Review of Educational Research
Month 202X, Vol. XX, No. X, pp. 1–41
DOI: 10.3102/00346543211019122
Article reuse guidelines: sagepub.com/journals-permissions

To Flip or Not to Flip? A Meta-Analysis of the Efficacy of Flipped Learning in Higher Education

Carrie A. Bredow
Patricia V. Roehling
Alexandra J. Knorp
Andrea M. Sweet
Hope College

*Although flipped classroom pedagogies have been widely touted for their ability to foster diverse 21st-century learning objectives, previous syntheses of flipped learning have focused almost exclusively on outcomes related to academic achievement. Using data from 317 studies, our research addresses this deficit by providing a comprehensive meta-analysis of the effects of flipped versus lecture-based learning on academic, intra-/interpersonal, and satisfaction-related outcomes in higher education. Overall, flipped classroom interventions produced positive gains across all three learning domains, and we found significant advantages of flipped over lecture-based instruction for seven out of eight outcomes (*gs *= 0.20–0.53). At the same time, there was substantial heterogeneity in flipped learning effects, and we identified several variables that influenced the relative efficacy of flipped versus traditional courses. Of the three types of moderators examined (contextual, design-based, and methodological), educational context (e.g., discipline, location) accounted for the most variability in flipped learning outcomes.*

Keywords: flipped classroom, higher education, learning outcomes, meta-analysis

Flipped learning, also known as inverted learning, is a relatively new teaching technique that has become increasingly popular over the past decade. The potential benefits of flipped learning have been touted in influential mainstream publications such as *The New York Times, Science*, and *The Chronicle of Higher Education* (Berrett, 2012; Fitzpatrick, 2012; Mazur, 2009). The flip, or inversion, occurs when students access lecture materials before class and engage in active learning or instructor-guided exercises during class time. Specific approaches vary, but the core philosophy is that by offloading the instruction of declarative or foundational knowledge, class time can be used to engage students in active

learning experiences that promote the development of more diverse and complex learning outcomes than are typically fostered in a lecture setting. Effective development of such "21st-century skills," including critical thinking, the ability to learn, and interpersonal competence, is considered a principle advantage of the flipped learning model (e.g., Hawks, 2014; Mazur, 2009).

But does flipped learning live up to all the hype? Although research comparing flipped learning with a traditional lecture format has burgeoned over the past decade (Al-Shabibi & Al-Ayasra, 2019; Yang et al., 2017), attempts to summarize this literature have been methodologically limited (e.g., do not use meta-analytic techniques; lax inclusion criteria) and/or narrow in scope (e.g., regional or discipline specific). Moreover, a majority of the reviews on the topic have focused exclusively on knowledge-based outcomes, yielding a growing—but narrow—body of evidence that flipped learning tends to be more effective than lecture-based learning for promoting academic achievement (e.g., Cheng et al., 2019; Karagöl & Esen, 2019). Little is known, in contrast, about the efficacy of flipped learning for producing other important outcomes, including the 21st-century skills it is most touted for improving. Our study addresses these deficits by providing a comprehensive meta-analysis of the efficacy of flipped versus lecture-based learning for fostering academic, intra-/interpersonal, and satisfaction-related outcomes in higher education. Specifically, we examine the following two research questions:

Research Question 1: Is flipped learning more effective than lecture-based learning for developing a range of learning outcomes in higher education?
Research Question 2: If efficacy varies across studies, to what extent can differences in educational context, course design, and study methodology account for this heterogeneity?

Defining the Flipped Classroom

Since its introduction into the education literature two decades ago (Lage et al., 2000), flipped learning has taken many forms. In the current meta-analysis, two essential features must be present to meet our criteria for flipped learning. First, content material, traditionally delivered during in-class lecture, must be delivered outside of the classroom via an audiovisual format (vodcast; O'Flaherty & Phillips, 2015). We exclude audio-only recordings (podcasts) because they do not contain the organizing information conveyed in a lecture outline or the complex information conveyed in graphs, figures, photographs, and illustrations. We also exclude courses that substitute in-class lectures with preclass readings or assignments instead of vodcasts. Although some have argued for including diverse forms of preclass material delivery under the flipped umbrella (Abeysekera & Dawson, 2015; van Alten et al., 2019), we believe video lectures are pedagogically distinct from these other forms of preclass content dissemination. Unlike podcasts or reading assignments, vodcasts are multisensory and tap into both visual and auditory modes of information processing, which may improve student learning and retention (J. L. Jensen et al., 2018; Moreno & Mayer, 2002). Moreover, instead of simply replacing direct instruction with out-of-class reading, courses utilizing vodcasts free up class time while retaining instructor-led content

delivery as a key pedagogical method. This combination of teacher- and student-centered teaching methods is an important feature of the flipped classroom that differentiates it from courses that utilize in-class active learning but rely primarily on outside texts for content delivery.

The second essential feature of flipped learning is that class time is used to engage in active learning experiences (Faculty Focus, 2015). At its core, active learning can be defined as any method that engages students in meaningful activities that require them to think about what they are doing (Prince, 2004). Although active learning experiences are quite varied, they often involve one or more of the following: (a) practicing or applying information through quizzes, presentations, or projects; (b) teaching others; (c) participating in engaging and challenging activities; and (d) exercising control over what is being learned or how it is learned (Wankat, 2002). Typical examples of active learning include discussion, writing, problem-solving, and practicing skills. A common thread running through all of these learning experiences is that students are engaged not just cognitively but also physically, and often interpersonally.

Notably, although we treat active and lecture-based class sessions as nonoverlapping forms of pedagogy, in practice, the distinction is not clear-cut. Flipped classes may include "just-in-time" mini-lectures and lecture-based classes may involve some active learning exercises (Lo et al., 2017). Indeed, it is not uncommon for "lecture-based" courses to intersperse the verbal presentation of facts and concepts with other pedagogical techniques (demonstrations, clicker questions, etc.; Hora & Ferrare, 2012). Nevertheless, college courses have been shown to differ in the *degree* to which they reflect an instructor-centered versus student-centered approach to constructing knowledge (Ferrare & Hora, 2014). Courses' relative emphasis on student-led learning over lecture during class is key to differentiating flipped and lecture-based instruction.

Finally, although flipped learning is used in a variety of educational settings—including primary and secondary schools—we limit our investigation to postsecondary students for several reasons. Given structural differences between higher education and primary/secondary education (e.g., degree of independent learning, frequency of contact with instructor), and the varied student abilities that must be accommodated, flipped classrooms (and the controls they are compared with) look markedly different across educational levels (Altemueller & Lindquist, 2017). Relevant learning outcomes also differ among these groups. Whereas some 21st-century learning outcomes are not germane in primary and secondary education (e.g., commitment to a field of study, professional skills), other skills of interest are not developmentally appropriate or are qualitatively different in lower-level students (e.g., higher-order thinking, metacognition).

The Case for Flipped Learning

Throughout the 20th century, education was primarily focused on teaching students skills identified as lower-order (foundational knowledge) and higher-order thinking skills in Bloom's learning taxonomy (Bloom, 1956). Today, there is a recognition among educators and employers that students of the 21st century need to develop a broader set of skills and abilities. Educational institutions are being called on to teach students to be flexible, work collaboratively in teams, adapt to

changing technology, and be lifelong learners (Hawks, 2014; Kivunja, 2014). The National Academies of Science, Engineering and Medicine (NASEM; 2017) recommends the following intra-/interpersonal skills be included as goals in higher education: ethics, lifelong learning/career orientation, intercultural/diversity competence, civic engagement, teamwork, and communication. In addition to being valuable outcomes in their own right, intra-/interpersonal competencies can provide a fairer assessment of college success across racial groups (NASEM, 2017) and are related to educational achievement (Fong et al. 2017; Martin & Dowson, 2009).

Reflecting these changing educational goals, Fink (2013) updated Bloom's taxonomy of learning to include the categories of application (including professional and academic skills), human dimension (knowledge of oneself and others), caring (developing feelings, values, and interests), and learning to learn (meta-cognitive skills needed to be a self-directed learner). In the present meta-analysis, we evaluate the effectiveness of flipped learning across a broad spectrum of learning outcomes that represent an integration of Bloom's and Fink's taxonomies of learning goals. These learning outcomes fall into two domains. The first domain, overall academics, includes foundational knowledge and higher-order thinking (both based on Bloom's cognitive domain), as well as academic/professional skills (Fink's application outcome). The second domain, overall intra-/interpersonal skills, consists of confidence/interpersonal skills (Fink's human dimension), engagement/identification (Fink's caring dimension), and meta-cognitive skills (Fink's learning to learn). See Table 1 for definitions and examples of each of these learning outcomes and Supplemental Table S1 (available in the online version of this article) for examples of representative measures used to assess intra-/interpersonal outcomes.

In addition to academics and intra-/interpersonal skills, we also examine the impact of flipped learning on a third domain pertinent to evaluating educational interventions: overall student satisfaction. Student satisfaction, as reflected in both course and instructor ratings, is relevant for two key reasons. First, is it related to achievement; students who are more satisfied with the course and professor tend to exhibit better class performance (Wright & Jenkins-Guarnieri, 2012). Second, student evaluations are widely used in higher education as a tool for assessing instructor performance and can have a significant influence on hiring, tenure and promotion decisions, as well as administrative investments (Beran et al., 2007). Thus, even if the flipped classroom is shown to be effective at fostering students' academic and intra-/interpersonal competencies, it is unlikely to be adopted and maintained if it has a consistently negative impact on course or instructor evaluations.

Several theoretical arguments have been advanced to support the position that flipped learning should be more effective than lecture-based learning at achieving the aforementioned outcomes, particularly more complex skills. Most prominent among these are cognitive load theory and constructivist learning theory. Cognitive load theory is an instructional design theory that uses an understanding of basic human cognitive architecture to inform the development of effective teaching methods (Sweller, 2020). According to cognitive load theorists, our ability to learn is constrained by the capacity and duration of our working memory, which

TABLE 1

Learning domains and outcomes

Learning domain	Overall academic			Overall intra-/interpersonal			Overall satisfaction	
Learning outcome	Foundational knowledge	Higher-order thinking	Academic/professional skills	Confidence/ interpersonal skills	Engagement/ identification	Meta-cognitive skills	Course satisfaction	Instructor satisfaction
Definition	Lower-order thinking skills, including recall of knowledge, conceptual understanding	Application of principles, analysis, synthesis, evaluation, creativity	Essential academic and professional skills	Skills necessary to work successfully with others, feelings of competence	Involvement, interest in, or identification with the course or discipline	Skills necessary for effective learning	Evaluation of aspects of the course	Evaluation of aspects of the instructor
Examples	Tests, exams, worksheets, in-class quizzes	Essay tests or exams, in-class exercises, written assignments, course projects, case analysis, work that involves solving problems, creating hypotheses, interpreting data, evaluating outcomes	Assessment of skills specific to a discipline: diagnostic skills, interpreting MRIs, computer programming, using engineering tools, dental wax carving, proficiency speaking or translating a nonnative language general skills: presentation, writing, speaking, library literacy	Collaborative skills, perspective taking, cultural awareness, respect for the opinion of others, social connectedness, efficacy, empowerment, feelings of competence, ethics	Course attendance, effort, attitude, enjoyment, value and meaningfulness of the discipline, identification with a discipline	Self-direction, self-regulation, learning strategies, time management, completion of assignments	Overall satisfaction with course, satisfaction with assignments, satisfaction with course structure, satisfaction with assessment	Overall satisfaction with instructor, instructor preparation, instructor responsiveness, instructor knowledge

Note. MRI = magnetic resonance imaging.

can only hold and manipulate a limited amount of *novel* information at a time. However, when processing information that has been retrieved from long-term memory, working memory has no known limits (Sweller, 2020; Van Merriënboer & Sweller, 2005). The flipped classroom optimizes learning opportunities, given this cognitive architecture. Novel information is taught online at an individualized pace, allowing students to integrate it into long-term memory before class. During class, students can then transfer the newly integrated information from long-term to working memory and use it to form more complex ideas, draw conclusions, and make connections. Thus, not only does flipped learning reduce cognitive load during initial instruction (theoretically leading to better retention and understanding of foundational knowledge), but it also affords students greater cognitive capacity to engage in complex learning and skill development during class.

Constructivist learning theory also provides theoretical support for the superiority of flipped over lecture-based learning for developing 21st-century outcomes. A central tenet of constructivist theory is that learning occurs not by passively absorbing information, but through actively constructing one's own knowledge (Tobin & Tippins, 1993). Cognitive constructivists posit that learning occurs when information or experiences interact with a learner's prior knowledge and experiences in ways that build on one's schemas to create new knowledge (Howe & Berv, 2000). Through this process, knowledge becomes increasingly complex as new information and experiences are incorporated into already established constructs. Social constructivists also focus on the cultural and social aspects of this process, emphasizing that knowledge construction does not occur in a vacuum (Windschitl, 2002). Rather, culture provides the context on which personal knowledge is built and knowledge is often co-constructed with others; learning to work with others can thus facilitate the social construction of knowledge.

Acts of knowledge construction can be weak, with fragile and loose connections to only a narrow range of contexts, or strong with connections across topics and knowledge that can form the basis for further connections (Windschitl, 2002). While it is possible to facilitate strong knowledge construction in a lecture format, the flipped classroom's focus on active and collaborative learning should make it particularly well-suited to facilitating "strong" acts of individual and social knowledge construction. By allowing students to integrate foundational information into their schemas before class, flipped class sessions can be devoted to building on this knowledge using student-centered activities shown to foster deep learning and higher-order skill development (Prince, 2004; Roehling, 2018). Unlike homework, these flipped classroom exercises can be guided by the instructor, which tends to lead to greater learning than when conducted independently (Furtak et al., 2012; Lazonder & Harmsen, 2016). Finally, through discussions, group work, and team projects, the flipped classroom provides more abundant opportunities for students to co-create knowledge than can occur in a lecture-based setting.

In summary, both cognitive load theory and constructivist learning theory suggest that the flipped classroom should produce greater learning gains and more complex knowledge and skill development than lecture-based courses. To engage in deep learning, students must have the requisite underlying knowledge structures as well as opportunities to actively manipulate and build on this prior

knowledge—ideally with the aid of peers and mentors. By using vodcasts to teach novel information before class, the flipped classroom capitalizes on working memory's differential capacity for processing new versus previously stored information (Sweller, 2020). This, in turn, provides students the cognitive capacity needed to fully engage and benefit from instructor-guided active learning experiences during class (Windschitl, 2002).

These theoretical arguments are supported by considerable research demonstrating the efficacy of active learning. A meta-analysis of 225 studies in STEM (science, technology, engineering, and mathematics) found that, compared with lecture-based learning, students who engaged in active learning performed better on exams and were less likely to fail the course (Freeman et al., 2014). Other research has touted the benefits of active learning for developing a broad range of higher-order learning outcomes (critical thinking, interpersonal skills, etc.) and improving student satisfaction (Prince, 2004). Moreover, although all students tend to do better with active learning, a growing literature suggests that the benefits may be particularly pronounced for students from underrepresented groups. Compared with their "majority" classmates, racially minoritized and first-generation students reap greater academic gains from active learning interventions, narrowing the achievement gap between these groups (Eddy & Hogan, 2014; Theobald et al., 2020). Underrepresented students also tend to thrive in more interdependent learning environments and in courses with more scaffolding and student-faculty interaction (Ives & Castillo-Montoya, 2020; Sanchez, 2000), all of which characterize the flipped classroom. Thus, in addition to fostering better learning outcomes overall, flipped learning has the potential to be an inclusive pedagogy that levels the playing field for students who have been underrepresented and underserved in higher education.

Previous Research

Although Lage et al. (2000) are believed to have published the first peer-reviewed article on flipped learning, empirical articles on flipped pedagogies were scarce until around 2012. Since then, there has been a dramatic increase in research comparing the efficacy of flipped and lecture-based learning (Al-Shabibi & Al-Ayasra, 2019; Yang et al., 2017). In an attempt to synthesize this growing literature, a number of qualitative reviews have been published in recent years. Overall, these reviews tend to conclude that flipped learning is associated with increased student achievement, motivation, and engagement (O'Flaherty & Phillips, 2015; Zainuddin & Halili, 2016). In the health sciences, four separate reviews have found a positive effect of flipped learning on achievement and professional skills (F. Chen et al., 2017; Gianoni-Capenakas et al., 2019; Hughes & Lyons, 2017; Presti, 2016). Positive conclusions also have been drawn by reviews of flipped English as a foreign language courses (Turan & Akdag-Cimen, 2020) and in flipped learning research conducted in Spain (Galindo-Domínguez & Bezanilla, 2019).

But not all reviews of the flipped classroom have been uniformly positive. A review of flipped learning in engineering courses found that only 13 of the 30 studies comparing students' performance in flipped versus traditional classrooms reported consistently positive effects for the flipped approach (Karabulut-Ilgu

et al., 2018). Kozikoğlu's (2019) review of articles written in Turkish and English also produced equivocal results and Evans et al. (2019) found limited support for the superiority of flipped learning in health care education. Outcomes regarding students' satisfaction with flipped versus lecture-based courses have been similarly mixed (Brewer & Movahedazarhouligh, 2018; Galindo-Domínguez & Bezanilla, 2019).

Unfortunately, the aforementioned reviews suffer from a number of drawbacks that make it difficult to reconcile this conflicting evidence. Many of the reviews included studies that did not employ a control condition. The reviews also relied almost exclusively on published articles, which, due to publication bias (Dickersin & Min, 1993; Kicinski et al., 2015), may overestimate the positive effects of flipped learning. Finally, although narrative reviews can provide suggestive evidence for the efficacy of flipped learning, they cannot quantitatively assess whether flipped classrooms are, overall, significantly more effective than traditional instruction.

Many of the above problems and inconsistencies can be addressed using meta-analysis. To this end, 14 meta-analyses on flipped learning have been published since 2017 (see online Supplemental Table S2 for a summary). However, most of these studies are small ($k < 30$) or limited in scope. Only four meta-analyses have examined the effectiveness of flipped learning across disciplines and regions and the vast majority have focused exclusively on academic achievement.

Previous meta-analyses have painted a largely positive picture regarding the effect of flipped learning on achievement-based outcomes. Across disciplines and educational levels, flipped learning is associated with small to moderate increases in test and homework scores (Cheng et al., 2019; Låg & Sæle, 2019; Shi et al., 2020; van Alten et al., 2019). Meta-analyses focusing on specific regions or disciplines have confirmed the benefits of flipped over traditional learning on achievement in science (K.-S. Chen et al., 2018), mathematics (Lo et al., 2017), health professions (Hew & Lo, 2018), nursing courses in China (Hu et al., 2018; Tan et al., 2017; Xu et al., 2019), and in studies conducted primarily in Turkey (Karagöl & Esen, 2019; Orhan, 2019). Only Gillette et al.'s (2018) meta-analysis of pharmaceutical courses produced a nonsignificant effect of flipped learning on achievement, and this analysis was likely underpowered ($k = 6$).

In contrast, very little can be gleaned from this spate of meta-analyses about the utility of flipped pedagogies for learning outcomes outside of achievement-based measures. Regarding student satisfaction, Låg and Sæle (2019) found a weak positive relationship between flipped learning and satisfaction ($g = 0.16$, $k = 69$), whereas van Alten et al. (2019) found no association ($g = 0.05$, $k = 22$). Beyond this, only three meta-analyses have examined outcomes other than knowledge-based assessments, all of which were focused on nursing courses in China. These meta-analyses provide preliminary evidence that flipped classes are superior to traditional courses at developing students' self-directed learning skills (Liu et al., 2018; Tan et al., 2017) and at teaching skills like higher-order thinking and teamwork (Xu et al., 2019). However, due to the small number of studies included (ks < 16) and their narrow population, the generalizability of these findings is unclear. As a result, the efficacy of flipped learning for fostering 21st-century learning outcomes, arguably its most touted benefit, remains largely unknown.

To address this gap, a major goal of our meta-analysis was to assess whether flipped learning does, indeed, better prepare students to meet a broad range of learning outcomes in the context of higher education. Do students in flipped classrooms display more advanced higher-order thinking, practical application, and intra-/interpersonal skills than those in well-controlled comparison groups? And is the advantage of flipped over lecture-based learning larger for 21st-century learning outcomes than for foundational knowledge? Although numerous studies have examined the effects of flipped learning on outcomes like meta-cognitive skills, critical thinking, and teamwork, our study is the first to synthesize these findings meta-analytically. Examining multiple outcomes in the same study also allowed us to compare the magnitude of flipped learning effects within and across different domains.

Our meta-analysis improves on previous meta-analyses in other ways as well. Låg and Sæle (2019) and van Alten et al. (2019) published the two largest meta-analyses to date, with 272 and 114 studies, respectively. However, their inclusion criteria were considerably more lax than in the present investigation. Whereas we limited our analyses to studies that used the same instructor and assessments in both conditions, Låg and Sæle (2019) and van Alten et al. (2019) included studies that compared flipped and lecture-based courses that were taught by different instructors or given different assessments. We also used a stricter definition of flipped learning than Låg and Sæle (2019) or van Alten et al. (2019), who included studies that did not utilize recorded lectures. As a result, there were only 140 overlapping studies between our meta-analysis and Låg and Sæle's meta-analysis (we excluded a number of their studies and identified an additional 178).

Potential Moderators

The second major objective of our investigation was to identify specific conditions under which flipped learning may be more or less effective relative to traditional pedagogies. Although previous meta-analyses have revealed consistently positive effects of flipped learning on academic achievement, the magnitude of these estimates has varied substantially, ranging from nonsignificant (Gillette et al., 2018) or weak ($g = 0.19$; Cheng et al., 2019) to quite large ($d = 1.39$–1.79; Xu et al., 2019). To date, relatively little is known about the source of this variability or which factors are most important in determining the success of flipped interventions. Most previous meta-analyses have lacked the power needed to adequately test moderators or have examined only a handful of variables. Past investigations also have failed to test and/or identify significant moderators for learning outcomes other than academic achievement. Our study addresses these issues by examining three broad factors that may influence the efficacy of flipped versus lecture-based learning: (a) differences in educational context, (b) differences in course design, and (c) methodological differences between studies.

Contextual Characteristics

One possible source of heterogeneity in the outcomes of flipped learning is differences in the context where the flipped intervention is administered. Academic discipline, for instance, has been forwarded as potentially relevant to the success of flipped models, with some arguing that applied STEM fields

such as engineering or health professions may be particularly well-suited to flipped learning (F. Chen et al., 2017; Karabulut-Ilgu et al., 2018). Subject areas that require practicing specific skills or problem sets (e.g., languages, mathematics) also have been identified as good candidates (Lo et al., 2017; Turan & Akdag-Cimen, 2020). But evidence for these assertions has been mixed. Of the four meta-analyses comparing effect sizes across disciplines, two found no differences (Shi et al., 2020; van Alten et al., 2019), and one found that courses in arts/humanities, natural and social sciences, and mathematics were all more effective than those in engineering (Cheng et al., 2019). Låg and Sæle (2019), who had the most power to detect significant differences, compared four categories of courses and found that flipped humanities courses were more effective than flipped STEM courses. However, their broad classification of STEM and humanities fields did not permit exploration of differences *within* these categories, potentially obscuring important distinctions between more versus less applied disciplines.

A second contextual variable that may influence the efficacy of flipped learning is the level of the course. Although little is known about the effect of course level on flipped learning outcomes, there is reason to believe that flipped pedagogies may be more effective in upper-level and graduate courses than in introductory classes. The goal of many introductory-level courses is to impart foundational knowledge, potentially making these courses better suited to lecture-based teaching. Upper-level and graduate courses, in contrast, are more likely to have learning goals that involve applying and extending the knowledge learned in earlier courses. Students in upper-level or graduate courses may also be more motivated and have better self-regulated learning skills than those in introductory courses, enabling them to benefit more from the flipped model (Lim & Morris, 2009; Wigfield et al., 2011). To date, only one meta-analysis has compared undergraduate with graduate courses (graduate $k = 4$; Cheng et al., 2019), and none have examined flipped learning outcomes in introductory versus advanced undergraduate courses.

Last, the effectiveness of flipped learning may be affected by the culture in which it is applied. Although studies of flipped learning have not directly assessed students' cultural orientations, researchers have identified country-level differences on several cultural dimensions (e.g., uncertainty avoidance, individualism vs. collectivism) that could influence the efficacy of flipped pedagogies (Hofstede, 2001; Shcheglova, 2018). Given its Western origins, flipped learning may be associated with more positive learning outcomes in countries that are culturally similar to the United States—namely, Anglo and European nations where pedagogies involving self-directed learning are more common and (perhaps) more ideologically compatible (Frambach et al., 2012; Phuong-Mai et al., 2005). But it is also possible that the benefits of flipped learning may be more pronounced in countries where active learning is less common and the flipped class is a greater departure from the norm (Middle Eastern and Asian nations; Tan et al., 2017). To date, only one meta-analysis has examined location as a moderator of flipped learning: Karagöl and Esen (2019) found that studies conducted in Turkey demonstrated greater benefits of flipped learning than those conducted in other (mostly North American) countries. However, given the limited number of countries examined, it is unclear if this regional difference can be generalized.

Course Design Characteristics

A second probable source of heterogeneity in flipped learning outcomes is differences in course design. Although flipped learning studies vary wildly in the number and type of pedagogical details they provide (Låg & Sæle, 2019; van Alten et al., 2019), there are a handful of design features reported in most studies that may be related to the efficacy of the approach. One such factor is whether and how students are held accountable for engaging the preclass material. According to cognitive load theory, learning the foundational content *prior* to attending in-class sessions is key to students' ability to engage the material during class (Sweller, 2020). To increase compliance, many flipped instructors hold students accountable for preclass learning through quizzes or other assignments. Prior meta-analyses seem to support the efficacy of this strategy, showing that courses that assess students' preclass preparation yield marginally to significantly larger effect sizes for flipped learning than those that do not (Hew & Lo, 2018; Låg & Sæle, 2019; Lo et al., 2017; van Alten et al., 2019). Interestingly, Hew and Lo (2018) also examined preclass and in-class assessments separately and found that assessing learning at the beginning of class was associated with higher achievement but that preclass assessments had no effect. This provides suggestive evidence that some methods of accountability may be more effective than others, but the pattern needs to be explored in a larger, more diverse sample.

The efficacy of the flipped classroom may also be shaped by the amount of time students spend in class and the duration of the flipped intervention. Although most flipped courses meet the same number of hours as their lecture-based counterparts, some instructors have decreased facetime in the flipped condition to maintain a consistent workload (e.g., Day, 2018; Smallhorn, 2017). Only one meta-analysis has examined time in class as a moderator of flipped learning (van Alten et al., 2019), yielding preliminary evidence that decreasing class time relative to the control reduces the efficacy of the approach. But when it comes to the *duration* of the flipped intervention, it is unclear whether fully flipped classes produce better results than courses that flip only certain components (partially flipped). Despite some evidence that students' compliance with flipped learning increases as they gain experience with the pedagogy (McLaughlin et al., 2014), all five meta-analyses examining this variable have produced null results (Cheng et al., 2019; Karagöl & Esen, 2019; Shi et al., 2020; van Alten et al., 2019; Xu et al., 2019).

Methodological Characteristics

Finally, variability in the effects of flipped learning may be rooted in differences in study methodology. Given evidence that meta-analytic conclusions depend strongly on the quality of the studies included (Borenstein et al., 2009; Cheung & Slavin, 2016), we took several steps to reduce quality-based noise, including restricting our analyses to studies where the instructor and assessments were identical in both conditions. However, our sample still varied on a number of methodological variables, including the presence of peer review, the purity of the comparisons being made, and the extent to which pre-intervention equivalence was established between conditions. Although there is reason to believe such factors may influence effect size estimates (cf. Cheung & Slavin,

2016), prior meta-analyses do not yield a clear picture regarding the prevalence or magnitude of these issues. For instance, despite finding substantial publication bias in their analyses, Låg and Sæle (2019) did not examine publication type as a moderator, and smaller studies looking at this variable have found no effect (Cheng et al., 2019; Orhan, 2019; van Alten et al., 2019). With respect to group equivalence, Låg and Sæle (2019) found that randomized studies produced larger effects than those with unequal or untested groups, but other studies have found no differences (Hew & Lo, 2018; Lo et al., 2017; van Alten et al., 2019). By utilizing a larger, more rigorously selected sample of primary studies, we seek to clarify the influence of these and other methodological variables on flipped learning effect size estimates.

Method

To examine our two overarching research questions, we conducted a systematic review and meta-analysis using standard collection procedures and synthesis techniques (Higgins & Green, 2011; Lipsey & Wilson, 2001). Details of our methodological approach are described in the following sections.

Eligibility Criteria and Search Strategy

Prior to performing any search procedures, we identified seven criteria that studies needed to meet to be included in the present meta-analysis. Specifically, eligible studies had to (a) be implemented at a postsecondary institution, (b) include experimental (flipped classroom) and control (traditional classroom) conditions taught by the *same* instructor, (c) use identical or highly similar outcome measures in each condition, (d) utilize vodcasts and active learning strategies in their flipped methodology, (e) include quantitative data on at least one outcome of interest, (f) provide sufficient data to calculate an effect size, and (g) be published in English. Duplicates were excluded, as were studies where the equivalency of the outcomes could not be verified. We also excluded studies where one condition was taught in-person and the other was taught entirely online (e.g., distance learning), and where conditions were nonequivalent in other ways that may influence the results (e.g., drastically reduced class size; different course levels).

To identify qualifying studies, we conducted an iterative and systematic literature search involving several databases and strategies. First, we searched the following electronic databases: Education Resources Information Center (ERIC), ProQuest Social Sciences, PsycINFO, Web of Science, and Wilson Education. Within each database, we used a combination of search terms to identify articles related to the flipped classroom ([flip* OR invert*] AND [class* OR learn]). To increase our probability of finding unpublished studies, we then expanded our search to Web of Science Conference Proceedings, ResearchGate, and ProQuest Dissertations and Theses. Finally, we searched the reference lists of relevant primary and review studies to retrieve articles not captured in our database searches. This procedure was completed during 2018 and 2019, with the last search occurring in May 2019. Altogether, the search process uncovered a total of 10,702 potentially relevant records.

In the initial screening stage, we reviewed the title and abstract of each citation identified and excluded studies that were clearly irrelevant ($n = 9{,}896$). A majority

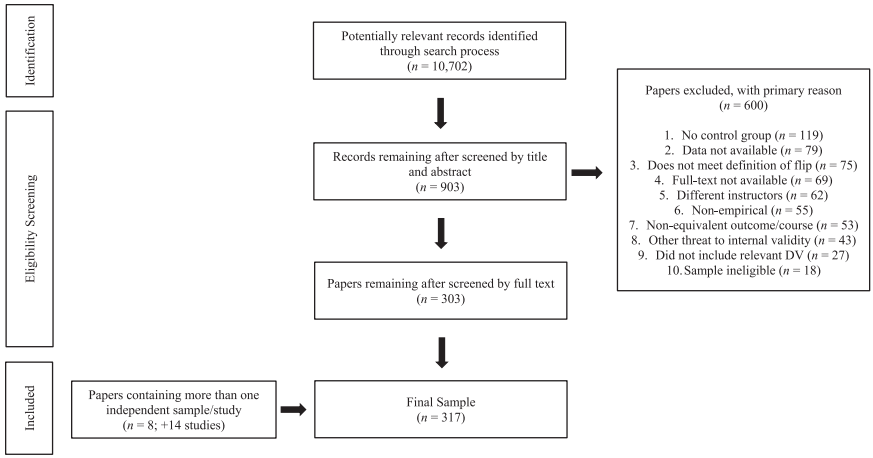

FIGURE 1. *Flow of study selection through search and screening process.*
Note. Exemplars of excluded studies, by criterion, are available in online Supplemental Table S3.

of these studies were unrelated to flipped learning, were nonempirical, were duplicates, and/or were not in higher education. Articles that passed this preliminary screen or whose abstracts lacked sufficient information to make a determination were retrieved for a more detailed evaluation ($n = 903$). Seventy records (mostly conference papers) were excluded at this step because the full-text could not be retrieved ($n = 56$) or they were not available in English ($n = 13$). On review of the full-text papers, we excluded an additional 531 studies that did not meet the inclusion criteria. The most common reason for exclusion was the lack of a control group ($n = 119$), followed by insufficient data for effect size calculation ($n = 79$) and failing to meet our definition of flipped learning ($n = 75$). Of the 303 reports remaining, eight examined learning outcomes in more than one group (adding a net of 14 studies), yielding a total of 317 independent studies (see Figure 1).

Coding Process and Framework

A total of four researchers participated in the coding process, with at least two authors independently coding each of the 317 studies. The final coding scheme included: (a) source descriptors (author names, publication date, manuscript type), (b) type of outcome (foundational knowledge, higher-order thinking, academic/professional skills, confidence/interpersonal skills, engagement/identification, meta-cognitive skills, course satisfaction, instructor evaluations),[1] (c) contextual characteristics (discipline, class level, country), (d) design characteristics (time in class, accountability assignments, duration of intervention), and (e) methodological variables (design type, pre-intervention equivalency, outcome dilution). We also recorded the data needed to compute effect sizes for each outcome. All discrepancies were resolved through consensus. The coders' reliability, which was found by dividing the number of observations agreed on by the total number of observations, was 95%.

Analytic Strategy

Effect Size Calculation and Data Synthesis

Following procedures outlined by Lipsey and Wilson (2001), we extracted effect sizes and variances from each study using relevant summary statistics. Prior to calculating effect sizes, we averaged the scores of two or more groups in each condition when applicable (Cortina & Nouri, 2000). In most cases, studies reported posttest means and standard deviations for the flipped and traditional groups, thus a standardized mean difference effect size was calculated. Because differential sample size across studies can bias effect sizes obtained from unadjusted standardized mean differences (i.e., Cohen's *d*), we used Hedges's *g* to adjust effect sizes to correct bias arising from small samples (Hedges, 1981; Hedges & Olkin, 1985). For studies with missing summary data, we calculated the standardized mean difference from *t* values, *F* ratios, or chi-squared values and converted them to Hedges's *g* (Borenstein et al., 2009). Several studies presented both posttest comparisons and pretest-posttest results; when available we chose calculations that accounted for pretest measures. Although none of the studies reported pre-post correlations (which can influence effect sizes for pretest-posttest designs; Dunlap et al., 1996), we were able to obtain these data from five authors and calculate an average pre-post correlation ($r = .53$) to use in studies where these data were missing. Apart from pretest adjustments, we used raw scores to compute effect sizes, requesting unadjusted estimates from the authors when possible.

When studies included multiple measures of a single outcome (e.g., several exams testing foundational knowledge), we aggregated the effect sizes to provide a single estimate of the effect (Gleser & Olkin, 2009). For studies that included multiple outcomes of interest, we used a shifting unit of analysis approach (Cooper, 2010), which involves averaging effects within a study when appropriate (e.g., knowledge and higher-order thinking when examining overall academics) and splitting effects when testing variables on which they differ (e.g., comparing the effects of different academic outcomes). This allowed us to retain as many data points as possible without violating the assumption that effects are independent. In addition to calculating separate effect sizes for each outcome, we computed three combined effects capturing overall academics (foundational knowledge, higher-order thinking, academic/professional skills), overall intra-/interpersonal (confidence/interpersonal, engagement/identification, meta-cognitive), and overall satisfaction (course satisfaction, instructor evaluations). When averaging effect sizes within a study, we assumed a correlation of .70 between nonindependent measures.

Given the broad scope of our review, we used a random-effects model to compute the effects of flipped learning on our outcomes of interest. Unlike the fixed-effects model, which assumes that each study is an approximation of a single true population parameter, the random-effects model assumes that studies are drawn from a distribution of effect sizes and that the true effect varies from study to study (Borenstein et al., 2009; Hedges & Vevea, 1998). A random-effects model also allows for broader generalization of the results and is more appropriate when methodological and contextual variability are high. To account for sample size,

we weighted the effect sizes by multiplying each effect by its inverse variance (Borenstein et al., 2009). All effect size integration and moderator analyses were conducted using Comprehensive Meta-Analysis.

Moderator Analyses

To better understand the scope of the intervention effects, we conducted a series of analyses exploring effect size heterogeneity. First, we computed effect size variability for each synthesis using Q, τ^2, and I^2. Whereas the Q statistic provides a test of *whether* average weighted effect sizes vary more than would be expected based on sampling error alone, τ^2 and I^2 estimate the *amount* of true heterogeneity present (Borenstein et al., 2009). Specifically, τ^2 is an estimate of the absolute amount of variation in the true effects (i.e., the actual range of dispersion around the population mean) and I^2 reflects the proportion of the observed variance that is due to true differences in effect size (more than 75% is considered high; Higgins et al., 2003).

When heterogeneity in average effect sizes exceeded sampling error alone (as indicated by a statistically significant Q), we conducted moderation analyses to determine if the excess variability could be accounted for by substantive or methodological differences between studies. A total of nine categorical moderators were examined, which include the following.

Discipline

Courses were categorized into one of eight disciplines: (a) engineering, (b) health sciences, (c) humanities, (d) foreign languages, (e) mathematics, (f) natural sciences, (g) social sciences, and (h) technology. Classes that straddled two or more categories (e.g., Math for Engineering) were placed in the group that was most appropriate based on the content covered.

Class Level

Courses were designated as introductory (entry-level undergraduate course; e.g., Introduction to Psychology), upper-level (advanced undergraduate course; e.g., Calculus II), or graduate (any course in a post-undergraduate program; e.g., Clinical Epidemiology).

Location

Given the large number of countries represented, studies were categorized into five groups based on regional and cultural similarities: (a) Asia, (b) Europe, (c) Latin America, (d) Middle East, and (e) North America/Australia (Gupta et al., 2002).

Accountability

Studies were categorized based on whether students were required to verify their engagement with the preclass material and, if so, whether the accountability took the form of preclass assignments (e.g., notes checked in class, worksheets), preclass quizzes, or in-class quizzes. When information about accountability was not addressed, and could not be obtained from the authors, the variable was coded as missing.

Time in Class

Although studies varied in the absolute amount of time spent in class each week, we were primarily interested in whether in-class time was equivalent for the flipped and lecture-based groups. Accordingly, we categorized studies as either "equivalent" (same amount of time in both conditions) or "less in flipped" (flipped group spent less time in class than the traditional group). There were only two instances in which the flipped group spent more time in class than the control, thus these two studies were excluded from this subset of analyses.

Duration

To capture differences in the duration of the flipped interventions, we created a dichotomous measure indicating whether studies utilized flipped instruction the vast majority of days for the entire term (fully flipped) or whether flipped methodologies were only used in a portion of the course (e.g., flipped some units but not others, flipped some days a week while teaching the other days traditionally, partially flipped).

Publication Type

Studies were categorized as one of four publication types: (a) peer-reviewed journal articles, (b) dissertations, (c) conference presentations, or (d) book chapters. Only one master's thesis was eligible and thus was included in the dissertation category.

Outcome Dilution

Some studies compared traditional courses with courses where only one or two units were flipped. Although most of these studies used outcomes specific to the flipped unit (e.g., unit tests), others compared the flipped and control groups on broader measures, such as final exam scores or overall course evaluations. Outcomes in the latter group are considered "diluted" because they provide a weaker (more watered down) test of the intervention effects than outcomes that were only influenced by one of the two methodologies (flipped or traditional).

Pre-Intervention Equivalence

Outcomes were divided into four categories based on the extent to which there was evidence of pre-intervention equivalence between the two conditions. These groups included (a) nonequivalent, (b) no information, (c) general equivalence (equivalence on GPA, SAT scores, etc.), and (d) outcome-specific equivalence (equivalence on the outcome of interest). Outcomes were only classified as nonequivalent when pre-intervention differences were *not* accounted for statistically (e.g., measures utilizing pre-post change scores were categorized as having outcome-specific equivalence regardless of pretest equivalence).

For each moderator, between-class homogeneity analyses (Q_b) were run to determine whether the average effect sizes of different categories of the moderator (e.g., effects estimated separately by discipline) vary more than would be predicted based on sampling error alone. Following these analyses, we also ran mixed-model meta-regressions where moderators were entered simultaneously as predictors of each outcome. In addition to limiting the number of statistical tests

conducted (and thus minimizing the risk of type I error; Polanin & Pigott, 2015), meta-regression addresses the tendency for study characteristics to correlate (e.g., studies from certain disciplines being more common in certain countries), which can potentially confound univariate relationships (Lipsey & Wilson, 2001).

Results

Descriptive Data

A total of 317 studies, yielding effect sizes for 614 outcomes, were included in our meta-analysis. Participants were divided fairly evenly between conditions, with 26,114 receiving the flipped learning treatment and 25,323 students participating in lecture-based control groups. Although our analyses involved a wide range of courses, a majority of the studies were drawn from STEM-related disciplines (70.0%) and used undergraduate populations (85.2%). Studies were conducted in 42 different countries, with a majority coming from North America/Australia (61.5%), followed by Asia (17.4%), and the Middle East (12.6%). Almost one quarter (22.1%) of the included reports were unpublished. The earliest study meeting our criteria was published 9 years ago (S. A. Jensen, 2011), and only six well-controlled studies of flipped learning were published in 2011–2012 (2.0% of the sample). In contrast, 41 eligible reports were produced in 2013–2014 (13.5%), 96 in 2015–2016 (31.7%), 115 in 2017–2018 (38.0%), and 44 in the first 5 months of 2019 (14.5%). Online Supplemental Table S4 provides a list of the studies included in the meta-analysis and their key characteristics; full references for all 317 studies also are available online.

Overall Effect Sizes for Learning Domains and Outcomes

Our first research question addressed the extent to which flipped learning is more effective than lecture-based learning across a range of pertinent outcomes. Table 2 presents the overall synthesis results for the three learning domains and eight specific outcomes examined. Consistent with our expectation that flipped learning would generally have a positive influence on academic performance, intra- and interpersonal development, and satisfaction, we found significant benefits for flipped over lecture-based learning on all three composites and seven of the eight individual outcomes. Below, we discuss the results for each learning domain in turn.

Academic Performance

Using a random-effect model to integrate the results of 282 studies, the mean effect size for overall academic performance was 0.39. According to Cohen (1988), effects of 0.80, 0.50, and 0.20 can be (tentatively) considered large, medium, and small, respectively. Thus, our findings suggest that flipped learning produces small to moderate academic gains relative to lecture-based learning. However, this effect was not uniform across the three outcomes tested (Q_b = 14.02, p = .001). Whereas a medium effect of 0.53 was found for outcomes involving varied academic and professional skills, effect sizes for foundational knowledge and higher-order thinking were significantly smaller at 0.34 and 0.20, respectively. The effect for higher-order thinking was, in turn, significantly smaller than the effect for foundational knowledge. Heterogeneity analyses

TABLE 2

Overall synthesis results for learning domains and specific outcomes

Outcome	*k*	*g*	95% CI	*Q* (*p*)	I^2, τ^2
Overall academic	282	0.39***	[0.34, 0.43]	1813.60 (.000)	84.51, .13
Foundational knowledge	234	0.34***	[0.29, 0.39]	1406.63 (.000)	83.44, .11
Higher-order thinking	43	0.20***	[0.11, 0.30]	170.56 (.000)	75.38, .08
Academic/professional skills	75	0.53***	[0.39, 0.68]	820.05 (.000)	90.98, .35
Overall intra-/interpersonal	96	0.43***	[0.33, 0.53]	867.14 (.000)	89.05, .22
Meta-cognitive skills	34	0.37***	[0.21, 0.53]	190.54 (.000)	82.68, .17
Confidence/interpersonal	40	0.52***	[0.35, 0.68]	272.79 (.000)	85.70, .24
Engagement/identification	65	0.41***	[0.27, 0.54]	714.90 (.000)	91.05, .26
Overall satisfaction	66	0.22**	[0.09, 0.36]	679.53 (.000)	90.44, .25
Instructor evaluations	29	0.16	[−0.05, 0.37]	238.99 (.000)	88.28, .28
Course satisfaction	59	0.23**	[0.09, 0.37]	637.24 (.000)	90.90, .26

Note. k = number of studies; g = average weighted effect size; CI = confidence interval; Q is the amount of variance not accounted for by sampling error; I^2 and τ^2 are measures of effect size variability.

**$p < .01$. ***$p < .001$.

revealed significant between-study variance in both overall academics and the individual outcomes, with I^2 values exceeding 75% and τ^2 values ranging from 0.08 to 0.35. Taken together, these statistics indicate considerable systematic variability in effect sizes.

Intra-/Interpersonal Skills

The overall average effect of the 96 studies that contributed outcomes pertaining to intra- and interpersonal competencies was 0.43. Flipped learning also had a significant and positive influence on all three specific outcomes in this domain, with effects of 0.37 for meta-cognitive skills, 0.41 for engagement/identification, and 0.52 for confidence/interpersonal skills. There was not a difference in the strength of these effects ($Q_b = 1.74$, $p = .420$). Heterogeneity in the effect estimates exceeded sampling error for all four variables, and I^2 values were consistently over 75% (τ^2: .17–.26), indicating a large amount of true between-study variance in effect sizes.

Satisfaction

Sixty-six studies contributed effects for overall satisfaction, yielding a small, but significant, average effect of 0.22. Although the effect sizes for instructor evaluations and course satisfaction did not differ ($Q_b = 0.30$, $p = .585$), only the effect for course satisfaction was statistically significant ($g = 0.23$). Heterogeneity tests were significant for both overall satisfaction and the individual outcomes and I^2 and τ^2 values were high for all three variables, confirming a large and significant amount of true variability in the effect estimates.

Moderation Analyses

Our second research question addressed the extent to which heterogeneity in the efficacy of flipped learning can be explained by a number of theoretically and methodologically relevant factors. To enhance the reliability of our analyses and maximize our power to detect potentially small moderation effects, we focused our subgroup and meta-regression analyses on the three broad learning domains. We also limited our analyses to subgroups with a minimum of four studies to reduce unreliable estimation due to small sample sizes (Fu et al., 2011). Tables 3 and 4 present the results of the subtype analyses and meta-regression models separately by domain.

Overall Academic

Given sufficient between-study heterogeneity in the effect sizes for overall academics, we proceeded with subgroup analyses to identify conditions where flipped learning was more or less effective than lecture-based learning. Beginning with educational context, discipline and location significantly moderated, and class level marginally moderated, the effect of flipped learning on overall academic performance. Although estimated effects were statistically greater than zero for all fields except humanities, the strength of these effects ranged from a small mean effect of 0.19 for engineering to a large average effect of 0.76 for languages. Post hoc analyses revealed that flipped learning effects were significantly larger in language courses than in other disciplines (except technology), and that effect sizes were larger in technology and health science courses than in mathematics and engineering. Social and natural sciences also produced larger mean effects than engineering. For location, flipped learning had a positive and significant effect on academic outcomes in four out of five regions (all but Latin America), but there were marked differences in the size of these effects. Effect estimates for studies conducted in Middle Eastern, Asian, and European countries were significantly larger than for studies conducted in North America/Australia and Latin America. Effect sizes from Middle Eastern countries also were higher than those from Europe, whereas effects in North America/Australia and Latin America did not differ. Although differences in the efficacy of flipped learning across class levels did not meet conventional levels of significance ($p = .067$), follow-up pairwise comparisons revealed that effects were significantly larger in upper-level versus introductory undergraduate courses.

For characteristics related to course design, both accountability and duration of treatment significantly moderated the effect of flipped learning on overall academics. Counter to our predictions, the positive effect of flipped learning was actually *weaker* among students who were held accountable for class preparation via preclass assignments or quizzes than for those who were not held accountable at all. No differences were found between courses using in-class quizzes and those with no accountability, but classes using in-class quizzes were significantly more effective than those using preclass quizzes. Partially flipped courses demonstrated greater academic gains relative to lecture-based controls than did courses that were fully flipped. Time in class did not significantly moderate the efficacy of flipped learning on academic outcomes.

TABLE 3

Subgroup analyses for proposed moderators

	Overall academic				Overall intra-/interpersonal				Overall satisfaction			
Moderator	*k*	*g*	95% CI	Q_b (*p*)	*k*	*g*	95% CI	Q_b (*p*)	*k*	*g*	95% CI	Q_b (*p*)
Discipline				42.62 (.000)				24.29 (.000)				10.21 (.069)
Engineering	46	0.19***	[0.11, 0.27]		16	0.09	[−0.07, 0.24]		15	0.17	[−0.09, 0.43]	
Health sciences	46	0.45***	[0.31, 0.60]		12	0.42*	[0.08, 0.76]		5	0.44	[−0.52, 1.41]	
Humanities	8	0.18	[−0.19, 0.55]		0				1			
Language	35	0.76***	[0.57, 0.95]		6	1.10***	[0.59, 1.60]		1			
Mathematics	36	0.23***	[0.13, 0.33]		17	0.33**	[0.14, 0.52]		10	0.53***	[0.34, 0.72]	
Natural sciences	47	0.35***	[0.26, 0.44]		14	0.41**	[0.15, 0.67]		13	0.12	[−0.12, 0.35]	
Social sciences	41	0.36***	[0.24, 0.49]		21	0.53***	[0.27, 0.78]		17	0.18	[−0.05, 0.40]	
Technology	23	0.52***	[0.33, 0.72]		10	0.64***	[0.30, 0.97]		4	0.09	[−0.47, 0.65]	
Class level				5.41 (.067)				3.05 (.217)				2.02 (.364)
Introductory	109	0.33***	[0.26, 0.39]		32	0.33***	[0.19, 0.46]		22	0.16*	[0.00, 0.32]	
Upper level	131	0.45***	[0.37, 0.53]		53	0.52***	[0.35, 0.69]		35	0.20^{+}	[−0.11, 0.46]	
Graduate	42	0.36***	[0.23, 0.49]		11	0.37*	[0.07, 0.66]		9	0.57*	[0.03, 1.12]	
Location				68.36 (.000)				29.68 (.000)				11.52 (.003)
Asia	50	0.66***	[0.52, 0.80]		22	0.80***	[0.59, 1.01]		9	0.67**	[0.25, 1.09]	
Europe	16	0.48***	[0.26, 0.71]		9	0.17*	[0.00, 0.34]		9	0.51**	[0.20, 0.83]	
Latin America	5	0.06	[−0.35, 0.46]		3				1			
Middle East	35	0.80***	[0.63, 0.97]		16	0.75**	[0.32, 1.19]		1			
North America/Australia	176	0.24***	[0.19, 0.29]		46	0.21**	[0.10, 0.33]		46	0.07	[−0.07, 0.22]	
Accountability				12.49 (.014)				2.64 (.620)				4.88 (.300)
None	60	0.52***	[0.39, 0.65]		24	0.36***	[0.16, 0.55]		20	0.25	[−0.07, 0.56]	
Preclass assignment	79	0.35***	[0.26, 0.44]		27	0.41***	[0.24, 0.59]		15	0.05	[−0.11, 0.22]	
Preclass quiz	71	0.28***	[0.20, 0.36]		23	0.40**	[0.12, 0.68]		20	0.31**	[0.09, 0.54]	
In-class quiz	49	0.46***	[0.35, 0.58]		10	0.49**	[0.19, 0.79]		7	0.36^{+}	[−0.03, 0.75]	
Missing	22	0.36***	[0.16, 0.55]		12	0.63***	[0.34, 0.92]		4	0.09	[−0.24, 0.41]	

(continued)

TABLE 3 (continued)

Moderator	Overall academic				Overall intra-/interpersonal				Overall satisfaction			
	k	*g*	95% CI	Q_b (*p*)	*k*	*g*	95% CI	Q_b (*p*)	*k*	*g*	95% CI	Q_b (*p*)
Time in class				2.02 (.155)				3.18 (.074)				0.23 (.632)
Equivalent	249	0.39***	[0.34, 0.45]		80	0.47***	[0.35, 0.59]		56	0.24**	[0.09, 0.38]	
Less in flipped	31	0.30***	[0.18, 0.42]		16	0.24*	[0.01, 0.47]		10	0.15	[−0.20, 0.49]	
Duration				4.58 (.032)				1.67 (.197)				1.59 (.208)
Full flip	191	0.35***	[0.29, 0.40]		70	0.38***	[0.27, 0.49]		47	0.17*	[0.03, 0.31]	
Partial flip	91	0.47***	[0.37, 0.57]		26	0.55***	[0.32, 0.78]		19	0.39*	[0.08, 0.71]	
Publication type				16.94 (.001)				13.39 (.001)				3.99 (.046)
Peer-reviewed	212	0.42***	[0.36, 0.47]		80	0.48***	[0.36, 0.60]		56	0.26*	[0.11, 0.41]	
Dissertation	14	0.14*	[0.02, 0.26]		6	0.08	[−0.29, 0.46]		2			
Conference	50	0.34***	[0.22, 0.45]		10	0.17**	[0.04, 0.30]		8	0.05	[−0.08, 0.19]	
Book chapter	6	0.29***	[0.10, 0.48]		0				0			
Diluted				1.77 (.184)				0.14 (.712)				0.28 (.599)
No	260	0.39***	[0.34, 0.44]		88	0.43***	[0.32, 0.54]		58	0.21**	[0.07, 0.35]	
Yes	22	0.30***	[0.17, 0.43]		8	0.38**	[0.10, 0.66]		8	0.33	[−0.09, 0.74]	
Group equivalence				7.59 (.055)				4.54 (.104)				1.80 (.408)
Nonequivalent	13	0.61**	[0.25, 0.98]		2				0			
No information	120	0.31***	[0.24, 0.37]		38	0.55***	[0.36, 0.74]		38	0.29**	[0.11, 0.46]	
General equivalence	41	0.35***	[0.20, 0.49]		26	0.41***	[0.21, 0.60]		24	0.20$^+$	[−0.00, 0.41]	
Outcome-specific equivalence	118	0.43***	[0.35, 0.51]		29	0.30***	[0.16, 0.44]		4	−0.18	[−0.88, 0.51]	

Note. *k* = number of studies; *g* = average weighted effect size; CI = confidence interval. Q_b is the amount of between-study variance not accounted for by sampling error.
$^+p < .10$. $^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

TABLE 4

Meta-regression models for proposed moderators

	Overall academic				Overall intra-/interpersonal				Overall satisfaction			
Moderator	Estimate (*SE*)	95% CI	*Q* (*df*)	*p*	Estimate (*SE*)	95% CI	*Q*(*df*)	*p*	Estimate (*SE*)	95% CI	*Q* (*df*)	*p*
Intercept	0.24 (0.10)	[0.04, 0.44]		.017	−0.20 (0.17)	[−0.54, 0.15]		.261	0.02 (0.20)	[−0.38, 0.41]		.939
Discipline			16.33 (7)	.022			7.43 (6)	.283			7.13 (5)	.211
Engineering (ref.)												
Health sciences	0.26 (0.12)	[0.03, 0.50]		.030	0.31 (0.27)	[–0.23, 0.84]		.255	−0.00 (0.35)	[–0.70, 0.69]		.993
Humanities	−0.04 (0.16)	[−0.36, 0.28]		.810								
Language	0.28 (0.11)	[0.07, 0.49]		.010	0.67 (0.28)	[0.12, 1.23]		.017				
Mathematics	0.06 (0.10)	[−0.14, 0.25]		.563	0.30 (0.19)	[−0.07, 0.67]		.113	0.31 (0.23)	[−0.14, 0.77]		.173
Natural sciences	0.17 (0.09)	[0.00, 0.34]		.046	0.40 (0.19)	[0.02, 0.77]		.037	0.06 (0.22)	[−0.38, 0.50]		.784
Social sciences	0.01 (0.09)	[−0.17, 0.19]		.917	0.31 (0.19)	[−0.05, 0.67]		.095	−0.15 (0.21)	[−0.56, 0.26]		.472
Technology	0.02 (0.11)	[−0.20, 0.25]		.849	0.33 (0.24)	[−0.15, 0.80]		.184	−0.45 (0.33)	[−1.10, 0.21]		.181
Class level			2.73 (2)	.255			4.13 (2)	.127			0.38 (2)	.826
Introductory (ref.)												
Upper level	0.04 (0.06)	[−0.07, 0.15]		.466	0.25 (0.12)	[0.01, 0.48]		.043	0.08 (0.16)	[−0.22, 0.38]		.605
Graduate	−0.13 (0.11)	[−0.34, 0.08]		.217	0.13 (0.26)	[−0.38, 0.64]		.611	0.13 (0.27)	[−0.39, 0.66]		.617
Location			44.16 (4)	.000			10.57 (3)	.014			13.29 (2)	.001
North America/Australia (ref.)												
Asia	0.33 (0.08)	[0.17, 0.48]		.000	0.46 (0.16)	[0.14, 0.78]		.005	0.65 (0.23)	[0.20, 1.11]		.005
Europe	0.27 (0.11)	[0.05, 0.48]		.016	0.13 (0.18)	[−0.24, 0.49]		.498	0.61 (0.23)	[0.17, 1.06]		.007
Latin America	−0.25 (0.18)	[−0.59, 0.10]		.168								
Middle East	0.52 (0.09)	[0.35, 0.69]		.000	0.42 (0.17)	[0.09, 0.74]		.012				
						Model:	38.05 (11)	.000		Model:	19.67 (9)	.020
[Intercept]					0.35 (0.12)	[0.11, 0.58]		.004	0.21 (0.15)	[−0.08, 0.50]		.153
Accountability			9.22 (4)	.056			1.55 (4)	.818			2.72 (4)	.606

(continued)

TABLE 4 (continued)

Moderator	Overall academic				Overall intra-/interpersonal				Overall satisfaction			
	Estimate (*SE*)	95% CI	*Q* (*df*)	*p*	Estimate (*SE*)	95% CI	*Q*(*df*)	*p*	Estimate (*SE*)	95% CI	*Q* (*df*)	*p*
No (ref.)												
Preclass assignment	−0.18 (0.07)	[−0.32, −0.05]		.009	0.06 (0.15)	[−0.23, 0.36]		.667	−0.23 (0.20)	[−0.63, 0.17]		.256
Preclass quiz	−0.20 (0.07)	[−0.34, −0.05]		.008	0.07 (0.16)	[−0.24, 0.38]		.674	0.06 (0.19)	[−0.31, 0.43]		.753
In-class quiz	−0.09 (0.08)	[−0.25, 0.06]		.226	0.09 (0.20)	[−0.31, 0.49]		.652	0.11 (0.27)	[−0.42, 0.63]		.694
Missing	−0.10 (0.11)	[−0.31, 0.11]		.342	0.24 (0.20)	[−0.14, 0.63]		.216	−0.17 (0.33)	[−0.82, 0.48]		.612
Time in class: Less	−0.07 (0.08)	[−0.22, 0.08]		.379	−0.20 (0.15)	[−0.49, 0.10]		.190	−0.07 (0.20)	[−0.47, 0.32]		.716
Duration: Full flip	−0.11 (0.06)	[−0.22, 0.01]		.074	−0.16 (0.12)	[−0.40, 0.08]		.194	−0.22 (0.16)	[−0.53, 0.10]		.181
						Model:	5.79 (6)	.447		Model:	4.52 (6)	.607
[Intercept]					0.58 (0.09)	[0.41, 0.75]		.000	0.31 (0.10)	[0.11, 0.50]		.003
Publication type			1.92 (3)	.590			3.88 (2)	.144				
Peer review (ref.)												
Dissertation	−0.16 (0.12)	[−0.39, 0.07]		.176	−0.36 (0.24)	[−0.82, 0.11]		.137				
Conference	−0.03 (0.07)	[−0.17, 0.10]		.649	−0.25 (0.18)	[−0.60, 0.10]		.159	−0.14 (0.22)	[−0.56, 0.28]		.508
Book chapter	−0.02 (0.16)	[−0.34, 0.31]		.923								
Diluted outcome	−0.10 (0.10)	[−0.30, 0.09]		.282	−0.09 (0.20)	[−0.47, 0.29]		.634	0.06 (0.21)	[−0.36, 0.47]		.782
Group equivalence			4.00 (3)	.406			2.44 (2)	.295			2.73 (2)	.255
No information (ref.)												
Nonequivalent	0.18 (0.13)	[−0.07, 0.44]		.158								
General equivalence	0.11 (0.08)	[−0.04, 0.25]		.161	−0.13 (0.13)	[−0.39, 0.13]		.310	−0.08 (0.15)	[−0.36, 0.21]		.607
Outcome-specific equivalence	−0.01 (0.06)	[−0.12, 0.11]		.906	−0.19 (0.13)	[−0.45, 0.06]		.137	−0.48 (0.29)	[−1.04, 0.09]		.100
		Model:	123.24 (27)	.000		Model:	6.97 (5)	.223		Model:	3.52 (4)	.475

Note. CI = confidence interval; *SE* = standard error; *df* = degrees of freedom; ref. = reference.

For methodological characteristics, publication type significantly moderated the impact of flipped learning on academic outcomes and pre-intervention equivalence of the conditions had a marginal effect. Although an advantage for flipped over lecture-based learning was found for all four publication types, the effect sizes for peer-reviewed journals ($g = 0.42$) and conference papers ($g = 0.34$) were significantly larger than the trivial effect for dissertations ($g = 0.14$). For pre-intervention equivalence, post hoc analyses indicated that the mean effect produced by studies documenting pre-intervention differences between conditions did not differ significantly from that of studies in the other three groups. However, the effect of flipped learning on academics *was* significantly larger for studies that established the outcome-specific equivalence of the two groups than for studies that did not provide any information about equivalence ($g = 0.43$ vs. 0.31). Using outcomes that were "diluted" (e.g., flipping only a portion of the class, but comparing final exam scores) did not significantly influence the magnitude of the effect sizes.

Given the tendency of study characteristics to correlate with one another (Lipsey & Wilson, 2001), we next estimated the joint impact of the proposed moderators using mixed-level meta-regression (see Table 4). When all nine moderators were considered simultaneously, four made a unique contribution to predicting variability in effect sizes. Whereas publication type was no longer significant, discipline and location remained significant predictors and accountability and intervention duration were marginal. Taken together, these variables accounted for 23% of the variance in effect sizes for academic outcomes, but a significant amount of true variation remained unexplained ($I^2 = 84.55\%$; $\tau^2 = .13$)

Overall Intra-/Interpersonal Skills

After verifying that there was adequate heterogeneity in effect sizes for overall intra-/interpersonal skills, we conducted subgroup analyses to identify study attributes that may help explain this variability. For educational context, both discipline and location significantly moderated the effect of flipped learning on intra-/interpersonal outcomes. The influence of flipped learning was positive and significant for every discipline except engineering, but these effect sizes varied considerably, ranging from 0.33 for mathematics to 1.10 for foreign language courses. Post hoc analyses revealed that the benefits of flipped learning for intra-/interpersonal skills were more pronounced in language classes than in social science, health science, natural science, mathematics, and engineering courses. Effect sizes in technology, social science, and natural science also were significantly larger than those in engineering. For location, positive effects of flipped learning were observed in all four regions that could be estimated (Latin America had too few studies), but gains were significantly larger in studies conducted in Asia and the Middle East than in Europe or North America/Australia. Class level did not significantly moderate the efficacy of flipped learning for intra-/interpersonal outcomes.

In contrast to contextual moderators, only one of the three design-related variables had even a marginal influence on the efficacy of flipped learning for intra-/interpersonal outcomes. Studies where students in the flipped and traditional

conditions spent an equivalent amount of time in class demonstrated marginally larger effects for intra-/interpersonal skill development than studies where the flipped condition met for fewer hours ($p = .074$). Neither the presence of formal accountability for watching vodcasts prior to class nor the duration of the flipped learning intervention moderated the effect of flipped learning on overall intra-/interpersonal skills.

For methodological characteristics, only publication type significantly moderated the effect of flipped learning on intra-/interpersonal outcomes. Whereas a medium-sized benefit of flipped learning was found for studies published in peer-reviewed journals ($g = 0.48$), the effect of flipped learning was trivial (albeit still significant) for conference papers ($g = 0.17$) and close to zero for dissertations ($g = 0.08$). Post hoc analyses confirmed that flipped learning effects were significantly larger for peer-reviewed articles than for the other two publication sources. Neither the pre-intervention equivalence of the flipped and lecture conditions nor the dilution of the outcome variable influenced the magnitude of the effects for intra-/interpersonal skills.

Although there is not a universally accepted minimum number of studies required for meta-regression, the *Cochrane Handbook* recommends a minimum of 10 studies for each study-level variable (Higgins & Green, 2011). Applying this benchmark, we would need a total of 220 studies to run a meta-regression with all nine moderators, which is substantially greater than the number of studies reporting intra-/interpersonal outcomes ($n = 96$). Accordingly, we attempted to balance the risk of type I error (multiple univariate models) and type II error (an underpowered multivariate model) by conducting separate meta-regressions for the three types of moderators. For educational context, only study location predicted significant variability in effect sizes for intra-/interpersonal skills net of the other moderators, but several *individual* dummy variables made a significant contribution (e.g., natural science and language vs. engineering; $R^2 = .12$). The models for design-related and methodological moderators were both nonsignificant.

Overall Satisfaction

Finally, given sufficient between-study heterogeneity, we conducted subgroup analyses to identify conditions where flipped learning has more or less influence on students' satisfaction. Similar to the other two domains, both discipline and location had at least a marginal effect on effect size estimates for overall satisfaction ($p = .069$ and $p = .002$, respectively). For discipline, although all subgroup effects were positively valanced, only the effect for mathematics was significant. Follow-up analyses revealed that the effect of flipped learning on student satisfaction was significantly greater for mathematics than for social science, engineering, or natural science classes. For location, flipped learning produced greater student satisfaction in Asia ($g = 0.67$) and Europe ($g = 0.51$), but the effect of flipped learning on satisfaction in North America/Australia was close to zero ($g = 0.06$). Post hoc analyses confirmed that the effect for North American countries and Australia was significantly smaller than the effects for Asian or European countries. There were no statistical differences in effect sizes based on class level.

Of the six design-based or methodological moderators tested, only publication type significantly influenced the effect of flipped versus lecture-based learning on

overall student satisfaction. Whereas flipped learning was positively and significantly associated with satisfaction in peer-reviewed articles ($g = 0.26$), the effect was null and close to zero in conference papers ($g = 0.05$). The link between flipped learning and overall student satisfaction was not significantly moderated by whether or how students were held accountable for preclass learning, the amount of time they spent in class, the duration of the flipped intervention, whether or not the outcome variable was diluted, or the pre-intervention equivalence of the conditions.

As with intra-/interpersonal outcomes, there were too few studies reporting satisfaction ($n = 66$) to examine the influence of all nine moderators simultaneously, so we conducted meta-regressions separately for each group of moderators. Similar to intra-/interpersonal skills, only the model testing the joint impact of contextual variables was significant, and only study location explained significant variability in effect sizes net of other characteristics in the model ($R^2 = .09$).[2] When tested collectively, neither design characteristics nor methodological features accounted for significant variance in the effects of flipped learning on overall satisfaction.

Publication Bias

To examine potential publication bias, we first conducted moderation analyses comparing the magnitude of the effect sizes produced by published (peer-reviewed articles, book chapters) versus unpublished (conference papers, dissertations) studies for each domain (Polanin et al., 2016). Consistent with our univariate analyses, published studies reported significantly larger effect sizes than unpublished studies for all three outcomes: academic ($g = 0.41$ vs. 0.29; $Q_b = 4.20$, $p = .040$), intra-/interpersonal ($g = 0.48$ vs. 0.15; $Q_b = 12.72$, $p < .001$), and satisfaction ($g = 0.26$ vs. 0.02; $Q_b = 5.49$, $p = .019$). We next conducted Duval and Tweedie's (2000) funnel plot analysis for each domain to assess the likelihood that our analyses are missing unpublished studies showing trivial, null, or negative effects. For all three outcomes, the trim and fill procedure identified zero missing effects, revealing no evidence of publication bias. Finally, we computed the fail-safe N for each domain. The classic fail-safe N (number of null studies needed to cause the effect to become nonsignificant) was 83,370 for academic outcomes, 10,725 for intra-/interpersonal skills, and 1,410 for satisfaction, all of which exceed the $5n + 10$ benchmark (Rosenthal, 1979). Using Orwin's (1983) more conservative test, we found that 505, 205, and 75 missing null studies would be needed to reduce the effect sizes to a trivial level of .10. These findings suggest that, despite evidence of publication bias in the *field*, our meta-analytic sample is robust, and publication bias does not pose a significant threat to the validity of our results.

Sensitivity Analyses

Last, to examine the extent to which our results are robust to outliers, we identified observations with standardized residuals of greater than 3 in absolute magnitude (Viechtbauer & Cheung, 2010). A total of 14 studies were identified as outliers on one or more outcomes. Examining these studies did not reveal any methodological flaws, thus we first tested the impact of Windsorizing the effects

(adjusting the value to the next closest effect size in the distribution; Tabachnick & Fidell, 2013). These adjustments had minimal influence on the estimated synthesis effects (Δg ranged from −0.04 to +0.04) and did not change any patterns of significance (see online Supplemental Table S5). Dropping versus Windsorizing the detected outliers produced a nearly identical pattern; although the changes in g were larger (−0.11 to +0.05), no significance levels were reduced (see online Supplemental Table S6).

Discussion

The present study breaks new ground by providing an unprecedentedly rigorous and comprehensive investigation of the effects of flipped learning on academic, intra-/interpersonal, and satisfaction-related outcomes in higher education. Overall, our results are promising, indicating that flipped classes do, on average, outperform their lecture-based counterparts across a wide range of 21st-century learning outcomes. At the same time, our work highlights the heterogeneity of flipped learning outcomes and identifies several contexts where the benefits of flipped interventions may be less pronounced or even nonexistent. These patterns and their potential implications for adopting and implementing flipped learning are discussed below.

Overall Efficacy of the Flipped Classroom

Past meta-analyses of flipped learning, which have focused almost exclusively on academic achievement, have generally found that flipped pedagogies are associated with small to moderate increases in students' performance on knowledge-based assessments (K.-S. Chen et al., 2018; Cheng et al., 2019; Hew & Lo, 2018; Karagöl & Esen, 2019; Låg & Sæle, 2019; Lo et al., 2017; Shi et al., 2020; van Alten et al., 2019). Consistent with this work, we found that students in flipped classrooms outperformed their traditional counterparts by 0.39 standard deviations in overall academics. Although small by Cohen's (1988) standards, this effect size is close to Hattie's (2012) cutoff for meaningful educational interventions (>0.4) and comparable with the efficacy of strategies like peer learning and time management (Schneider & Preckel, 2017).

With the exception of studies focused on nursing courses in China (Hu et al., 2018; Tan et al., 2017; Xu et al., 2019), previous meta-analyses of flipped learning have treated academic outcomes as a uniform entity, paying little attention to whether outcomes differ based on their position in the learning taxonomy. The assumption that flipped learning should be particularly adept at fostering higher-taxonomy skills was only partially supported by our results. Although the advantage for flipped learning was greater for skills-based outcomes than for assessments of foundational knowledge ($g = 0.53$ vs. 0.34), flipped learning produced significantly *smaller* improvements in higher-order thinking ($g = 0.20$) than in the other two categories. Given that many studies did not specify the types of exam questions used, some measures categorized as foundational knowledge likely involved some higher-order thinking, potentially exaggerating the advantage for knowledge over higher-level thought. However, it is also possible that higher-order thinking is simply a more difficult skill to teach, regardless of pedagogy. To the extent that higher-level cognitive aptitudes take years to master (Bloom, 1956;

Fink, 2013), the benefits of a single flipped course on their development may be less apparent.

In addition to elucidating the influence of flipped learning on academic outcomes, our study provides the first robust evidence that flipped courses outperform lecture-based courses in developing students' intra-/interpersonal skills (g = 0.43). Compared with students in traditional courses, students in flipped classes demonstrated greater meta-cognitive gains (g = 0.37), greater confidence and interpersonal skills (g = 0.41), and greater engagement and identification with the discipline (g = 0.52). Although these effects are smaller than the preliminary estimates garnered from Chinese nursing courses (d = 1.18–1.60; Liu et al., 2018; Tan et al., 2017; Xu et al., 2019), they are just as large as the effects of flipped learning on academic outcomes. Thus, consistent with popular and theoretical arguments for flipped learning's utility for teaching 21st-century skills, flipped interventions appear to yield moderate to medium-sized gains in a variety of intra-/interpersonal proficiencies needed to succeed in a changing, globally connected world.

Finally, our analyses revealed a small positive effect of flipped learning on students' overall satisfaction (g = 0.22). When examining course and instructor satisfaction independently, only course evaluations were significantly related to learning condition, with students in flipped classes reporting higher course satisfaction than their traditionally taught counterparts. Although hardly a silver bullet for improving course evaluations, our findings paint a more optimistic picture of the effect of flipped learning on satisfaction than previous meta-analyses, which found raw or adjusted effect sizes of close to zero (Låg & Sæle, 2019; van Alten et al., 2019). Despite being smaller than the observed academic and intra-/interpersonal gains (cf. Uttl et al., 2017), these results can provide some reassurance to instructors that converting to a flipped classroom is unlikely to hurt their individual evaluations and could slightly improve their course ratings.

Moderators of Flipped Learning

A second objective of our research was to identify theoretical and methodological moderators of the efficacy of flipped versus lecture-based learning. Although previous meta-analyses have detected substantial heterogeneity in flipped learning outcomes, our study is the first to rigorously test a wide range of moderators using an adequately powered sample. Our research is also the first to examine factors that may influence the efficacy of flipped versus traditional courses for teaching intra-/interpersonal competencies. As expected, we found significant and substantial heterogeneity in flipped learning outcomes across all three learning domains. Of the nine contextual, design-based, or methodological moderators examined, all but one (outcome dilution) accounted for at least marginal variance in one or more domains. At the same time, only geographic region predicted variability in all three domains, and several univariate predictors became nonsignificant in the meta-regression models, revealing a complex picture of how factors independently and jointly influence different learning outcomes.

Across all three domains, the most robust predictors of differences in flipped learning outcomes were characteristics related to the context of the flipped intervention. Based on theory and past research, we postulated that flipped learning

may be most effective in applied, skill-based disciplines; in upper-level and graduate courses; and (perhaps) in cultures where active learning is used less routinely in higher education. With a few key exceptions, our analyses largely supported these predictions. The most notable deviation was the minimal influence of class level on flipped learning outcomes. Despite univariate evidence that flipped pedagogies produced marginally better academic outcomes in upper-level versus introductory undergraduate courses, this difference did not hold up in the meta-regression, and class level did not account for variability in the development of intra-/interpersonal skills or satisfaction with flipped learning. This lack of robust effects may be due, at least in part, to the difficulty of accurately coding class level from course names and (sometimes very limited) descriptions. The results may also be muddied by differences in the nature of introductory, upper-level, and graduate courses across countries, universities, and even instructors.

In contrast, several disciplinary differences were observed across learning outcomes. Consistent with Låg and Sæle's (2019) discovery that flipped learning was more effective in the humanities than in STEM, the largest academic and intra-/interpersonal benefits were found in foreign language courses and the smallest effects were in engineering. However, a closer look reveals considerable nuance in this humanities versus STEM dichotomy. First, by distinguishing languages from other humanities classes, our work shows that superior outcomes in humanities (see Cheng et al., 2019; Låg & Sæle, 2019) are largely driven by foreign language courses. Language courses seem to be particularly well-suited to the flipped classroom, perhaps because their objectives include speaking, writing, and listening skills that are easily practiced via active learning. Second, we found that, rather than reflecting a uniform group, STEM courses varied considerably in flipped learning outcomes. Whereas students in flipped technology and health science classes outperformed their traditionally taught counterparts by about half a standard deviation, effect sizes for mathematics and engineering were the smallest of all eight disciplines.

Overall, these results point toward greater efficacy of flipped learning in more applied, skill-based disciplines, with an interesting exception: engineering. Although some have argued that the hands-on nature of engineering should make it particularly amenable to a flipped model (Karabulut-Ilgu et al., 2018), engineering courses consistently produced smaller flipped gains than other disciplines. It is possible that engineering courses rely more on difficult to teach higher-order thinking skills than courses in other applied disciplines or that flipped engineering courses differ less markedly from their traditional controls. Further research is needed to explore these and other possibilities.

Intriguingly, the most robust contextual predictor was the location where the flipped intervention was administered. Although flipped learning was more effective than lecture-based learning in four out of five regions, the relative advantage of flipped pedagogies was higher in Eastern than in Western countries. For academic outcomes, flipped courses showed the greatest gains in Middle Eastern and Asian countries, followed by European, and then North American/Australian countries. The only region not to show significant learning gains was Latin America; however, given that only five studies were available, this finding should be interpreted with caution. For intra-/interpersonal skills, our results followed a

similar pattern, with flipped learning demonstrating greater benefits in the Middle East and Asia than in Europe and North America/Australia. Students in Asia and Europe also reported greater satisfaction with flipped than lecture classes, whereas North American and Australian students showed no preference. Notably, location remained the strongest predictor of variability in flipped learning outcomes in the meta-regressions. Thus, even after accounting for the large proportion of foreign language courses taught in the Middle East and Asia, location still accounted for the greatest share of heterogeneity in each of the three learning domains.

Given the dominance of geographic locale in our predictive models, it is striking that only one other meta-analysis has examined location as a moderator of flipped learning outcomes. Consistent with our results, Karagöl and Esen (2019) found that flipped learning was more effective in studies conducted in Turkey than in North American countries. To our knowledge, no studies have directly examined cross-cultural variability in the implementation of flipped pedagogies or attempted to identify mechanisms that may lead to disparate efficacy in different countries. Identifying the specific variables underlying the observed regional effects—cultural, instructional, or otherwise—is thus a key direction for future research. However, we suspect our results are at least partly due to the fact that flipped classes may provide a greater departure from traditional courses in non-Anglo countries. In North American higher education, even classes that adhere to the lecture model often incorporate some aspects of active learning (cf. Hora & Ferrare, 2012). In Asia and the Middle East, where classes tend to be more teacher-centric and less reliant on active learning (Ly & Brew, 2010; Phuong-Mai et al., 2005), adopting a flipped model may produce more pronounced benefits.

Relative to contextual characteristics, differences in the design of flipped interventions accounted for notably less heterogeneity in flipped learning outcomes. Whereas van Alten et al. (2019) found that reducing class time significantly reduced the effectiveness of flipped learning, results from our larger sample revealed that time in class had no effect on academic outcomes or satisfaction, and had only a marginal influence on intra-/interpersonal skills. These findings suggest that modest reductions in the number of hours spent in class are unlikely to undermine academic outcomes *or* boost satisfaction with the flipped model. When instructors' learning goals include intra-/interpersonal aptitudes, reducing class time *may* diminish the efficacy of the flipped intervention, but the link is tenuous. In a similar vein, we found some evidence that partially flipping a course may produce better academic outcomes than flipping the entire class, but this effect was only marginal net of other study characteristics. Nevertheless, this pattern raises the intriguing possibility that judiciously flipping select segments of a course may have some advantages over flipping a course in its entirety. Given the time and skill required to design effective flipped class sessions (McLaughlin et al., 2016), partially flipped courses may be easier to implement successfully, particularly during early iterations. Partially flipped courses may also afford instructors more flexibility to flip content that lends itself best to the model, while saving more complex topics for in-class instruction (Lo et al., 2017; Simmons et al., 2020).

Unexpectedly, our results regarding the benefits of accountability assignments diverged from those of previous studies. Consistent with the idea that preclass

learning is key to students' ability to benefit from in-class exercises (Sweller, 2020), past meta-analyses have found that flipped interventions are more effective when they hold students accountable for watching the vodcasts before class (Hew & Lo, 2018; Låg & Sæle, 2019; Lo et al., 2017; van Alten et al., 2019). In contrast, we found that formally assessing students' preclass preparation was, at best, neutral, and at worst, deleterious to students' academic outcomes. Although utilizing in-class quizzes did not influence students' outcomes one way or another, courses that held students accountable via preclass assessments (e.g., worksheets, quizzes) actually demonstrated fewer gains from flipped learning than courses with no accountability. Unlike in-class quizzes, preclass quizzes, or worksheets may encourage students to approach vodcasts with the goal of completing the assessment rather than learning the material. Yet, despite avoiding this pitfall, in-class quizzes still did not yield any *advantage* in our study. Given that most of the meta-analyses showing a positive effect for accountability included primary/secondary students in their sample, it may be that accountability is less important for flipped learning outcomes in higher education, where self-regulation abilities are generally higher (Wigfield et al., 2011). However, it is also possible that the association is spurious (e.g., instructors of highly motivated learners may not need to assign accountability and may also have students who are most likely to benefit from flipped learning), suggesting that more research is needed before firm conclusions can be drawn.

Last, we examined the extent to which estimates of the efficacy of flipped interventions were influenced by methodological features of the studies. Although we found no evidence that studies utilizing diluted outcome measures produced lower estimates than studies making more "pure" comparisons, subtype analyses revealed significant differences based on publication type and marginal differences for group equivalence. Confirming the presence of publication bias in the field, peer-reviewed articles reported greater academic benefits of flipped learning than did dissertations, and larger gains in intra-/interpersonal skills and satisfaction than both conference presentations and dissertations. We also found more pronounced benefits of flipped learning among studies that established outcome-specific group equivalence prior to the intervention than those that did not provide information on equivalency, but only in the academic domain. This is largely consistent with Låg and Sæle's (2019) discovery that randomized studies produced larger effects than those with unequal or unknown equivalence, and suggests that the noise introduced into poorly controlled studies may attenuate estimates of flipped learning effects (cf. Cheung & Slavin, 2016). However, both publication type and pre-intervention group equivalence fell below significance in our meta-regression models, suggesting that these methodological features do not account for notable heterogeneity in flipped learning outcomes net of other study characteristics.

Limitations and Future Directions

Despite the numerous strengths of our meta-analysis, we are cognizant of several limitations that we hope will be catalysts for additional work in this area. First, because most studies in our meta-analysis were instructors' first attempts at implementing a flipped classroom, flipped learning may actually be more

effective than our estimates suggest. Most curricular changes require adjustments after their first implementation and there is evidence that learning outcomes improve as instructors gain experience with flipped pedagogies (e.g., Dove & Dove, 2017; Gorres-Martens et al., 2016; Overmyer, 2014). Indeed, given that flipped interventions are generally compared with traditional courses that have been taught and revised multiple times, it is quite notable that consistent learning gains were detected. To provide a more accurate test of flipped learning's relative merits, researchers need to continue to evaluate the efficacy of flipped courses as they are refined and redesigned over time.

Second, although differences in course design had a relatively modest influence on the efficacy of flipped learning in our study, the apparent superiority of contextual over design-based predictors should be interpreted with caution. Although we were able to explain more variance in flipped learning outcomes than past meta-analyses (e.g., Låg & Sæle, 2019; van Alten et al., 2019), more than 75% of between-study heterogeneity in academic effects could not be accounted for by our fairly robust set of moderators—and explanatory power was even lower for satisfaction and intra-/interpersonal skills. We suspect that much of this unidentified heterogeneity stems from uncodable differences in the implementation of the flipped model. Studies in our meta-analysis used a wide range of teaching strategies, active learning exercises, and technologies, some of which were likely more effective than others. Indeed, both qualitative research and review articles suggest a number of design factors that are likely to influence the efficacy of the flipped classroom, including the quality, duration, and style of the vodcasts; the use of collaborative versus independent learning; the regularity and timeliness of instructor guidance and feedback; and the constructive alignment of pre- and in-class activities (Akçayır & Akçayır, 2018; Kim et al., 2014; Lo et al., 2017; McLaughlin et al., 2016).

Nevertheless, identifying the specific features of flipped learning that contribute most to its success has proven difficult, largely due to a lack of systematic reporting. Although both Låg and Sæle (2019) and van Alten et al. (2019) attempted to examine whether the nature of active learning exercises influenced the efficacy of flipped courses, they were unable to code instructional differences in enough detail to explain any variance in outcomes. We similarly found that a minority of eligible studies included the information needed to reliably code these variables. To remedy this issue, it is imperative that future studies of flipped learning's efficacy report more consistent and thorough implementation details, including the design and application of their vodcasts and in-class activities, as well as the activities of the control condition.

Third, given the notorious difficulty of detecting small moderation effects in highly heterogeneous meta-analytic samples (Hempel et al., 2013), the relative dearth of moderation effects for intra-/interpersonal skills and satisfaction should be interpreted with caution. Although it may be that the tested moderators truly account for more variability in academic outcomes than in the other two domains, we had substantially more power to detect moderation effects in the former, highlighting the need for additional, well-controlled studies of flipped learning effects on nonacademic outcomes. It also would be beneficial to examine other potentially important sources of variability in learning outcomes, including both

student and instructor attributes (e.g., baseline meta-cognitive skills; previous experience with flipped pedagogies; instructor training and resources; McLaughlin et al., 2016; Simmons et al., 2020).

Finally, although the scope and rigor of our meta-analysis permits greater confidence in the results than previous investigations, there are several limits to the generalizability of our findings that should be noted. First, despite including research from more than 40 different countries, our analyses were limited (for pragmatic reasons) to studies published in English, which may have influenced our effects. Although excluding non-English articles is unlikely to have substantially altered our *overall* meta-analytic conclusions (Morrison et al., 2012; Nussbaumer-Streit et al., 2020), the impact of potential language bias on *regional* effect estimates may be greater—particularly in regions where few eligible studies were available (e.g., Latin America). Additional, well-resourced studies of flipped learning effects are needed to remedy this potential issue. Future research also is needed regarding the generalizability of our findings to diverse subpopulations and instructional mediums. For instance, in light of the documented benefits of active learning for racially minoritized and first-generation students (Eddy & Hogan, 2014; Theobald et al., 2020), future studies should examine whether flipped learning has similarly pronounced benefits for underrepresented students. Many questions also exist regarding the application of flipped learning to fully online spaces, including whether in-person active learning can be replaced by virtual learning exercises with similarly positive results. To date, flipped learning studies have been conducted almost exclusively with in-person courses, thus expanding this research to include online applications is an important (and timely) direction for future work.

Conclusion

Is the flipped classroom a mecca for developing higher-taxonomy proficiencies or a pedagogical fad with few measurable benefits for the average postsecondary student? Using a large, carefully selected sample of studies, our research demonstrates that flipped learning is superior to lecture-based learning for fostering a range of academic, intra-/interpersonal, and satisfaction-related outcomes. However, our work also suggests that not all flipped interventions are created equal and that educational context plays a significant role in shaping the relative benefits of adopting a flipped approach. Flipped pedagogies may produce sizable gains in foreign language, technology, and medical science courses and when implemented in Asian and Middle Eastern countries. But in other contexts—including engineering and mathematics classes at North American/Australian colleges—the educational benefits of flipped learning may be much more modest.

Notes

[1] We originally coded final grade as an outcome when grades were calculated identically across conditions, but ended up dropping this category due to its low frequency in the eligible data set ($n = 18$) and redundancy with other academic outcomes (e.g., final grades that consisted mostly of exam scores already included as foundational knowledge). This resulted in the exclusion of four articles in which only final grades were reported.

[2] Given that nine dummy variables are needed to represent the three contextual moderators, the meta-regression testing this subset of characteristics may still be underpowered. To explore the impact this might have on the results, we condensed study discipline into three categories—thus yielding a total of six dummy variables in the model—and obtained the same pattern of results.

References

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

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

Al-Shabibi, T. S., & Al-Ayasra, M. A. (2019). Effectiveness of the flipped classroom strategy in learning outcomes (bibliometric study). *International Journal of Learning, Teaching and Educational Research*, *18*(3), 96–127. https://doi.org/10.26803/ijlter.18.3.6

Altemueller, L., & Lindquist, C. (2017). Flipped classroom instructions for inclusive learning. *British Journal of Special Education*, *44*(3), 341–358. https://doi.org/10.1111/1467-8578.12177

Beran, T. N., Violato, C., & Kline, D. W. (2007). What's the "use" of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, *37*(1), 27–43. https://doi.org/10.47678/cjhe.v37i1.183545

Berrett, D. (2012, February 19). How "flipping" the classroom can improve the traditional lecture. *Chronicle of Higher Education*. https://www.chronicle.com/article/how-flipping-the-classroom-can-improve-the-traditional-lecture/

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. David McKay.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

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*(4), 409–416. https://doi.org/10.1111/jcal.12250

Chen, F., Lui, A. M., & Martinelli, S. M. (2017). A systematic review of the effectiveness of flipped classrooms in medical education. *Medical Education*, *51*(6), 585–597. https://doi.org/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 in Review*, *52*(9), 910–924. https://doi.org/10.1111/medu.13616

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development*, *67*(4), 793–824. https://doi.org/10.1007/s11423-018-9633-7

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. Sage.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Sage. https://doi.org/10.4135/9781412984010

Day, L. J. (2018). A gross anatomy flipped classroom effects performance, retention, and higher-level thinking in lower performing students. *Anatomical Sciences Education*, *11*(6), 565–574. https://doi.org/10.1002/ase.1772

Dickersin, K., & Min, Y.-I. (1993). NIH clinical trials and publication bias. *Online Journal of Current Clinical Trials*, *2*(1), Article 50. http://access.portico.org/stable?au=phzcxbg8x

Dove, A., & Dove, E. (2017). How flipping much? Consecutive flipped mathematics courses and their influence on students' anxieties and perceptions of learning. *Journal of Computers in Mathematics and Science Teaching*, *36*(2), 129–141. https://www.learntechlib.org/primary/p/178271/

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE Life Sciences Education*, *13*(3), 453–468. https://doi.org/10.1187/cbe.14-03-0050

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*(2), 74–78. https://doi.org/10.1097/NNE.0000000000000554

Faculty Focus. (2015). *Flipped classroom trends: A survey of college faculty*. Magna.

Ferrare, J. J., & Hora, M. T. (2014). Cultural models of teaching and learning in math and science: Exploring the intersections of culture, cognition, and pedagogical situations. *Journal of Higher Education*, *85*(6), 792–825. https://doi.org/10/1353/jhe.2014.0030

Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses, revised and updated*. Jossey-Bass.

Fitzpatrick, M. (2012, June 24). Classroom lectures go digital. *The New York Times*. https://www.nytimes.com/2012/06/25/us/25iht-educside25.html

Fong, C. J., Davis, C. W., Kim, Y., Kim, Y. W., Marriott, L, & Kim, S. (2017). Psychosocial factors and community college student success: A meta-analytic investigation. *Review of Educational Research*, *87*(2), 388–424. https://doi.org/10.3102/00346653479

Frambach, J. M., Driessen, E. W., Chan, L.-C., & van der Vleuten, C. P. M. (2012). Rethinking the globalisation of problem-based learning: How culture challenges self-directed learning. *Medical Education*, *46*(8), 738–747. https://doi.org/10.1111/j.1365-2923.2012.04290.x

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 United States of America*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., Wilt, T. J., Griffith, L., Oremus, M., Raina, P., Ismaila, A., Santaguida, P., Lau, J., & Trikalinos, T. A. (2011). Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *Journal of Clinical Epidemiology*, *64*(11), 1187–1197. https://doi.org/10.1016/j.jclinepi.2010.08.010

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300–329. https://doi.org/10.3102/0034654312457206

Galindo-Domínguez, H., & Bezanilla, M. J. (2019). A systematic review of flipped classroom methodology at university level in Spain. *International Journal of Technology and Educational Innovation*, *5*(1), 81–90. https://doi.org/10.24310/innoeduca.2019.v5i1.4470

Gianoni-Capenakas, S., Lagravere, M., Pachêco-Pereira, C., & Yacyshyn, J. (2019). Effectiveness and perceptions of flipped learning model in dental education: A systematic review. *Journal of Dental Education*, *83*(8), 935–945. https://doi.org/10.21815/JDE.019.109

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*(5), Article 6898. https://doi.org/10.5688/ajpe6898

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). Russell Sage Foundation.

Gorres-Martens, B. K., Segovia, A. R., & Pfefer, M. T. (2016). Positive outcomes increase over time with the implementation of a semiflipped teaching model. *Advances in Physiological Education*, *40*(1), 32–37. https://doi.org/10.1152/advan.00034.2015

Gupta, V., Hanges, P. J., & Dorfman, P. (2002). Cultural clusters: Methodology and findings. *Journal of World Business*, *37*(1), 11–15. https://doi.org/10/1016/S1090-9516(01)00070-0

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge. https://doi.org/10.4324/9780203181522

Hawks, S. J. (2014). The flipped classroom: Now or never? *AANA Journal*, *82*(4), 264–269. https://pubmed.ncbi.nlm.nih.gov/25167605/

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., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

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), Article 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), Article 38. https://doi.org/10.1186/s12909-018-1144-z

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). http://handbook.cochrane.org

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviours, institutions, and organizations across nations* (2nd ed.). Sage.

Hora, M. T., & Ferrare, J. J. (2012). Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching. *Journal of the Learning Sciences*, *22*(2), 212–257. https://doi.org/10.1080/10508406.2012.729767

Howe, K. R., & Berv, J. (2000). Constructing constructivism, epistemological and pedagogical. In D. C. Phillips (Ed.), *Constructivism in education: Opinions and second opinions on controversial issues* (pp. 19–40). University of Chicago Press.

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*(March), 94–103. https://doi.org/10.1016/j.ijnurstu.2017.11.012

Hughes, Y., & Lyons, N. (2017). Does the flipped classroom improve exam performance in medical education? A systematic review. *MedEdPublish*, *6*(2), Article 38. https://doi.org/10.15694/mep.2017.000100

Ives, J., & Castillo-Montoya, M. (2020). First-generation college students as academic learners: A systematic review. *Review of Educational Research*, *90*(2), 139–178. https://doi.org/10.3102/0034654319899707

Jensen, J. L., Holt, E. A., Sowards, J. B., Ogden, T. H., & West, R. E. (2018). Investigating strategies for pre-class content learning in a flipped classroom. *Journal of Science Education and Technology*, *27*(6), 523–535. https://doi.org/10.1007/s10956-018-9740-6

Jensen, S. A. (2011). In-class versus online video lectures: Similar learning outcomes, but a preference for in-class. *Teaching of Psychology*, *38*(4), 298–302. https://doi.org/10.1177/0098628311421336

Karabulut-Ilgu, A., Cherrez, N. J., & Jahren, C. T. (2018). A systematic review of research on the flipped learning method in engineering education. *British Journal of Educational Technology*, *49*(3), 398–411. https://doi.org/10.1111/bjet.12548

Karagöl, I., & Esen, E. (2019). The effect of flipped learning approach on academic achievement: A meta-analysis study. *Hacettepe University Journal of Education*, *34*(3), 708–727. https://doi.org/10.16986/HUJE.2018046755

Kicinski, M., Springate, D. A., & Kontopantelis, E. (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine*, *34*(20), 2781–2793. https://doi.org/10.1002/sim.6525

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*(July), 37–50. https://doi.org/10.1016/j.iheduc.2014.04.003

Kivunja, C. (2014). Innovative pedagogies in higher education to become effective teachers of 21st century skills: Unpacking the learning and innovations skills domain of the new learning paradigm. *International Journal of Higher Education*, *3*(4), 37–48. https://doi.org/10.5430/ijhe.v3n4p37

Kozikoğlu, I. (2019). Analysis of the studies concerning flipped learning model: A comparative meta-synthesis study. *International Journal of Instruction*, *12*(1), 851–868. https://doi.org/10.29333/iji.2019.12155a

Låg, T., & Sæle, R. G. (2019). Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis. *AERA Open*, *5*(3), 1–17. https://doi.org/10.1177/2332858419870489

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *Journal of Economic Education*, *31*(1), 30–43. https://doi.org/10.2307/1183338

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

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. https://www.jstor.org/stable/jeductechsoci.12.4.282

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (2nd ed.). Sage.

Liu, Y.-Q., Li, Y.-F., Lei, M.-J., Liu, P.-X., Theobald, J., Meng, L.-N., Liu, T.-T., Zhang, C.-M., & Jin, C.-D. (2018). Effectiveness of the flipped classroom on the development of self-directed learning in nursing education: A meta-analysis. *Frontiers of Nursing*, *5*(4), 317–329. https://doi.org/10.1515/FON-2018-0032

Lo, C. K., Hew, K. F., & Chen, G. W. (2017). Toward a set of design principles for mathematics flipped classrooms: A synthesis of research in mathematics education. *Educational Research Review*, *22*(November), 50–73. https://doi.org/10.1016/j.edurev.2017.08.002

Ly, B. H., & Brew, C. (2010). Philosophical and pedagogical patterns of beliefs among Vietnamese and Australian mathematics preservice teachers: A comparative study. *Australian Journal of Teacher Education*, *35*(2), 67–86. https://doi.org/10.14221/ajte.2010v35n2.5

Martin, A. J., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of Educational Research*, *79*(1), 327–365. https://doi.org/10.3102/0034654308325583

Mazur, E. (2009). Farewell, lecture? *Science*, *323*(5910), 50–51. https://doi.org/10.1126/science.1168927

McLaughlin, J. E., Roth, M. T., Glatt, D. M., Gharkholonarehe, N., Davidson, C. A., Griffin, L. M., Esserman, D. A., & Mumper, R. J. (2014). The flipped classroom: A course redesign to foster learning and engagement in a health professions school. *Academic Medicine*, *89*(2), 236–243. https://doi.org/10.1097/ACM.0000000000000086

McLaughlin, J. E., White, P. J., Khanova, J., & Yuriev, E. (2016). Flipped classroom implementation: A case report of two higher education institutions in the United States and Australia. *Computers in the Schools*, *33*(1), 24–37. https://doi.org/10.1080/07380569.2016.1137734

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Education & Psychology*, *94*(1), 156–163. https://doi.org/10.1037//0022-0663.94.1.156

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., Mierzwinski-Urban, M., Clifford, T., Hutton, B., & Rabb, D. (2012). The effect of English language restriction on systematic review-based meta-analyses: A systematic

review of empirical studies. *International Journal of Technology Assessment in Health Care*, *28*(2), 138–144. https://doi.org/10.1017/S0266462312000086

National Academies of Sciences, Engineering, and Medicine. (2017). *Supporting students' college success: The role of assessment of intrapersonal and interpersonal competencies*. National Academies Press. https://doi.org/10.17226/24697

Nussbaumer-Streit, B., Klerings, I., Dobrescu, A. I., Persad, E., Stevens, A., Garritty, C., Kamel, C., Affengruber, L., King, V. J., & Gartlehner, G. (2020). Excluding non-English publications from evidence-syntheses did not change conclusions: A meta-epidemiological study. *Journal of Clinical Epidemiology*, *118*, 42–54. https://doi.org/10.1016/j.jclinepi.2019.10.011

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *Internet and Higher Education*, *25*(April), 95. https://doi.org/10.1016/j.iheduc.2015.02.002

Orhan, A. (2019). The effect of flipped learning on students' academic achievement: A meta-analysis study. *Cukurova University Faculty of Education Journal*, *48*(1), 368–396. https://doi.org/10.14812/cufej.400919

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159. https://doi.org/10.2307/1164923

Overmyer, G. R. (2014). *The flipped classroom model for college algebra: Effects on student achievement* [Doctoral dissertation, Colorado State University]. CSU Theses and Dissertations. http://hdl.handle.net/10217/83800

Phuong-Mai, N., Terlouw, C., & Pilot, A. (2005). Cooperative learning vs Confucian heritage culture's collectivism: Confrontation to reveal some cultural conflicts and mismatch. *Asia Europe Journal*, *3*(3), 403–419. https://doi.org/10.1007/s10308-005-0008-4

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

Presti, C. R. (2016). The flipped learning approach in nursing education: A literature review. *Journal of Nursing Education*, *55*(5), 252–257. https://doi.org/10.3928/01484834-20160414-03

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93*(3), 223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Roehling, P. V. (2018). *Flipping the college classroom: An evidence-based guide*. Palgrave. https://doi.org/10.1007/978-3-319-69392-7

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sanchez, I. M. (2000). Motivating and maximizing learning in minority classrooms. *New Directions for Community Colleges*, *2000*(112), 35–44. https://doi.org/10.1002/cc.11203

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098

Shcheglova, I. A. (2018). A cross-cultural comparison of the academic engagement of students. *Russian Education & Society*, *60*(8–9), 665–681. https://doi.org/10.1080/10609393.2018.1598163

Shi, Y., Ma, Y., MacLeod, J., & Yang, H. H. (2020). College students' cognitive learning outcomes in flipped classroom instruction: A meta-analysis of the empirical literature. *Journal of Computers in Education*, *7*(1), 79–103. https://doi.org/10.1007/s40692-019-00142-8

Simmons, M., Colville, D., Bullock, S., Willems, J., Machado, M., McArdle, A., Tare, M., Kelly, J., Ali Taher, M., Sallyann Middleton, S., Shuttleworth, M., & Reser, D. (2020). Introducing the flip: A mixed method approach to gauge student and staff perceptions on the introduction of flipped pedagogy in pre-clinical medical education. *Australasian Journal of Educational Technology*, *36*(3), 163–175. https://doi.org/10.14742/ajet.5600

Smallhorn, M. (2017). The flipped classroom: A learning model to increase student engagement not academic achievement. *Student Success*, *8*(2), 43–53. https://doi.org/10.5204/ssj.v8i2.381

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

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

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., II, Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., . . . Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

Tobin, K., & Tippins, D. (1993). Constructivism as a referent for teaching and learning. In K. Tobin (Ed.), *The practice of constructivism in science education* (pp. 1–20). Lawrence Erlbaum.

Turan, Z., & Akdag-Cimen, B. (2020). Flipped classroom in English language teaching: A systematic review. *Computer Assisted Language Learning*, *33*(5–6), 590–606. https://doi.org/10.1080/09588221.2019.1584117

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*(September), 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review*, *28*(November), Article 100281. https://doi.org/10.1016/j.edurev.2019.05.003

Van Merriënboer, J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

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

Wankat, P. C. (2002). *The effective, efficient professor: Teaching, scholarship and service*. Allyn & Bacon.

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). Routledge.

Windschitl, M. (2002). Framing constructivism in practice as the negotiation of dilemmas: An analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of Educational Research*, *72*(2), 131–175. https://doi.org/10.3102/00346543072002131

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education*, *37*(6), 683–699. https://doi.org/10.1080/02602938.2011.563279

Xu, P., Chen, Y., Nie, W., Wang, Y., Song, T., Li, H., Li, J., Yi, J., & Zhao, L. (2019). The effectiveness of a flipped classroom on the development of Chinese nursing students' skill competence: A systematic review and meta-analysis. *Nurse Education Today*, *80*(September), 67–77. https://doi.org/10.1016/j.nedt.2019.06.005

Yang, L., Sun, T., & Liu, Y. (2017). A bibliometric investigation of flipped classroom research during 2000–2015. *International Journal of Emerging Technologies*, *12*(6), 178–186. https://doi.org/10.3991/ijet.v12i06.7095

Zainuddin, Z., & Halili, S. H. (2016). Flipped classroom research and trends from different fields of study. *International Review of Research in Open and Distributed Learning*, *17*(3), 313–340. https://doi.org/10.19173/irrodl.v17i3.2274

Authors

CARRIE A. BREDOW is an associate professor of psychology at Hope College, PO Box 9000, Holland, MI 49422, USA; email: *bredow@hope.edu*. Her primary line of research focuses on the social-cognitive processes underlying the development and maintenance of adult romantic relationships. As a developmental psychologist at an undergraduate liberal arts college, she is also interested in the scholarship of teaching, including the efficacy of flipped learning and other alternative pedagogies in higher education.

PATRICIA V. ROEHLING is a professor of psychology at Hope College, PO Box 9000, Holland, MI 49422, USA; email: *roehling@hope.edu*. Her research areas are the scholarship of teaching, work-life issues, and weight discrimination. She is author of the book *Flipping the College Classroom: An Evidence-Based Guide*. She has also published articles on flipped learning, facilitating class discussions, and the effective use of PowerPoint.

ALEXANDRA J. KNORP received her BA in psychology from Hope College; email: *aknorp@gmail.com*. She is currently a graduate student at the University of Detroit Mercy where she has obtained her master's degree in school psychology and is working toward her school psychology specialist degree.

ANDREA M. SWEET received her BA in psychology from Hope College; email: *andrea.sweet1397@gmail.com*. She plans to pursue a master's degree in counseling psychology and work as a school guidance counselor.