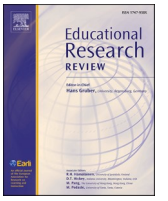
Educational Research Review 34 (2021) 100394

Contents lists available at ScienceDirect

Educational Research Review

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



Thematic Review

Facilitating flexible learning by replacing classroom time with an online learning environment: A systematic review of blended learning in higher education

Claude Müller *, Thoralf Mildenberger

Zurich University of Applied Sciences, Switzerland

ARTICLE INFO

Keywords:
Blended learning
Learning effectiveness
Higher education
Meta-analysis
Flexible learning

ABSTRACT

Higher education institutions are trying to provide more flexibility and individualization, which is mainly realized through the use of new technologies and implemented in online or blended learning designs. This systematic review aims to investigate the impact of replacing classroom time with an online learning environment. The meta-analysis ($k = 21$ effect sizes) applied strict inclusion criteria concerning research design, measurement of learning outcomes and implementation of blended learning. The estimated effect size (Hedge's g) was positive, although not significantly different from zero and the confidence interval [-0.13, 0.25], suggesting that overall differences between blended and conventional classroom learning are small, and, at best, very small negative or moderate positive effects are plausible. This means that despite a reduction in classroom time between 30 and 79 per cent, equivalent learning outcomes were found. Consequently, blended learning with reduced classroom time is not systematically more or less effective than conventional classroom learning.

1. Introduction

The digitalization of society means that the demands placed on employee skills increase or change throughout their careers (OECD, 2019). To meet the growing need for highly qualified employees in the labour market, higher education should be made accessible to broader sections of the population (Orr et al., 2020). Owing to the changing requirements of the labour market, it can also be assumed that the prevalence of multi-careers will increase; a large number of people will be active in various occupational fields during their lifetime (OECD, 2017) and that they will have to expand their skills in the sense of continuous lifelong learning. To meet the demands of a digital society, educational institutions are expected to provide greater flexibility and individualization so that learners have the opportunity to adapt the learning process to their own needs and specific life phases (Barnett, 2014). Flexible learning, which is frequently mentioned in this context, is a broad term with different interpretations (Hrastinski, 2019). In a more general sense, flexible learning provision should meet the diverse needs of learners and enable them to take more personal responsibility for the learning process (Wade, 1994). Central to flexible learning are learners and their needs, and the educational services on offer should allow them to decide for themselves what, when, how and where they learn (Higher Education Academy, 2015). Most flexible learning initiatives focus on aspects of temporal and spatial flexibility in learning, which is nowadays realized primarily through the use of new technologies (Tucker & Morris, 2012) and implemented didactically in an online or blended learning environment (Andrade &

* Corresponding author. Zürich University of Applied Sciences, St. Georgenplatz 2, 8401, Winterthur, Switzerland.
E-mail address: muew@zhaw.ch (C. Müller).

https://doi.org/10.1016/j.edurev.2021.100394
Received 11 November 2020; Received in revised form 25 April 2021; Accepted 4 June 2021
Available online 8 June 2021

Alden-Rivers, 2019). Current research confirms (again) that computer technology can create interactive and engaging (additional) learning environments which may have positive effects on knowledge gain, skill acquisition and student perception (e.g. Chen, Wang, Kirschner, & Tsai, 2018). However, blended learning also claims not only to enrich classroom learning but also to redesign the learning environment with higher degrees of freedom for learners (Smith & Hill, 2019). Students should be able to study more independently of time and place and determine content and learning pace individually.

The decisive question is whether online elements are able to replace some aspects of classroom time and enable greater flexibility without compromising educational quality and performance (Owston & York, 2018). This is especially important in the context of the COVID-19 pandemic. Many universities have been considering replacing some or even all their classroom teaching with an online learning environment, not only in the short term but also in the future (Peters et al., 2020; Saichaie, 2020). However, only if face-to-face classroom time can be replaced with more flexible learning conditions without reducing student performance will universities be able to offer and expand these learning formats with any long-term success. This review systematically examines the issue by investigating the impact of replacing classroom instruction with blended learning in a higher education setting and is guided by the following question: Can classroom time be reduced and replaced with an online learning environment to offer more flexibility in the learning process without compromising learning outcomes?

In the first section, blended learning is defined and existing research about the effectiveness of blended learning discussed. In the second section, within the framework of a meta-analysis, the learning effect of courses with reduced face-to-face classroom time and a higher online share is examined and compared with those offering solely face-to-face classroom learning.

1.1. Definition of blended learning

Learning can take place in different modalities, often distinguishing between face-to-face classroom instruction and virtual learning, as well as asynchronous and synchronous learning (Chaeruman, Wibawa, & Syahrial, 2018). Today, (online) technologies are primarily used for virtual and asynchronous learning, while integration with face-to-face classroom instruction is referred to as blended learning. Blended learning is often used interchangeably with terms such as hybrid, mixed-mode or flexible learning. According to Hrastinski (2019), the definitions of blended learning most frequently used in scientific publications are those by Graham (2006): "blended learning systems combine face-to-face instruction with computer-mediated instruction" (p. 5) and by Garrison and Kanuka (2004): "the thoughtful integration of classroom face-to-face learning experiences with online learning experiences." According to Smith and Hill (2019), blended learning definitions are problematic because they are ambiguous and include various teaching practices with little consensus on what they cover. Thus, blended learning encompasses all technically supported learning environments except pure online learning environments and pure classroom instruction. Since practically all universities today use an online learning management system through which at least teaching materials are made available, blended learning is also referred to as "the new traditional model" or the "new normal" (Dziuban, Graham, Moskal, Norberg, & Sicilia, 2018). In comparing the two definitions, that of Garrison and Kanuka (2004) is slightly narrower as it includes a qualitative dimension, requiring a "thoughtful integration" of classroom instruction and online learning. To ascertain the "right mix" between classroom instruction and online learning, according to So and Bonk (2010), the following questions need to be answered. Firstly, what part of the interaction should take place in classroom-based face-to-face or online settings? Secondly, when should online or classroom-based face-to-face learning be used, and thirdly, how can the two modalities be combined to optimize the learning process? Allen, Seaman, and Garrett (2007) employ the online proportion of a learning environment as a differentiation criterion for the four modalities: traditional, web-facilitated, blended/hybrid and online learning (Table 1).

According to the authors, blended learning contains an online proportion of between 30 and 79 per cent of content delivery – and this must take place online rather than in the traditional format. Other authors suggest comparable proportions when discussing blended learning. Bernard, Borokhovski, Schmid, Tamim, and Abrami (2014), for example, use a 1:1 ratio for online and classroom instruction as an inclusion criterion for their meta-analysis. This ratio is also supported by several studies which report that students prefer high and medium online proportions rather than low or supplementary online parts in blended learning (Asarta & Schmidt, 2015; Hilliard & Stewart, 2019; Owston & York, 2018). In contrast, if we assume a broad understanding of blended learning with a supplemental online environment, the integration of online components can lead to an increase in the workload compared with a traditional course, resulting in the so-called "course and a half" syndrome (Garrison & Vaughan, 2008, p. 202).

This review examines in greater detail the effects of replacing face-to-face classroom time with an online learning environment to offer more flexibility in the learning process, and the definition of blended learning by Allen et al. (2007) is considered more precise (compared with other definitions). The lower end of 30 per cent ensures that a substantial part of the teaching takes place online, and

Table 1
Classifications of Courses According to the Proportion of Content Delivery Online (based on Allen et al., 2007, p. 5).

Type of Course	Proportion of Online Delivery Content	Description
Traditional	0%	A course with no online technology used – content is delivered in writing or orally.
Web-facilitated	1–29%	A course that uses web-based technology to facilitate what is essentially a face-to-face course. For example, it uses a course management system (CMS) or web pages to post the syllabus and assignments.
Blended/Hybrid	30–79%	A course that blends online and face-to-face delivery. A substantial proportion of the content is delivered online, typically features online discussions, and usually has some face-to-face meetings.
Online	>80%	A course where most or all the content is delivered online. It typically has no face-to-face meetings.

with an upper end of 79 per cent, blended learning can be differentiated from pure online learning. In this review, blended learning is defined as follows: A course that blends online and classroom learning, with a proportion of between 30 and 79 per cent of the content delivered online.

1.2. Effectiveness of blended learning

Over the last ten years, various meta-analyses about the effectiveness of blended learning have been carried out, and these are discussed here with respect to the central question of this paper.

In their study, Means, Toyama, Murphy, and Baki (2013) compared learning outcomes for either fully online or blended learning environments with those of classroom-based, face-to-face learning. Their analysis between 1996 and 2008 included 45 studies with 50 effects, most of them at higher education level (42 effects). The duration varied according to the study programs and exceeded one month in most cases. The meta-analysis corpus consisted of experimental studies with randomized allocation and quasi-experiments with statistical control for existing group differences. The meta-analysis showed that students under online learning conditions (online and blended learning together) performed slightly better on average than those receiving classroom courses ($g = +0.20$; $p < 0.001$). The advantage over the classroom course was significant in those studies that contrasted blended learning with traditional classroom courses ($g = +0.35$; $p < 0.0001$), but not in those studies that contrasted purely online with classroom conditions ($g = +0.05$; $p = 0.46$). In addition, the researchers investigated several variables of practice, condition and methods if they moderate blended versus classroom-based learning to explain the difference in performance. The "time on task" moderator is especially interesting for this review. The analysis favoured more time spent online than face-to-face (i.e. >50% online) compared with less time spent online than face-to-face (i.e. ≤ 50% online) and the results came close to significance ($Q = 3.62$, $p = 0.06$).

In the meta-analysis by Bernard et al. (2014), the authors integrated 117 effects of 96 studies between 1990 and 2010 comparing blended learning with classroom instruction in higher education. Blended learning includes all lessons with technology where the online share of blended learning does not exceed 50%; purely online courses were not included. The authors emphasize that blended learning is a relatively broad concept: "We argue that this a conservative test of blended learning that explores the lower limits of the addition of OL components" (p. 91). Only studies with real experiments and quasi-experiment designs were included, while studies with two-group pre-experiments and one-group pretest-posttest designs were excluded. All the studies also had to test for selection bias somehow, and the results were similar to those of Means et al. (2013) with a significant positive effect of blended learning compared to conventional classroom instruction ($g = 0.33$, $k = 117$, $p < 0.001$). Bernard et al. (2014) also examined the proportion of time spent online as a moderating variable. They considered two categories: up to 30 per cent of course time and 30 to 50 per cent of course time. They found a "definite trend towards higher effect sizes for longer versus shorter time spent online" (p. 112), but this was not significant (Q-Between = 0.47, $df = 1$, $p = 0.49$), as was the case for Means et al. (2013). The meta-analyses by Borokhovski, Bernard, Tamim, Schmid, and Sokolovskaya (2016) and Schmid et al. (2014) were based on the same data set and are not discussed in detail.

Spanjers et al. (2015) focused on more recent studies (since 2003) in their meta-analysis but included all levels of education (five on K12 level, 19 on upper-secondary level). They insisted that the online aspect should not simply supplement the time, resources or activities of the conventional learning environment but also replace at least part of it. Their analysis favoured blended learning over conventional learning environments and revealed an effect size for objective measurements that was slightly higher than for subjective measurements ($g = 0.34$, $p < 0.01$ versus $g = 0.27$, $p = 0.01$). Student satisfaction showed a small positive effect for blended learning ($g = 0.11$, $p < 0.01$). The authors summarise their findings as follows: "On average, blended learning was somewhat more effective and about equally attractive" (Spanjers et al., 2015, p. 71). However, they also acknowledge that large differences were found between the studies, some of which produced negative results for blended learning.

In the most recent meta-analysis by Vo, Zhu, and Diep (2017), the authors adopt the definition of blended learning used in Bernard et al. (2014) but only included studies conducted between 2001 and 2015. In addition, an objective performance measurement had to be available for the studies, so those using subjective measurements – including student estimations of learning gains – were excluded. A total of 51 effects from 40 studies were incorporated into this meta-analysis. Once again, their findings showed a positive and significant effect similar to the meta-analyses presented above ($g = 0.39$, $p < 0.001$) for blended learning, when compared with conventional classroom instruction.

All the meta-analyses discussed above demonstrate a small to medium positive effect of blended learning compared with conventional face-to-face classroom instruction. However, there are limitations to these studies, which may prevent a generalization of the results. Firstly, some studies integrate subjective and objective performance measures. Mixing these measurements is a sensitive issue because comparisons of self-assessments of learning outcomes in different learning environments include bias effects and should be treated with caution. Some meta-analyses also include studies up to the 1990s, although technologies for teaching and learning have developed enormously since the advent of the internet (e.g. web-based tools and online learning videos). Additionally, the availability and capability of computers, including online access, has improved significantly since then. At the same time, most students have adapted to using computers for learning and no longer struggle to operate digital devices (Chen et al., 2018). Therefore, mixing studies from different generations of teaching and learning with technologies is not conducive to meaningful conclusions.

Most importantly and as the result of unclear definitions of blended learning, a wide variety of blended learning environments have been compared directly with other learning environments. For example, Dziuban et al. (2018) analysed the meta-analysis by Means et al. (2013) and identified studies with different blending techniques such as learning management systems, online instruction, computer laboratories, electronic portfolios and e-mail. Furthermore, some meta-analyses require an inclusion criterion that technology is not used only to supplement the conventional learning environment. However, in all the reviews, no clear distinction is made between studies in which classroom instruction is replaced by online learning and studies in which online learning augments the

conventional learning environment. We analysed the 40 blended learning studies included in the meta-analysis conducted by Vo et al. (2017) and found that only 22 of these compared traditional classroom learning with blended learning (with reduced classroom time), while the other 18 studies either examined e-learning-enriched classroom courses or compared conventional classroom courses with purely online sequences. In studies comparing traditional learning environments with supplemented online learning environments, additional input in the form of teaching resources, course elements with social interactions and associated workload, and time-on-task in the sense of the "course and a half" syndrome (Garrison & Vaughan, 2008, p. 202) may explain the positive results.

1.3. Moderating effects

Various factors influence the effectiveness of blended learning, and these can be divided into the moderator categories of methods, practice and condition (Means et al., 2013). From a methodological perspective, learning outcomes may depend on the research design and the extent to which the conditions of the groups studied are comparable (learning objectives, learning environment, instructor equivalence, type of performance evaluation). Means et al. (2013) investigated several methodological factors, while Vo et al. (2017) studied the level of performance evaluation, and these variables were not found to be significant moderators. The present study has strict inclusion criteria regarding methodology; however, there is some variance in research design, type of performance evaluation and whether the same instructor taught in the groups studied – and these are coded and analysed as methodological moderator variables.

Regarding conditions, the main interest is whether the discipline or course level (undergraduate vs graduate course) influences effectiveness and also whether the effectiveness of blended learning depends on the publication year, i.e. whether there is a trend in effectiveness due to technical development. Means et al. (2013) and Bernard et al. (2014) examined these moderators and found no significance across the levels. Vo et al. (2017) focused on the subject and found a significantly higher mean effect size in STEM disciplines ($g = 0.50$) compared to that of non-STEM ($g = 0.21$). This review examines a limited period and therefore does not conduct a moderator analysis of the publication year but codes the study level and subject variables as possible moderators.

In the past, different moderators concerning the use of technology or types of media used in the experimental and control group were examined, and at times significant differences between the levels were found (Bernard et al., 2014). This particular analysis is not pursued further in our review because it can be assumed that similar technologies (LMS) and media were used in the experimental and control groups during the review period, regardless of the learning format. From a practice perspective, however, it is interesting to analyse whether the educational design moderates the learning effectiveness of blended learning. The moderator analysis by Spanjers et al. (2015) shows that the use of quizzes, in particular, has a significant and positive influence on the effectiveness and attractiveness of blended learning. The flipped classroom approach has received special attention in the educational design of blended learning (Thai, De Wever, & Valcke, 2017), and this has proven effective (Strelan, Osborn, & Palmer, 2020), also in combination with blended learning (Baepler, Walker, & Driessen, 2014). Consequently, we investigate whether the educational design of flipped classroom moderates the learning effectiveness of blended learning. In addition, the classroom reduction time is also analysed as a moderator variable for the practice.

1.4. Research question of the study

This study systematically examines whether classroom time can be reduced and replaced with an online learning environment to offer more flexibility in the learning process without compromising learning outcomes. It addresses the following two questions:

(1) What is the impact of blended learning on higher education student achievement in formal educational settings with a reduced classroom time of between 30 and 79 per cent?
(2) How do variables of study condition (subject, study level), methods (research design, instruction equivalence, performance evaluation) and practice (educational design, classroom-time reduction) moderate the overall average effect size?

2. Methodology

Methodologically, this systematic review follows the policies and guidelines of the Campbell Collaboration (Kugley et al., 2017) and the best practice guidelines for a meta-analysis by Bernard et al. (2014), and was conducted according to the quality standards for meta-analyses in the PRISMA statement by Moher, Liberati, Tetzlaff, Altman, and The Prisma Group (2009). We also report on the data source and literature research, the literature coding and analysis, and the meta-analytic methods used. Details concerning the calculations of effect sizes are given in Appendix A.

2.1. Inclusion/exclusion criteria

As discussed in the literature review, the conceptualization of blended learning is crucial as an inclusion criterion for selecting appropriate studies for meta-analysis. However, for a scientific comparison between the effectiveness of blended learning and conventional classroom instruction, teacher input and student workload should be about the same. To analyse the effect of blended learning in offering greater flexibility in the learning process within a review, only studies that partly replace classroom time with online learning sequences should be included. According to the online proportion of a learning environment, an appropriate differentiation of blended learning was developed by Allen et al. (2007) in their classification of traditional, web-facilitated, blended/hybrid

and online learning (Table 1). According to Allen et al. (2007), blended learning contains a proportion of between 30 and 79 per cent of online content delivery that must occur online rather than in the traditional format. As this review examines in detail the effects of replacing face-to-face classroom time with an online learning environment to offer more flexibility in the learning process, their definition of blended learning is considered accurate and suitable in this context. Hence, studies were identified that analyse student performance using blended learning (or equivalent conditions such as hybrid learning, online learning, distance education, digital learning, technology-enhanced learning, flexible learning) in the experimental condition and classroom instruction in the control group, with a reduction of 30–79% of classroom time in comparison with the control group. Student performance must also be evaluated in the same way across study conditions and objectively through measures such as final exams or tests (or a combination of both), projects, mid-term exams and final tests (integrated assessment) in a controlled design containing at least two independent samples (including real experiments and quasi-experiments). All studies must control for selection bias and sufficient statistical information must be provided to calculate effect size. Studies reporting self-ratings and attitudes and two or more group pre-experiments as well as one-group, pretest-posttest designs were omitted. Primary research had to be conducted in a formal higher education context (course or program leading to a certificate, diploma, or degree) with an intervention duration of at least one term. Studies focusing on informal education – defined as education without formal curriculum requirements and no formal degrees earned after completion (Cilasun, Demir-Şeker, Dincer, & Tekin-Koru, 2018) such as vocational training and professional certification, computer courses and language courses – were excluded. Since the internet has only become widely accessible to most learners in the last ten years (and with sufficient capacity to support learning, such as using learning videos in remote areas), we only included studies published from 2008 onwards in French, German or English in our meta-analysis. Table 2 summarises our meta-analysis criteria for inclusion and exclusion.

2.2. Data source and literature research

Based on the Campbell Method Guide "Searching for studies: a guide to information retrieval for Campbell systematic reviews" (Kugley et al., 2017), we primarily searched the following scientific electronic resources: Education Resources Information Centre (ERIC, $N = 1204$), FIS Bildung ($N = 1016$), PsycInfo ($N = 298$), PubMed ($N = 201$), Web of Science ($N = 280$), Lista EBSCO ($N = 64$). In searches for relevant studies, the following key terms with some variations were used (*blended OR online OR hybrid OR flexible OR "e-learning" OR elearning OR "web-augmented" OR flipped OR "Web*-supported*" OR distance OR computer-based) AND (learning OR instruction OR class) AND ("face-to-face" OR traditional OR "lecture-based" OR "classroom instruction" OR "lecture courses") AND (outcome OR performance OR effect OR impact OR achievement OR effectiveness)*.

The key terms were translated for searches in German (FIS Bildung) and French (HAL archives ouvertes) databases. Furthermore, Australian Education Index (AEI), HAL archives ouvertes, Learning & Technology Library (LearnTechLib), Networked Digital Library of Theses and Dissertations (NDLTD), OpenGrey, Springer Link, Taylor & Francis, Wiley Online Library were screened, and potential articles added to the data source. Additionally, articles cited in recent meta-analyses (Bernard et al., 2014; Borokhovski et al., 2016; Mahmud, 2018; Means et al., 2013; Schmid et al., 2014; Vo et al., 2017) were also added. Previous meta-analyses and reviews were also used for branching of leading journals in the field of educational technology (e.g. Internet and Higher Education). Finally, a search on Google Scholar, Open Grey and the website for not significant studies (www.nosignificantdifference.org) was carried out for grey literature. All articles published before 2008 were filtered out. The initial electronic database searches yielded 4035 articles, of which 2552 were excluded owing to duplication.

The selection process was designed in two stages and conducted by two researchers. Initially, the abstracts were screened and then the remaining studies were subjected to an in-depth, full-paper review. Selection was a criterion-by-criterion process (see Fig. 1).

If the first criterion was not met, that study was immediately excluded; for example, 239 studies in the abstract screening were excluded because they were not primary studies. Of course, it is also possible that these studies would not have fulfilled other criteria.

Table 2
Criteria for inclusion and exclusion in the meta-analysis.

Criterion	Inclusion	Exclusion
Type of course (according to Allen et al., 2007)	Blended learning.	Traditional, web-facilitated and online learning.
Reduced classroom time compared with the control group	30–79%	1–29% and >80%
Student performance	Objective tests and assessments.	Self-ratings, attitudes.
Study design	True and quasi-experiments containing at least two independent samples with control for selection bias.	Two or more group pre-experiments and one-group pretest-posttest designs
Statistical information	Sufficient statistical information for calculating effect sizes.	Studies offering insufficient statistical information.
Publication date	2008–2019	Before 2008.
Publication language	English, French or German (including abstracts).	All other languages.
Publication type	All original research, including grey literature	–
Study context	Higher education.	Vocational training, primary or secondary education.
Course or programme	Leading to a certification, diploma, or degree.	Without certification, diploma, or degree.
Intervention duration	At least one term.	Courses covering a few lessons.
Subjects	All subjects	–

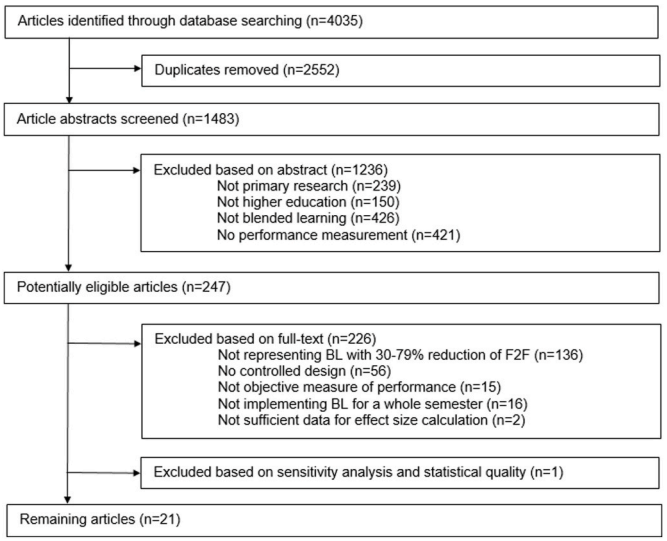


Fig. 1. Study selection process.

The inter-rater agreement of a hundred randomly selected studies rated by both researchers was 97.0%, Cohen's $K = 0.884$, $p < 0.001$. Six authors were requested to provide further information about their studies during the screening.

2.3. Sensitivity analysis

Extreme effect sizes – so-called outliers – can distort the weighted average effect size to a certain degree (Borenstein, 2009). Among the 22 studies, one outlier (Melton, Bland, & Chopak-Foss, 2009) with an effect size of 2.879 stands out in particular. Apart from this value being implausible, closer inspection of Table 2 in Melton et al. (2009) raised some serious doubts about the values reported there, as it was impossible to reproduce the T-statistics and *p*-values from the means and standard deviations provided. We, therefore, tested whether this study had strongly influenced the results, i.e. showed a high leverage value. For the random-effects model, with Melton et al. (2009) included, we obtained a small estimate of the effect size, and the result was not significant ($g =$

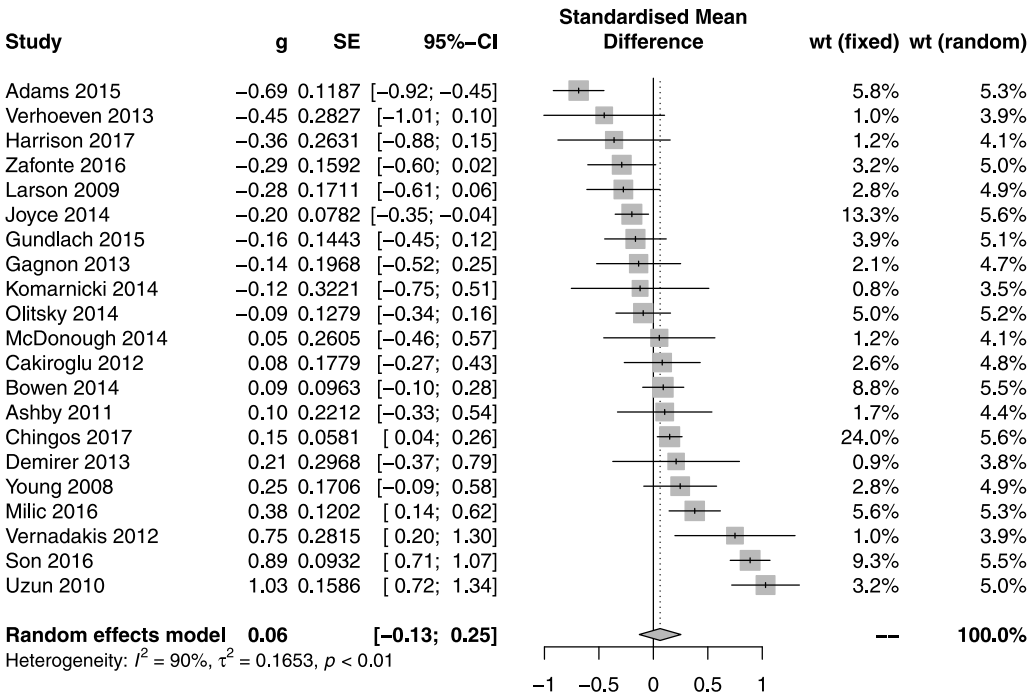


Fig. 2. Forest plot of effect sizes of the studies included ($k = 21$).

Table 3
Characteristics of studies included in the meta-analysis.

Author	Subject	STEM	Study Level	Reduction in Classroom Time (%)	Performance Evaluation	Educational Design	Instructor Equivalence	Research Design	*N* BL	*N* Control
Adams, Randall, and Traustadóttir (2015)	Microbiology	STEM	Undergraduate	50	Exam/posttest	Flipped	Yes	Quasi (randomized groups)	197	120
Ashby, Sadera, and McNary (2011)	Mathematics	STEM	Undergraduate	33	Course score	Other	Yes	Quasi (self-selection)	32	54
Bowen, Chingos, Lack, and Nygren (2014)	Statistics	STEM	Undergraduate	66	Exam/posttest	Flipped	No	Randomized	223	208
Cakiroglu (2012)	Programming	STEM	Undergraduate	60	Exam/posttest	Flipped	Not reported	Quasi (self-selection)	61	64
Chingos, Griffiths, Mulhern, and Spies (2017)	Various	Several	Undergraduate	30–79	Exam/posttest	Not reported	No	Quasi (randomized groups)	578	609
Demirer et al. (2013)	Multimedia Design	Other	Undergraduate	50	Exam/posttest	Other	Yes	Randomized	22	22
Gagnon, Gagnon, Desmartis, and Njoya (2013)	Nursing	Other	Undergraduate	54	Exam/posttest	Not reported	Yes	Randomized	52	50
Gundlach, Richards, Nelson, and Levesque-Bristol (2015)	Statistics	STEM	Undergraduate	50	Exam/posttest	Flipped	Yes	Quasi (self-selection)	56	331
Harrison, Saito, Markee, and Herzog (2017)	Mechanics	STEM	Undergraduate	50	Course score	Flipped	Yes	Quasi (self-selection)	24	35
Joyce, Crockett, Jaeger, Altindag, and O'Connell (2015)	Microeconomics	Other	Undergraduate	50	Course score	Other	Yes	Randomized	331	325
Komarnicki (2014)	Operations Management	Other	Undergraduate	50	Exam/posttest	Other	Yes	Quasi (self-selection)	18	19
Larson et al. (2009)	Management	Other	Undergraduate	31	Course score	Other	Yes	Quasi (self-selection)	79	60
McDonough, Roberts, and Hummel (2014)	Psychology	Other	Undergraduate	50	Course score	Not reported	Yes	Quasi (self-selection)	26	32
Milic et al. (2016)	Medical Statistics	STEM	Undergraduate	66	Course score	Not reported	Yes	Quasi (self-selection)	87	353
Olitsky et al. (2014)	Economics	Other	Undergraduate	33	Exam/posttest	Other	No	Quasi (self-selection)	82	236
Son, Narguizian, Beltz, and Desharnais (2016)	General Education Science	Other	Undergraduate	50	Course score	Flipped	No	Quasi (self-selection)	344	197
Uzun et al. (2010)	Computer Literacy	STEM	Not reported	50	Exam/posttest	Other	Not reported	Quasi (matching)	86	93
Verhoeven et al. (2013)	Microeconomics	Other	Undergraduate	50	Exam/posttest	Other	Yes	Quasi (self-selection)	36	19
Vernadakis, Giannousi, Derri, Michalopoulos, and Kioumourtzoglou (2012)	Physical Education	Other	Undergraduate	30–79	Course score	Other	Yes	Quasi (self-selection)	24	29
Young (2008)	Foreign Language	Other	Undergraduate	33	Exam/posttest	Other	Yes	Quasi (matching)	68	69
Zafonte et al. (2016)	APA style	Other	Undergraduate	50	Exam/posttest	Other	Not reported	Quasi (randomized groups)	79	79

0.19, $k = 22$, $SE = 0.13$, lower 95th = −0.07 and upper 95th = 0.45, $p = 0.16$). Excluding Melton et al. (2009) leads to much smaller estimate of the effect size (see Fig. 2), although still positive but not significant. The relatively large quantitative difference between results which include and exclude the study, combined with serious doubts about the statistical quality of the study, led to the exclusion of Melton et al. (2009) for further data analysis.

2.4. Computing effect sizes

Two researchers independently calculated the effect sizes for the different studies using two separate methods. One primarily used the online effect-size-calculator from the Campbell Collaboration, and the other calculated the effect sizes with R. For 19 papers, the same effect sizes were calculated (86.36 per cent agreement), and in three cases deviating results were found. This was due to the use of different primary data or different calculation methods, and these different results were discussed and reconciled. R (R Core Team, 2019) and functions implemented in the package meta (Schwarzer, 2020; Schwarzer, Carpenter, & Rücker, 2015) were used for most calculations and plots. To make the calculation of effect sizes transparent and reproducible, the relevant methodology is explained in detail below.

As the studies report the outcomes on different scales (usually grades), the standardised mean difference is used as the effect size. Regardless of the scale used, higher scores always correspond to better performance. The following presentation and definitions follow Borenstein (2009). In the most common version – the two-independent-groups design – the population effect size is defined by

$$\delta = \frac{\mu_{treat} - \mu_{control}}{\sigma}$$

where μ_{treat} is the true mean of scores for the treated population, $\mu_{control}$ is the true mean of the control population and σ is the standard deviation of scores within both populations (assumed to be equal for both populations). Hence, the effect size is given by the difference of means between treated and untreated populations divided by the standard deviation. A positive sign means that the treatment population has higher outcomes than the control population, and the absolute value is interpreted in units of standard deviations; for example, $\delta = -0.2$ means that the treatment population scores on average 0.2 standard deviations worse than the control population.

All the quantities and hence also δ are unobserved but can be estimated by plugging in the corresponding sample counterparts. Including a correction factor, this leads to an (approximately unbiased) estimator of δ, usually called *Hedges' g*:

$$g = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot \frac{\overline{Y}_{treat} - \overline{Y}_{control}}{sd_{pooled}},$$

where $\overline{Y}_{treat}$ is the observed mean of scores in the treatment group, $\overline{Y}_{control}$ is the observed mean of scores in the control group and $sd_{pooled}$is the pooled standard deviation from both samples. The *degrees of freedom* are given by $df = n_{treat} + n_{control} - 2$ for the simple case of a two-group comparison but are calculated using slightly different formulas in cases where, for example, regression coefficients are used. Hedge's g as defined above (with appropriate modifications for other designs) is used throughout this meta-analysis. Details on the calculations and formulas for standard errors and confidence intervals are given in Appendix A.

In some cases, g had to be calculated from other values given in the original studies, such as F test scores. Also, several studies used regression analysis or analysis of covariance to adjust for pre-existing differences in the two groups, mainly arising from selection bias.

Generally, – and particularly in non-randomized studies – it is preferable to adjust the difference in means to correct for other variables. For this meta-analysis, adjusted values were always preferred to unadjusted ones if both were reported in the original study. Since not all studies performed adjustments, this may seem like comparing apples and oranges. However, an adjusted estimate from a non-randomized study is still generally more easily comparable with an unadjusted estimate from a randomized study than an unadjusted estimate from a non-randomized study since both randomization and adjustment are methods for correcting for selection bias. Consequently, we attempted to use the best estimate available for every study where "best" means the most thorough correction for selection bias. Details about the calculation of effect sizes from adjusted values or results given in other forms are also shown in Appendix A.

3. Results

3.1. Description of included studies

The characteristics of the 21 studies included in our analysis are listed in Table 3. The total sample size was 2505 participants for the experimental blended learning condition and 3004 participants for the control condition. The studies assigned to a study level were all carried out at the undergraduate level, and the subject could be categorized into STEM ($N = 8$) and non-STEM subjects ($N = 12$). One study included both STEM and non-STEM courses and was not categorized. Fig. 2 shows the forest plot of the confidence intervals of the effect sizes for the 21 studies included.

3.2. Publication bias

Possible publication bias was assessed by a so-called funnel plot of estimated standard error versus effect size (Sutton, 2009), as shown in Fig. 3. It is generally assumed that higher-quality studies (corresponding to small standard errors) are usually published

regardless of whether the results are significant or not and irrespective of the size and direction of the estimated effect. At the same time, smaller studies with non-significant results or small and/or negative effects might remain unpublished, thereby biasing the meta-analysis towards larger positive effects. Apart from publication bias, smaller studies may be biased towards higher estimated effect sizes for other possible reasons, including poorer design or less appropriate statistical evaluation. An indicator for a bias of small studies towards a larger positive effect would be a strong asymmetry in the lower part of the funnel plot.

While no asymmetry was visually apparent, more formal tests for asymmetry (Chapter 3.3.2 in Sutton, 2009) were conducted as well, but these were not significant (linear regression test: $p = 0.69$, rank correlation test: 0.86). Overall, there is no indication of small-study effects. However, the number of studies included is not large, so it cannot be ruled out that an existing asymmetry was undetected due to low power. On the other hand, the relatively strict criteria for inclusion of studies may have prevented bias to some extent.

3.3. Summary effect size

As there were substantial differences in treatments, student populations, performance evaluation measures and other characteristics of the studies, a random-effects model is appropriate for summarising treatment effects in this context (Bernard, Borokhovski, & Tamim, 2019). This is also supported by a formal test of study heterogeneity ($Q = 201.90$, $df = 20$, $p < 0.0001$) and the estimated between-study standard deviation is considerable ($\tau = 0.41$).

For the $k = 21$ studies included, the results of the mixed-effects meta-analysis are shown in Table 4. The estimated effect size is positive but rather small and not significantly different from zero; correspondingly, the 95% confidence interval [-0.13, 0.25] rules out neither positive nor negative effects. However, the confidence interval strongly suggests that the effect size is not large. On Cohen's scale, an effect size of 0.2 is small, so at best, a very small negative or a very moderate positive effect is plausible.

3.4. Analysis of possible moderator variables

Although this meta-analysis uses strict inclusion criteria regarding the conceptualization of blended learning, the measurement of learning outcomes, the implementation period and the year of publication, considerable heterogeneity of the effect size was found between the studies that were included. Therefore, we investigated whether this could in part be explained by moderating variables. Variables of study condition (subject, study level), methods (research design, instruction equivalence, performance evaluation) and practice (educational design, classroom-time reduction) were coded. Unfortunately, moderator analysis of the study level could not be performed because all the studies bar one (unspecified) were conducted at the undergraduate level (see Table 3). Since the number of studies $k = 21$ is rather small, moderator analysis for the rest of the variables was performed separately using a by-group, random-effects meta-analysis, allowing the between-study variance to vary by group. Results of the per-group meta-analysis are shown in Table 5.

Regarding the study conditions and in line with Vo et al. (2017), the subjects were categorized into STEM and non-STEM subjects. One study included a mix of different subjects from STEM and non-STEM fields and was excluded from this analysis, leaving $k = 20$ studies. The subgroup differences are not significant ($Q(\mathrm{I}) = 0.00$, $df = 1$, $p = 0.95$), and the estimated treatment effects in both groups differ only slightly from the global estimate.

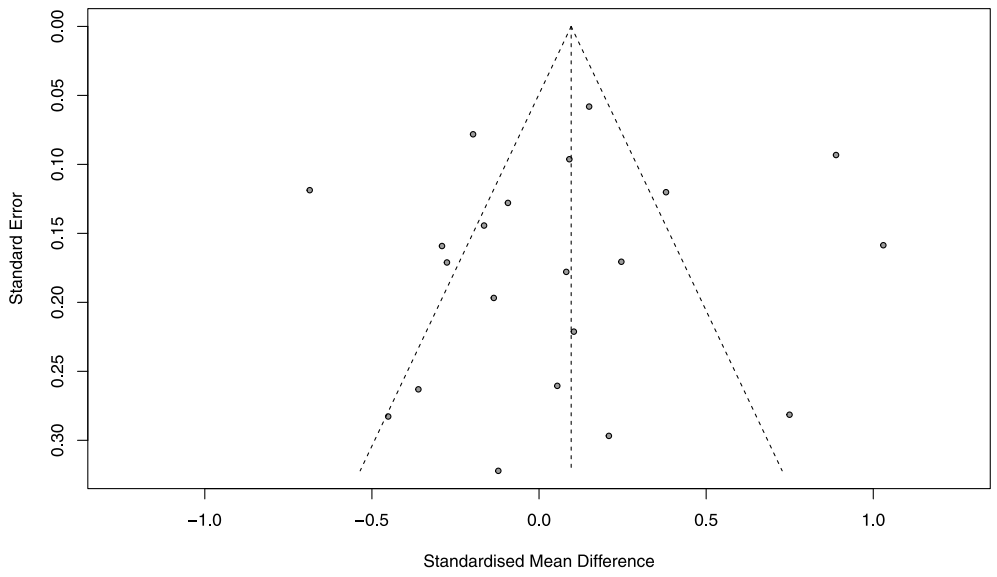


Fig. 3. Funnel plot of standard error according to Hedge's *g*.

Table 4
Summary effect size for random effect model and homogeneity statistics.

	Effect Size and Standard Error			Confidence Interval	
	K	*g*	*SE*	Lower 95th	Upper 95th
Effect	21	0.0621*	0.0976	−0.1292	0.2533
Heterogeneity	Q-total = 201.90, $df = 20$, $p < 0.0001$, $I^2 = 90.1\%$, $\tau^2 = 0.1653$				

Note: *$z = 0.64$, $p = 0.5246$.

Table 5
Analysis of potential moderator variables.

Moderator variable	Effect Size and SE			Confidence Interval	
	K	*g*	SE	Lower	Upper
Subject, $Q(I) = 0.0032$, $df = 1$, $p = 0.9549$					
STEM	8	0.0637	0.1847	−0.2982	0.4256
NON-STEM	12	0.0503	0.1500	−0.2437	0.3442
Research design, $Q(I) = 0.0497$, $df = 1$, $p = 0.8236$					
Randomized or matched	9	0.0355	0.1358	−0.2306	0.3017
Self-selection	12	0.0799	0.1453	−0.2049	0.3647
Instructor equivalence, $Q(I) = 2.4996$, $df = 2$, $p = 0.2866$					
Yes	14	−0.0613	0.1025	−0.2622	0.1397
No	4	0.2623	0.1992	−0.1282	0.6528
Not specified	3	0.2746	0.4066	−0.5223	1.0716
Performance evaluation, $Q(I) = 0.6112$, $df = 1$, $p = 0.4343$					
Exam/posttest	13	−0.0050	0.1108	−0.2222	0.2122
Course score	8	0.1716	0.1969	−0.2143	0.5576
Educational design, $Q(I) = 0.5511$, $df = 2$, $p = 0.7592$					
Flipped	6	−0.0161	0.2614	−0.5286	0.4963
Other	11	0.0789	0.1368	−0.1893	0.3471
Not specified	4	0.1606	0.0930	−0.0216	0.3429
Classroom-time reduction, $Q(I) = 2.1995$, $df = 2$, $p = 0.3330$					
30–40%	4	−0.0192	0.1117	−0.2382	0.1998
41–59%	12	−0.0117	0.1832	−0.3709	0.3474
60–79%	3	0.1904	0.1024	−0.0104	0.3911

As regards variables of methods, the research design was analysed as a moderator. While the inclusion criteria excluded studies not controlling for self-selection bias, there were considerable differences in how the issue was addressed. Four studies used a completely randomized design without any possibility of self-selection. Two studies used a randomized-groups design, where students signed up for classes, but classes were randomized to treatments. Two further studies allowed for self-selection but used matching techniques in the analysis to correct for self-selection bias. These nine studies were grouped together as they all used more effective means to control for self-selection bias. The 12 remaining studies used quasi-experimental designs that allowed for self-selection and did not explicitly correct the bias in the analysis. Instead, differences between control and treatment groups were assessed, either by descriptive analyses or formal tests for differences. These studies – with only weak control for selection bias – formed the second group. No studies were excluded as design information was available in every case. Subgroup differences are clearly not significant ($Q(I) = 0.05$, $df = 1$, $p = 0.82$), and the per-group estimates of the standardised mean difference are close to the results for all studies. Additionally, effect sizes did not vary significantly depending on whether the same instructor or instructors taught in the experimental blended learning and in the control conditions ($Q(I) = 2.50$, $df = 2$, $p = 0.29$) or depending on the type of performance evaluation ($Q(I) = 0.61$, $df = 1$, $p = 0.43$).

Although the flipped classroom is considered a promising didactic approach for blended learning (Thai et al., 2017), the effect size of blended learning is not moderated by this educational design compared to others ($Q(I) = 0.55$, $df = 2$, $p = 0.76$). As a second practice moderator variable, the amount of classroom-time reduction was analysed. While the replacement time is a continuous variable, it was decided to form groups rather than performing moderator analysis for a continuous variable (also known as meta-regression). The reasons for this are two-fold: in most of the included studies, 50% of classroom time had been replaced with online content, with only a few studies replacing a substantially smaller or larger fraction (often 33% or 66%, with very few other variants). Hence, the clustering around certain distinct values effectively makes the variable discrete. In addition, meta-regression assumes a linear and, therefore, monotone effect of replacement time, which should not be assumed a priori. Three groups were formed with online replacement rates of 30%–39%, 40%–59% and 60%–79%. Two studies had to be excluded – one included different replacement rates, and for the other, the exact replacement rate could not be determined. The estimated effect sizes differed between the groups with small and medium replacements, on the one hand, and the group with high replacement rates, on the other. However, the difference is not significant ($Q(I) = 2.20$, $df = 2$, $p = 0.33$).

Overall, no significant effects of the potential moderator variables considered were found. Power may generally be low with small numbers of studies, so a future meta-analysis featuring a larger number of studies might reveal relevant moderating effects. With a

greater number of studies being analysed, finer grouping could also be considered.

4. Discussion

In this meta-analysis, the estimated effect size (Hedge's *g*) for blended learning with reduced classroom time is positive, although not significantly different from zero. The confidence interval [-0.13, 0.25] suggests that overall differences between blended and conventional classroom learning are small, and at best, very small negative or very moderate positive effects are plausible. This means that despite reducing classroom time between 30 and 79 per cent, no adverse effects on learning outcomes were found. However, although on-average equivalent effectiveness was found, around half the studies reported positive and half negative results, and there were considerable differences between the studies. This implies that replacing classroom time with online learning does not always lead to a neutral or positive outcome.

Previous meta-analyses have found positive small to medium effect sizes for blended learning (Bernard et al., 2014; Means et al., 2013; Spanjers et al., 2015; Vo et al., 2017). At the same time, this review shows a considerably small – yet still positive – effect size of blended learning, although it is not significantly different from zero. There are several reasons for the differences in the magnitude of effect sizes between the previous reviews and this current meta-analysis. Firstly, and most importantly, there are methodological differences since this study employs much stricter inclusion criteria; only studies using objective measurements and controlled study designs were included. Additionally, classroom instruction needed to be replaced by online learning to some degree, thereby excluding studies where blended learning simply meant a technology-enhanced learning environment.

4.1. Limitations

All the studies included in this meta-analysis had to randomize the groups or control the characteristics of the groups. However, since it is hard to control all the characteristics in a real setting, differences in the input variables may have occurred. This is plausible because flexible study programs with reduced classroom time appeal to a particular student population with heavy demands on temporal and/or spatial flexibility, perhaps owing to an exacting job or domestic commitments. For example, academic achievement and competence in the area of self-regulation are significant predictors of student learning in both traditional and blended learning environments (Lim & Morris, 2009; Shea & Bidjerano, 2010). Concerning self-regulation, in particular, selection bias may exist insofar as students with better self-regulation skills choose flexible study programs. This personality trait positively impacts learning outcomes and should be given special attention in future comparative studies.

Apart from analysing the flipped classroom teaching format, this study did not examine the design of learning environments in greater detail. Teaching and learning are complex processes influenced by more than just the format. The quality of learning resources and tasks, the interaction between teachers and learners, and modes of feedback all have a major influence on learning success. In their literature review, Nortvig, Petersen, and Balle (2018) show that the following factors are of particular importance in blended learning environments – educator presence in online settings; interactions between students, teachers and content; and deliberate connections between online and offline activities and campus-related/practice-related activities. Many of the studies did not provide information about these features. Due to the small number of studies, no further analysis could be conducted on the process or input variables described above.

4.2. Methodical and practical issues in existing blended learning empirical studies

Currently, even in well-designed studies with sophisticated statistical analyses, findings are often not reported in a way that facilitates inclusion in a meta-analysis. For example, while ANCOVA or regression models allow correcting for the influence of factors other than treatment, to calculate an effect size it is also necessary to know the uncorrected sample standard deviations of both the treatment and control groups. In one case, a study had to be excluded because the direction of a non-significant effect was not provided. Also, in several studies, the numbers of levels of factor variables were not given, although these are required to calculate degrees of freedom. Sometimes, the values needed to calculate effect sizes and their standard errors could be reconstructed from other values, while, in other cases, the authors had to be asked for unreported values. In the worst cases, otherwise eligible studies had to be excluded because of missing information. Appendix A describes how some missing data issues were handled in this meta-analysis.

Research on blended learning would greatly benefit from more meta-analysis-friendly reporting. The inclusion of a greater number of relevant studies would facilitate the synthesis of evidence in this area and investigate the considerable heterogeneity of the results correctly, for example, by identifying and analysing other moderator variables. Higher standards for reporting results would also lead to more careful statistical analyses in primary studies and hence improve their quality.

4.3. Practical implications

Blended learning with reduced classroom time is not systematically more or less effective than conventional learning. This confirms current research findings and reviews, showing that it is not the learning format that is decisive for learning success but that teaching and learning are strongly circumstantial and context-dependent (Gillett-Swan, 2017; Nortvig et al., 2018). It is, therefore, hardly surprising that findings from the individual studies differ widely. Regarding the analysis of various forms of university course delivery, Hattie (2015) points out that of greater importance are how teachers – irrespective of the method of delivery – make their success criteria clear and offer challenge and feedback, coupled with the quality of the interaction among students and between students and

teacher. Therefore, from a practical point of view, there is no fundamentally superior teaching format; rather, this is contextual, and its effectiveness depends largely on the quality of its implementation.

At the same time and based on the current state of research and results of this study, it can also be concluded that educational institutions can offer blended learning study formats with reduced classroom time without fundamentally compromising the quality of the education being provided. Adapting study programs to offer greater flexibility and individualization enables more people with professional or domestic responsibilities to access higher education.

In many cases, however, the underlying motivation for offering flexible study formats is to reduce the cost of supplying higher education. Although efficiency is often mentioned as a key argument and driver for blended learning (Graham, 2019), there have been relatively few analyses of the cost-efficiency of blended learning (Galvis, 2018). These studies have not confirmed the widespread assumptions about the cost-efficiency of online learning; indeed, some indicate that the cost can be even higher than for traditional formats (Kennedy, Laurillard, Horan, & Charlton, 2015). It is, therefore, questionable to what extent it will be possible to reduce costs by replacing classroom time with online learning environments without compromising the quality and successful implementation of blended learning.

4.4. Future directions

The quality of a meta-analysis always depends on the quality of the studies it includes. As this study shows, the number of controlled studies in the field of blended learning is still limited and small for many moderator analyses. First and foremost, more primary studies of the highest methodological quality need to be conducted in various disciplines to validate the results further and investigate the effectiveness of blended learning in different disciplines and contexts. In particular, these studies should: (a) cover as many disciplines as possible. In this way, the frequently mentioned contextualization of blended learning can be examined to determine whether and to what extent it is suitable for the various subject areas. As Vo et al. (2017) mention, at least five studies per discipline should be available for this purpose. These studies should also (b) include different areas of competence. Depending on the subject areas, the desired learning outcomes may include more professional or generic competencies such as social and personal skills. A differentiation of competency effects would allow for further specification of learning formats, independent of the subject. In addition, these studies must (c) document the instructional design. As various authors mention (e.g. Hattie, 2015; Nortvig et al., 2018), the quality of a learning environment is determined less by its format (e.g. online, blended learning, classroom instruction) and more by its specific design. When explaining the context of a study, the instructional design for both the experimental and control conditions should be described precisely. Finally, they need to (d) report data appropriately. These studies should ensure that the reporting of data meets the requirements for inclusion in a meta-analysis. In particular, this includes uncorrected standard deviations as well as values for non-significant results.

In addition to enabling greater flexibility for students, blended learning with reduced classroom time is implemented particularly for cost reasons in times of a steadily increasing student population. However, evidence about the cost efficiency of online or blended learning is limited and requires more thorough analysis.

5. Conclusion

An increasing number of higher education institutions are considering replacing part of the face-to-face classroom instruction with an online learning environment by offering students a blended learning format. This meta-analysis, with its strict inclusion criteria regarding research design, measurement of learning outcomes and implementation of blended learning, shows that despite reduced classroom time between 30 and 79 per cent, such blended learning environments are not associated with poorer learning outcomes but are equivalent to conventional classroom instruction. Consequently, this study encourages higher education institutions to offer students greater flexibility in terms of time and place in their study programmes, thereby making higher education accessible to a broader section of society. However, it must be emphasized that there is considerable heterogeneity in the effect sizes between the various studies. From an empirical point of view, it is not yet possible to say in which disciplines or for which specific competencies a blended learning format is particularly suitable.

Furthermore, the quality of the instructional design could not be examined in either the blended learning format or the conventional classroom instruction format in detail. When a reduction of classroom time is made for cost reasons, this could impact the educational quality of the learning programme and negatively influence the outcome. When implementing blended learning featuring lower face-to-face classroom time, it is vital to ensure that empirical research findings on the successful design of blended learning are applied.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Claude Müller: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, preparation, and, Writing – review & editing. **Thoralf Mildenberger:** Methodology, Data curation, Formal analysis, Software, Writing – original draft, preparation, and, Writing – review & editing.

Appendix A. Details Regarding the Calculation of Effect Sizes

Calculation of hedges' g for two independent groups

The following presentation and definitions follow Borenstein (2009). In the most common version, the two-independent-groups design, the population effect size is defined by

$$\delta = \frac{\mu_{treat} - \mu_{control}}{\sigma}$$

where μ_{treat} is the true mean of scores for the treated population, $\mu_{control}$ is the true mean of the control population and σ is the standard deviation of scores within both populations (assumed to be equal for both populations).

All the quantities and hence also δ are unobserved but can be estimated by plugging in the corresponding sample counterparts. This leads to an estimator of δ, usually called *Cohen's d*:

$$d = \frac{\overline{Y}_{treat} - \overline{Y}_{control}}{sd_{pooled}},$$

where $\overline{Y}_{treat}$ is the observed mean of scores in the treatment group, $\overline{Y}_{control}$ is the observed mean of scores in the control group and sd_{pooled} is the pooled standard deviation from both samples:

$$sd_{pooled} = \sqrt{\frac{(n_{treat} - 1) \cdot sd^2_{treat} + (n_{control} - 1) \cdot sd^2_{control}}{n_{treat} + n_{control} - 2}}$$

where n_{treat} and $n_{control}$ are the sizes of the treatment and control groups, and sd_{treat} and $sd_{control}$ are the sample standard deviations. The pooled estimate is used because the true treatment and control standard deviations are assumed to be equal. It can be shown that d is not an unbiased estimate of δ, especially for small samples. Hence, a biased-corrected version known as *Hedges' g* is usually used:

$$g = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot d = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot \frac{\overline{Y}_{treat} - \overline{Y}_{control}}{sd_{pooled}},$$

where the *degrees of freedom* are given by $df = n_{treat} + n_{control} - 2$ for the simple case of a two-group comparison but are calculated using slightly different formulas in cases where, for example, regression coefficients are used. There are slightly different versions of the correction factor in the literature; the one used here is given in Borenstein (2009) and also used in Vo et al. (2017) and Bernard et al. (2014). Generally, the differences are small between different versions, and all versions of the correction factor are very close to 1 for larger samples.

The standard error (used for confidence intervals and p-values) of g is given (to a good approximation) by

$$SE_g = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot \sqrt{\frac{n_{treat} + n_{control}}{n_{treat} \cdot n_{control}} + \frac{d^2}{2(n_{treat} + n_{control})}}$$

In some studies, the effect sizes g had to be calculated differently from the formulas given above. In the following section, the most important cases are described briefly.

F-test for the comparison of three groups

In one study, three groups (blended learning, traditional instruction and fully online learning) were compared with the means of the groups given but not the standard deviations. The difference between the two means for blended learning and traditional learning was calculated, and the standard deviation was inferred from the F statistic for the global test for differences between the three groups using the formula given in Lipsey and Wilson (2001, Formula (17) in Table B10, p. 200):

$$sd_{pooled} = \sqrt{\frac{\sum_{i=1}^{k} n_i \overline{Y_i^2} - \frac{\left(\sum_{i=1}^{k} n_i \overline{Y_i}\right)^2}{\sum_{i=1}^{k} n_i}}{(k-1) \cdot F}}$$

Here, $k = 3$ is the number of groups, F is the reported F statistic, n_i is the size of the i-th group, $\overline{Y_i}$and $\overline{Y_i^2}$ are the means of responses and squared responses in the i-th group.

Note that while the third group (fully online learning) plays no role for the difference between the two other groups, it is only possible to estimate the standard deviation from all three groups (assuming all three population standard deviations are equal). Hence, in this case, data from the third group (otherwise irrelevant for the meta-analysis) play a role in calculating the effect size, but only for the denominator.

Means adjusted by regression or ANCOVA

Several studies control for other variables (e.g. gender or some measure of previous academic achievement) to adjust for possible differences between the treatment and control groups due to self-selection. This is achieved using a regression (also called analysis of covariance or ANCOVA in the case of adjusting for a single numerical variable). In fully randomized studies, this is not needed since any pre-existing differences between the two groups are due to chance. However, in some cases, precision may be increased when adjusting for covariates. The presentation generally follows Chapter 12.3.3 in Borenstein (2009). When the adjusted means are given, the following formula is used:

$$g = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot \frac{\overline{Y}^{adj}_{treat} - \overline{Y}^{adj}_{control}}{sd_{pooled}}$$

In some studies, instead of the two adjusted means, the treatment is given as a regression coefficient. For a factor variable with two levels, this is the same as the difference between the two means (taking into account the correct direction) and the following formula for g can be used:

$$g = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot \frac{\widehat{\beta}_{treatment}}{sd_{pooled}}$$

In one study, a so-called standardised regression coefficient was given; the regression coefficient after both the dependent and all independent variables were standardised to have a mean of zero and standard deviation of one. In this case, the non-standardised regression coefficient $\widehat{\beta}_{treatment}$ can be recovered by

$$\widehat{\beta}_{treatment} = \widehat{\beta}_{standardized} \cdot sd_Y \cdot \sqrt{\frac{(n_{treat} + n_{control})(n_{treat} + n_{control} - 1)}{n_{treat} \cdot n_{control}}}$$

where $\widehat{\beta}_{standardized}$ is the standardised regression coefficient and sd_Y is the (unconditional) sample standard deviation of the response (ignoring the grouping).

In all cases, the degrees of freedom are given by $df = n_{treat} + n_{control} - 2 - q$, where q is the number of co-variables adjusted for (factor variables count as $m–1$ variables where m is the number of factor levels). For some studies, q could not be determined exactly (see below).

Note that in any case, the adjustment is only made for the difference of means, not for the standard deviation. The unadjusted sample standard deviation was given in all the studies included here (at least for the control group), although it can be reconstructed from other quantities as well (cf. Borenstein, 2009). The standard error of g for adjusted differences also depends on this correlation and is given by

$$SE_g = \left(1 - \frac{3}{4 \cdot df - 1}\right) \cdot \sqrt{\frac{(n_{treat} + n_{control}) \cdot (1 - R^2)}{n_{treat} \cdot n_{control}} \cdot \frac{d^2}{2(n_{treat} + n_{control})}},$$

where R^2 is the multiple correlation between the variables adjusted for and the outcome. Since R^2 could not be inferred from the studies, it was set to 0, resulting in a slightly conservative estimate of the standard error.

Missing information

In several studies, some pieces of information were not provided but could be imputed or inferred indirectly using reasonable assumptions:

- *Dropout:* In some studies, the original group sizes at the beginning of the experiment were given, but some participants dropped out. If the total number of dropouts was given but not the numbers by group, it was assumed that dropout was proportional. For example, if the treatment and control groups had 50 and 100 participants at the beginning and 9 dropped out, the final group sizes were assumed to be 50–3 = 47 and 100–6 = 94, respectively. Assumed final group sizes were rounded to the nearest integer.
- *Standard deviation*: In some studies, only the sample standard deviation for the control group was given but was missing for the treatment group. In this case, since the overall assumption is that the population standard deviations are equal, all calculations were performed with the missing treatment standard deviation set equal to the control standard deviation.
- *Degrees of freedom*: The degrees of freedom in the correction terms for small-sample bias depend on the number of variables adjusted for in the regression or ANCOVA, and in particular also on the number of levels if factor variables are involved. If the number of variables was known (but not the number of levels for the factors), the number of variables was used as a lower bound on the number of estimated parameters q. Since the sample sizes were generally large in the studies affected, the corrections factor is close to 1 in any case, so the exact value of q is not needed.

For several other studies, essential information was missing and could not be reasonably imputed; for example, it is possible to calculate the absolute value of g from an F statistic, but if the group means are missing, it is not possible to infer the sign. In some cases,

the missing information could be obtained from the authors, but in other cases, the studies had to be excluded even when they satisfied all other selection criteria.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2021.100394.

References

Adams, A. E. M., Randall, S., & Traustadóttir, T. (2015). A tale of two sections: An experiment to compare the effectiveness of a hybrid versus a traditional lecture format in introductory microbiology. *14*(1), 1–8. https://doi.org/10.1187/cbe.14-08-0118

Allen, I. E., Seaman, J., & Garrett, R. (2007). *Blending in: The extent and promise of blended education in the United States.* Newsburyport, MA: Sloan Consortium.

Andrade, M. S., & Alden-Rivers, B. (2019). Developing a framework for sustainable growth of flexible learning opportunities. *Higher Education Pedagogies, 4*(1), 1–16. https://doi.org/10.1080/23752696.2018.1564879

Asarta, C. J., & Schmidt, J. R. (2015). The choice of reduced seat time in a blended course. *The Internet and Higher Education, 27*, 24–31. https://doi.org/10.1016/j.iheduc.2015.04.006

Ashby, J., Sadera, W. A., & McNary, S. W. (2011). Comparing student success between developmental math courses offered online, blended, and face-to-face. *The Journal of Interactive Online Learning, 10*(3), 128–140.

Baepler, P., Walker, J. D., & Driessen, M. (2014). It's not about seat time: Blending, flipping, and efficiency in active learning classrooms. *Computers & Education, 78*, 227–236. https://doi.org/10.1016/j.compedu.2014.06.006

Barnett, R. (2014). *Conditions of flexibility: Securing a more responsive higher education system.* York, UK: Higher Education Academy.

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education, 26*(1), 87–122. https://doi.org/10.1007/s12528-013-9077-3

Bernard, R. M., Borokhovski, E., & Tamim, R. M. (2019). The state of research on distance, online, and blended Learning: Meta-analyses and qualitative systematic reviews. In M. G. Moore, & W. C. Diehl (Eds.), *Handbook of distance education (4* (pp. 92–104). New York: Routledge.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd, pp. 221–235). New York: Sage.

Borokhovski, E., Bernard, R. M., Tamim, R. M., Schmid, R. F., & Sokolovskaya, A. (2016). Technology-supported student interaction in post-secondary education: A meta-analysis of designed versus contextual treatments. *Computers & Education, 96*, 15–28. https://doi.org/10.1016/j.compedu.2015.11.004

Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygren, T. I. (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management, 33*(1), 94–111. https://doi.org/10.1002/pam.21728

Cakiroglu, U. (2012). Comparison of novice programmers' performances: Blended versus face-to-face. *The Turkish Online Journal of Distance Education, 13*(3), 135–151.

Chaeruman, U. A., Wibawa, B., & Syahrial, Z. (2018). Determining the appropriate blend of blended learning: A formative research in the context of spada-Indonesia. *American Journal of Educational Research, 6*(3), 188–195. https://doi.org/10.12691/education-6-3-5

Chen, J., Wang, M., Kirschner, P. A., & Tsai, C.-C. (2018). The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. *Review of Educational Research, 88*(6), 799–843. https://doi.org/10.3102/0034654318791584

Chingos, M. M., Griffiths, R. J., Mulhern, C., & Spies, R. R. (2017). Interactive online learning on campus: Comparing students' outcomes in hybrid and traditional courses in the university system of Maryland. *The Journal of Higher Education, 88*(2), 210–233. https://doi.org/10.1080/00221546.2016.1244409

Cilasun, S. M., Demir-Şeker, S., Dincer, N. N., & Tekin-Koru, A. (2018). Adult education as a stepping-stone to better jobs. *An Analysis of the Adult Education Survey in Turkey, 68*(4), 316–346. https://doi.org/10.1177/0741713618783890

Demirer, V., & Sahin, I. (2013). Effect of blended learning environment on transfer of learning: An experimental study: Effect of BL on transfer of learning. *Journal of Computer Assisted Learning, 29*(6), 518–529. https://doi.org/10.1111/jcal.12009

Dziuban, C., Graham, C. R., Moskal, P. D., Norberg, A., & Sicilia, N. (2018). Blended learning: The new normal and emerging technologies. *International Journal of Educational Technology in Higher Education, 15*(3), 1–16. https://doi.org/10.1186/s41239-017-0087-5

Gagnon, M. P., Gagnon, J., Desmartis, M., & Njoya, M. (2013). The impact of blended teaching on knowledge, satisfaction, and self-directed learning in nursing undergraduates: A randomized, controlled trial. *Nursing Education Perspectives, 34*(6), 377–382. https://doi.org/10.5480/10-459

Galvis, Á. H. (2018). Supporting decision-making processes on blended learning in higher education: Literature and good practices review. *International Journal of Educational Technology in Higher Education, 15*(25), 1–38. https://doi.org/10.1186/s41239-018-0106-1

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education, 7*(2), 95–105. https://doi.org/10.1016/j.iheduc.2004.02.001

Garrison, D. R., & Vaughan, N. D. (2008). *Blended learning in higher education: Framework, principles, and guidelines:* San Francisco: Jossey-Bass.

Gillett-Swan, J. (2017). The challenges of online learning: Supporting and engaging the isolated learner. *Journal of Learning Design, 10*(1), 20–30. https://doi.org/10.5204/jld.v9i3.293

Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions. In C. J. Bonk, & C. R. Graham (Eds.), *The handbook of blended learning: Global perspectives, local designs* (pp. 3–21). San Francisco: Wiley & Sons.

Graham, C. R. (2019). Current research in blended learning. In M. G. Moore, & W. C. Diehl (Eds.), *Handbook of distance education* (4th ed., pp. 173–188). New York: Routledge.

Gundlach, E., Richards, K. A. R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and perceptions for web-augmented traditional, fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education, 23*(1), 1–33.

Harrison, D. J., Saito, L., Markee, N., & Herzog, S. (2017). Assessing the effectiveness of a hybrid-flipped model of learning on fluid mechanics instruction: Overall course performance, homework, and far- and near-transfer of learning. *European Journal of Engineering Education, 42*(6), 712–728. https://doi.org/10.1080/03043797.2016.1218826

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

Higher Education Academy. (2015). Framework for flexible learning in higher education. York, UK: Author. Retrieved from https://www.heacademy.ac.uk/system/files/downloads/flexible-learning-in-HE.pdf.

Hilliard, L. P., & Stewart, M. K. (2019). Time well spent: Creating a community of inquiry in blended first-year writing courses. *The Internet and Higher Education, 41*, 11–24. https://doi.org/10.1016/j.iheduc.2018.11.002

Hrastinski, S. (2019). What do we mean by blended learning? *TechTrends*, 1–6. https://doi.org/10.1007/s11528-019-00375-5

Joyce, T., Crockett, S., Jaeger, D. A., Altindag, O., & O'Connell, S. D. (2015). Does classroom time matter? *Economics of Education Review, 46*, 64–77. https://doi.org/10.1016/j.econedurev.2015.02.007

Kennedy, E., Laurillard, D., Horan, B., & Charlton, P. (2015). Making meaningful decisions about time, workload and pedagogy in the digital age: The course resource appraisal model. *Distance Education, 36*(2), 177–195. https://doi.org/10.1080/01587919.2015.1055920

Komarnicki, J. K. (2014). *How do they fare? A study of learning achievement and satisfaction with blended learning for traditional-age undergraduates at moderately selective colleges. ProQuest Dissertations and Theses*. Ann Arbor: Northeastern University.
Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A., Hammerstrøm, K., et al. (2017). *Searching for studies: A guide to information retrieval for Campbell systematic reviews*. https://doi.org/10.4073/cmg.2016.1
Larson, D. K., & Sung, C.-H. (2009). Comparing student performance: Online versus blended versus face-to-face. *Journal of Asynchronous Learning Networks, 13*(1), 31–42.
Lim, D. H., & Morris, M. L. (2009). Learner and instructional factors influencing learning outcomes within a blended learning environment. *Journal of Educational Technology & Society, 12*(4), 282–293.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. New York: Sage.
Mahmud, M. M. (2018). Technology and language–what works and what does not: A meta-analysis of blended learning research. *The Journal of AsiaTEFL, 15*(2), 365–382. https://doi.org/10.18823/asiatefl.2018.15.2.7.365
McDonough, C., Roberts, R. P., & Hummel, J. (2014). Online learning: Outcomes and satisfaction among underprepared students in an upper-level psychology course. *Online Journal of Distance Learning Administration, 17*(3).
Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record, 115*(3), 1–47.
Melton, B. F., Bland, H., & Chopak-Foss, J. (2009). Achievement and satisfaction in blended learning versus traditional general health course designs. *International Journal for the Scholarship of Teaching & Learning, 3*(1), 26. https://doi.org/10.20429/ijsotl.2009.030126
Milic, N. M., Trajkovic, G. Z., Bukumiric, Z. M., Cirkovic, A., Nikolic, I. M., Milin, J. S., … Marinkovic, J. M. (2016). Improving education in medical statistics: Implementing a blended learning model in the existing curriculum. *PloS One, 11*(2), Article e0148882. https://doi.org/10.1371/journal.pone.0148882
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097
Nortvig, A. M., Petersen, A. K., & Balle, S. H. (2018). A literature review of the factors influencing e-learning and blended learning in relation to learning outcome, student satisfaction and engagement. *Electronic Journal of e-Learning, 16*(1), 46–55.
OECD. (2017). *Key issues for digital transformation in the G20*. Paris: OECD Publishing. https://www.oecd.org/g20/key-issues-for-digital-transformation-in-the-g20.pdf.
OECD. (2019). *Going digital: Shaping policies, improving lives*. In Paris: OECD Publishing.
Olitsky, N. H., & Cosgrove, S. B. (2014). The effect of blended courses on student learning: Evidence from introductory economics courses. *International Review of Economics Education, 15*, 17–31. https://doi.org/10.1016/j.iree.2013.10.009
Orr, D., Luebcke, M., Schmidt, J. P., Ebner, M., Wannemacher, K., Ebner, M., et al. (2020). From lines of development to scenarios. In *Higher education landscape 2030: A trend analysis based on the AHEAD international horizon scanning* (pp. 5–24). Cham: Springer International Publishing.
Owston, R., & York, D. N. (2018). The nagging question when designing blended courses: Does the proportion of time devoted to online activities matter? *The Internet and Higher Education, 36*(Supplement C), 22–32. https://doi.org/10.1016/j.iheduc.2017.09.001
Peters, M. A., Rizvi, F., McCulloch, G., Gibbs, P., Gorur, R., Hong, M., et al. (2020). Reimagining the new pedagogical possibilities for universities post-Covid-19. *Educational Philosophy and Theory*, 1–44. https://doi.org/10.1080/00131857.2020.1777655
R Core Team. (2019). R: A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/.
Saichaie, K. (2020). Blended, flipped, and hybrid learning. *Definitions, Developments, and Directions*, (164), 95–104. https://doi.org/10.1002/tl.20428, 2020.
Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., et al. (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. *Computers & Education, 72*, 271–291. https://doi.org/10.1016/j.compedu.2013.11.002
Schwarzer, G. (2020). Meta: General package for meta-analysis R package version 4.11-0. Retrieved from https://CRAN.R-project.org/package=meta.
Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. London: Springer.
Shea, P., & Bidjerano, T. (2010). Learning presence: Towards a theory of self-efficacy, self-regulation, and the development of a communities of inquiry in online and blended learning environments. *Computers & Education, 55*(4), 1721–1731. https://doi.org/10.1016/j.compedu.2010.07.017
Smith, K., & Hill, J. (2019). Defining the nature of blended learning through its depiction in current research. *Higher Education Research and Development, 38*(2), 383–397. https://doi.org/10.1080/07294360.2018.1517732
So, H.-J., & Bonk, C. J. (2010). Examining the roles of blended learning approaches in computer-supported collaborative learning (CSCL) environments: A delphi study. *Journal of Educational Technology & Society, 13*(3), 189–200.
Son, J., Narguizian, P., Beltz, D., & Desharnais, R. (2016). Comparing physical, virtual, and hybrid flipped labs for general education biology. *Online Learning, 20*(3), 228–243.
Spanjers, I., Könings, K., Leppink, J., Verstegen, D., de Jong, N., Czabanowska, K., et al. (2015). The promised land of blended learning: Quizzes as a moderator. *Educational Research Review, 15*, 59–74.
Strelan, P., Osborn, A., & Palmer, E. (2020). The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. *Educational Research Review, 30*, 100314. https://doi.org/10.1016/j.edurev.2020.100314
Sutton, A. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd, pp. 435–452). New York: Sage.
Thai, N. T. T., De Wever, B., & Valcke, M. (2017). The impact of a flipped classroom design on learning performance in higher education: Looking for the best "blend" of lectures and guiding questions with feedback. *Computers & Education, 107*, 113–126. https://doi.org/10.1016/j.compedu.2017.01.003
Tucker, R., & Morris, G. (2012). By design: Negotiating flexible learning in the built environment discipline. *Research in Learning Technology, 20*(1). https://doi.org/10.3402/rlt.v20i0/14404
Uzun, A., & Senturk, A. (2010). Blending makes the difference: Comparison of blended and traditional instruction on students' performance and attitudes in computer literacy. *Contemporary Educational Technology, 1*(3), 196–207.
Verhoeven, P., & Rudchenko, T. (2013). Student performance in a principle of microeconomics course under hybrid and face-to-face delivery. *American Journal of Educational Research, 1*(10), 413–418. https://doi.org/10.12691/education-1-10-1
Vernadakis, N., Giannousi, M., Derri, V., Michalopoulos, M., & Kioumourtzoglou, E. (2012). The impact of blended and traditional instruction in students' performance. *Procedia Technology, 1*, 439–443. https://doi.org/10.1016/j.protcy.2012.02.098
Vo, H. M., Zhu, C., & Diep, N. A. (2017). The effect of blended learning on student performance at course-level in higher education: A meta-analysis. *Studies In Educational Evaluation, 53*(Supplement C), 17–28. https://doi.org/10.1016/j.stueduc.2017.01.002
Wade, W. (1994). Introduction. In W. Wade, K. Hodgkinson, A. Smith, & J. Arfield (Eds.), *Flexible learning in higher education* (pp. 12–17). Abdington: Routledge.
Young, D. J. (2008). An empirical investigation of the effects of blended learning on student outcomes in a redesigned intensive Spanish course. *CALICO Journal, 26*(1), 160–181.
Zafonte, M., & Parks-Stamm, E. J. (2016). Effective instruction in APA style in blended and face-to-face classrooms. *Scholarship of Teaching and Learning in Psychology, 2*(3), 208–218. https://doi.org/10.1037/stl0000064