finding, albeit an uncertain one, with possible implications for educators, is that there seem to be no gains on average in student satisfaction for the flipped classroom model over traditional teaching. This may be due to discrepancies between student expectations of workload and actual workload and/or to activities not being perceived as useful to

instructional goals. If so, educators who want to introduce the flipped classroom model should endeavor to design learning activities that are closely aligned with formulated learning outcomes and discuss the purposes of the learning activities with their students. Importantly, educators and stakeholders should systematically assess both learning gains and satisfaction when evaluating course quality, as improvements in learning may not be associated with corresponding increases in student satisfaction.

Future Research

Most of the included studies report results from first attempts at the flipped classroom model. We encourage researchers to replicate studies after a few years' run, when necessary adjustments and improvements have been made to the interventions and when teachers and students are more familiar with the model.

Given the substantial heterogeneity among studies of the flipped classroom, and the relatively few and insubstantial clues as to what causes this variability, there seems to be a need for studies targeting specific features of flipped classroom interventions, pitting them against each other in direct and well-controlled comparisons. This will likely help identify the working mechanisms and moderators of the positive effect.

The limitations identified above introduce some uncertainty concerning both the summary effect size estimates and any potential influence of moderators. Thus, we recommend a number of measures to future researchers and editors:

1. Ensure complete descriptions of interventions and control conditions. This is recommended in various reporting standards (American Educational Research Association, 2006; Appelbaum et al., 2018; Schulz, Altman, & Moher, 2010) but is not generally followed.
2. Use clear and comparable outcome measures.
3. Create stringent study designs. In education research, true randomization is difficult to attain because teaching takes place in groups and few educational institutions carry out several versions of the same course in the same semester. To establish group equality, adequate reports of group characteristics are essential.
4. Provide for blinding of assessment. Using, for instance, external examiners who are not aware of the experimental conditions should reduce detection bias.
5. Secure adequate power by using sufficiently large sample sizes and precise outcome measures.
6. To reduce publication bias, submit and publish methodologically sound studies with adequate power, also in cases of null or negative findings.

Conclusions

A meta-analysis of studies comparing the flipped classroom with traditional teaching revealed small summary effect size estimates on student learning in favor of the flipped classroom. We found clear evidence of publication or small-study bias, and reanalyses suggest that the true mean effect sizes are lower than the initial estimates indicated. One likely explanation for the discrepancy in summary estimates on the learning outcome between our analysis and previous meta-analyses is that previous analyses were based on considerably fewer studies and none of them were able to identify the publication bias.

There is large heterogeneity between studies, and our moderators account for very little of it. Hence, we know little about why a flipped classroom intervention may work very well in some situations and be less effective in others. There is therefore a need for more stringent study designs and more complete and accurate reporting.

Nevertheless, educators attempting the flipped classroom teaching model are likely to experience at least a small positive impact on student learning. There is some support for the notion that this positive impact may increase slightly if a test of student preparation is a part of the implementation.

Acknowledgments

We thank the anonymous reviewers for their thoughtful and constructive comments.

ORCID iDs

Torstein Låg https://orcid.org/0000-0002-1325-5235
Rannveig Grøm Sæle https://orcid.org/0000-0003-0834-5532

References

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*, 1–14. doi:10.1080/07294360.2014.934336

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education*, *126*, 334–345. doi:10.1016/j.compedu.2018.07.021

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40. doi:10.3102/0013189x035006033

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*, 3–25. doi:10.1037/amp0000191

Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education. *Change*, *27*(6), 12–26. doi:10.1080/00091383.1995.10544672

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101. doi:10.2307/2533446

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan: University of Kansas. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf

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

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. doi:10.1016/j.nedt.2015.12.010

Bishop, J. L. (2014). *A controlled study of the flipped classroom with numerical methods for engineers* (Doctoral dissertation). Available from Proquest Dissertations and Theses Global Database. (UMI No. 3606852)

Bishop, J. L., & Verleger, M. A. (2013, June 23–26). *The flipped classroom: A survey of the research*. Paper presented at the ASEE Annual Conference and Exposition, Atlanta, GA. Retrieved from https://peer.asee.org/22585

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Boring, A., Ottoboni, K., & Stark, P. (2016, January 17). Student evaluations of teaching (mostly) do not measure teaching effectiveness [Online]. *ScienceOpen Research*. doi:10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Brewer, R., & Movahedazarhouligh, S. (2018). Successful stories and conflicts: A literature review on the effectiveness of flipped learning in higher education. *Journal of Computer Assisted Learning*, *34*, 409–416. doi:10.1111/jcal.12250

Bruffee, K. A. (1987). The art of collaborative learning: Making the most of knowledgeable peers. *Change*, *19*, 42–47.

Burgess, A. W., McGregor, D. M., & Mellis, C. M. (2014). Applying established guidelines to team-based learning programs in medical schools: A systematic review. *Academic Medicine*, *89*, 678–688. doi:10.1097/ACM.0000000000000162

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365. doi:10.1038/nrn3475

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 495–518. doi:10.1023/a:1025492407752

Chen, F., Lui, A. M., & Martinelli, S. M. (2017). A systematic review of the effectiveness of flipped classrooms in medical education. *Medical Education*, *51*, 585–597. doi:10.1111/medu.13272

Chen, K. S., Monrouxe, L., Lu, Y. H., Jenq, C. C., Chang, Y. J., Chang, Y. C., & Chai, P. Y. C. (2018). Academic outcomes of flipped classroom learning: A meta-analysis. *Medical Education*, *52*, 910–924. doi:10.1111/medu.13616

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*, *67*, 793–824. doi:10.1007/s11423-018-9633-7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Laurence Erlbaum.

Della Ratta, C. B. (2015). Flipping the classroom with team-based learning in undergraduate nursing education. *Nurse Educator*, *40*, 71–74. doi:10.1097/NNE.0000000000000112

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11–33). New York, NY: Wiley.

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology*, *1*, 72–78. doi:10.1037/stl0000024

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58. doi:10.1177/1529100612453266

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98. doi:10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*, 629–634. doi:10.1136/bmj.315.7109.629

Evans, L., Vanden Bosch, M. L., Harrington, S., Schoofs, N., & Coviak, C. (2019). Flipping the classroom in health care higher education: A systematic review. *Nurse Educator*, *44*, 74–78. doi:10.1097/NNE.0000000000000554

Fear, F. A., Doberneck, D. M., Robinson, C. F., Fear, K. L., Barr, R. B., Van Den Berg, H., . . . Petrulis, R. (2003). Meaning making and "The Learning Paradigm": A provocative idea in practice. *Innovative Higher Education*, *27*, 151–168. doi:10.1023/a:1022351126015

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161–180. doi:10.1037/1082-989X.6.2.161

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444–467. doi:10.1037/1082-989X.10.4.444

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619. doi:10.1177/001316447303300309

Foldnes, N. (2016). The flipped classroom and cooperative learning: Evidence from a randomised experiment. *Active Learning in Higher Education*, *17*(1), 39–49. doi:10.1177/1469787415616726

Freeman, S., Eddy, S. L., Jordt, H., Smith, M. K., & Wenderoth, M. P. (2014). Reply to Hora: Meta-analytic techniques are designed to accommodate variation in implementation. *Proceedings of the National Academy of Sciences of the U S A*, *111*, E3025. doi:10.1073/pnas.1410405111

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning

increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the U S A*, *111*, 8410–8415. doi:10.1073/pnas.1319030111

Gillette, C., Rudolph, M., Kimble, C., Rockich-Winston, N., Smith, L., & Broedel-Zaugg, K. (2018). A meta-analysis of outcomes comparing flipped classroom and lecture. *American Journal of Pharmaceutical Education*, *82*, 433–440.

Goldstein, H., Lackey, K. C., & Schneider, N. J. B. (2014). A new framework for systematic reviews: Application to social skills interventions for preschoolers with autism. *Exceptional Children*, *80*, 262–286. doi:10.1177/0014402914522423

Goodwin, B., & Miller, K. (2013). Evidence on flipped classrooms is still coming in. *Educational Leadership*, *70*(6), 78–80.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20. doi:10.1037/h0076157

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, *89*, 743–751. doi:10.1037/0022-0663.89.4.743

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*: London, England: Routledge.

Hattie, J. (2011). Which strategies best enhance teaching and learning in higher education? In D. Mashek & E. Y. Hammer (Eds.), *Empirical research in teaching and learning: Contributions from social psychology* (pp. 130–142). Malden, MA: Blackwell.

Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*, 79–91. doi:10.1037/stl0000021

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. doi:10.1037/1082-989X.3.4.486

Herreid, C. F., Schiller, N. A., Herreid, K. F., & Wright, C. B. (2014). A chat with the survey monkey: Case studies and the flipped classroom. *Journal of College Science Teaching*, *44*, 75–80. Retrieved from https://www.jstor.org/stable/43631780

Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: A meta-analysis. *BMC Medical Education*, *18*(1), 38. doi:10.1186/s12909-018-1144-z

Higgins, J. P. T., & Green, S. (Eds.). (2011, March). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, The Cochrane Collaboration). Retrieved from https://handbook-5-1.cochrane.org/

Hu, R., Gao, H., Ye, Y., Ni, Z., Jiang, N., & Jiang, X. (2018). Effectiveness of flipped classrooms in Chinese baccalaureate nursing education: A meta-analysis of randomized controlled trials. *International Journal of Nursing Studies*, *79*, 94–103. doi:10.1016/j.ijnurstu.2017.11.012

Hung, H.-T. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning*, *28*, 81–96. doi:10.1080/09588221.2014.967701

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*, 275–292. doi:10.1111/1468-2389.00156

Ioannidis, P. A. J. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. doi:10.1097/EDE.0b013e31818131e7

Johnson, D. W., & Johnson, R. T. (1999). *Learning together and alone: Cooperative, competitive, and individualistic learning* (5th ed.). Boston, MA: Allyn & Bacon.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (2014). Cooperative learning: Improving university instruction by basing practice on validated theory. *Journal on Excellence in College Teaching*, *25*, 85–118. Retrieved from https://www.researchgate.net/publication/284471328_Cooperative_Learning_Improving_university_instruction_by_basing_practice_on_validated_theory

Karabulut-Ilgu, A., Jaramillo Cherrez, N., & Jahren, C. T. (2018). A systematic review of research on the flipped learning method in engineering education. *British Journal of Educational Technology*, *49*, 398–411. doi:10.1111/bjet.12548

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157–163. doi:10.1177/0963721412443552

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Retrieved from https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

Liu, S.-N. C., & Beaujean, A. (2017). The effectiveness of team-based learning on academic outcomes: A meta-analysis. *Scholarship of Teaching and Learning in Psychology*, *3*, 1–14. doi:10.1037/stl0000075

Love, B., Hodge, A., Grandgenett, N., & Swift, A. W. (2014). Student learning and perceptions in a flipped linear algebra course. *International Journal of Mathematical Education in Science & Technology*, *45*, 317–324. doi:10.1080/0020739X.2013.822582

Lundin, M., Bergviken Rensfeldt, A., Hillman, T., Lantz-Andersson, A., & Peterson, L. (2018). Higher education dominance and siloed knowledge: Systematic review of flipped classroom research. *International Journal of Educational Technology in Higher Education*, *15*(1), 20. doi:10.1186/s41239-018-0101-6

Låg, T., & Sæle, R. G. (2019). *Study level data for Does the Flipped Classroom Improve Student Learning and Satisfaction? A systematic review and meta-analysis* [Dataset]. DataverseNO, V1. doi:10.18710/QA8WBZ

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*, 183–212. doi:10.3102/00028312038001183

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education*, *30*, 159–167. doi:10.1152/advan.00053.2006

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (2002). *Team-based learning: A transformative use of small groups*. Westport, CT: Greenwood.

Missildine, K., Fountain, R., Summers, L., & Gosselin, K. (2013). Flipping the classroom to improve student performance and satisfaction. *Journal of Nursing Education*, *52*, 597–599. doi:10.3928/01484834-20130919-03

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. doi:10.1016/j.iheduc.2015.02.002

Pai, H.-H., Sears, D. A., & Maeda, Y. (2015). Effects of small-group learning on transfer: A meta-analysis. *Educational Psychology Review*, *27*, 79–102. doi:10.1007/s10648-014-9260-8

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*, 4544–4562. doi:10.1002/sim.2889

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*, 424–432. doi:10.3102/0013189x13507104

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93*, 223–231. doi:10.1002/j.2168-9830.2004.tb00809.x

Robinson, C. D., Scott, W., & Gottfried, M. A. (2019). Taking it to the next level: A field experiment to improve instructor-student relationships in college. *AERA Open*, *5*(1), 1–15. doi:10.1177/2332858419839707

Roediger, H. L., III. (2011). Using testing to improve learning and memory. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society*. New York, NY: Worth.

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395. doi:10.1037/a0026252

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. doi:10.1037/0033-2909.86.3.638

Schulz, K. F., Altman, D. G., & Moher, D. (2010). Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, c332. doi:10.1136/bmj.c332

Seifert, T. A., Bowman, N. A., Wolniak, G. C., Rockenbach, A. N., & Mayhew, M. J. (2017). Ten challenges and recommendations for advancing research on the effects of college on students. *AERA Open*, *3*(2), 1–12. doi:10.1177/2332858417701683

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*, 598–642. doi:10.3102/0034654313496870

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*, 1119–1129. doi:10.1016/S0895-4356(00)00242-0

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, *11*(6), 800–816. doi:10.1177/1745691616650284

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435–452). New York, NY: Russell Sage Foundation.

Talbert, R. (2017). *Flipped learning: A guide for higher education faculty*. Sterling, VA: Stylus.

Tan, C., Yue, W.-G., & Fu, Y. (2017). Effectiveness of flipped classrooms in nursing education: Systematic review and meta-analysis. *Chinese Nursing Research*, *4*, 192–200. doi:10.1016/j.cnre.2017.10.006

Tomcho, T. J., & Foels, R. (2012). Meta-analysis of group learning activities: Empirically based teaching recommendations. *Teaching of Psychology*, *39*, 159–169. doi:10.1177/0098628312450414

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. doi:10.1016/j.stueduc.2016.08.007

Wiggins, B. L., Eddy, S. L., Grunspan, D. Z., & Crowe, A. J. (2017). The ICAP active learning framework predicts the learning gains observed in intensely active classroom experiences. *AERA Open*, *3*(2), 1–14. doi:10.1177/2332858417708567

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*, 307–313. doi:10.1016/0361-476X(86)90027-5

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2011). *Handbook of self-regulation of learning and performance*. New York, NY: Routledge.

Authors

TORSTEIN LÅG is a senior academic librarian at UiT The Arctic University of Norway. His research interests include research synthesis, information literacy, teaching, and learning.

RANNVEIG GRØM SÆLE is an associate professor at UiT The Arctic University of Norway. Her research interests include student learning, school dropout, teaching, and research synthesis.