The Effect of Blended Learning on Student Achievement in Undergraduate Statistics Education: A Systematic Review and Meta-Analysis

Anaïs Fopma
6199356

Bachelor's Project in Methodology and Statistics
Faculty of Social and Behavioural Sciences
Utrecht University

Supervisor: Dr. Charlotte Rietbergen
Second assessor: Dr. Vera Toepoel

[Word count: 7119]

24 June 2020

Abstract

This paper presents a research synthesis on the effect of blended learning on student achievement in undergraduate education in general, and undergraduate statistics education specifically. Studies included in the synthesis made a between-groups comparison on quantitatively measured student achievement between blended learning interventions and face-to-face learning in undergraduate courses. Blended learning was defined as integrating online learning with face-to-face learning and requiring active participation of students. A systematic review of the literature resulted in the selection of 15 studies. A meta-analysis was conducted on the effects found in these studies. In addition, a meta-analysis of the studies on statistics courses was conducted. Due to a large amount of heterogeneity found in both these analyses, the summary effects could not be interpreted as meaningful. To find an explanation for the heterogeneity, studies with larger sample sizes were excluded in an exploratory post-hoc analysis, as it is possible that blended learning has a more consistent effect in smaller courses or classes. The homogeneous results show a small summary effect of $d = 0.32$, $p < .0001$. Both this result and the heterogeneity found are in line with previous research, because of the likely presence of moderators. Blended learning does not seem to offer a convenient solution to increasing achievement at a high rate. However, it may offer a small positive effect in terms of its efficiency and other capabilities. This gives reason to implement it in undergraduate courses, albeit with careful consideration of context and moderators.

Keywords: blended learning, undergraduate education, student achievement, statistics education, innovations in education, online learning environments

Contents

1. Introduction

1.1. Problem Definition

The use of online learning environments and digital tools in education has been on the rise ever since access to the Internet has become common for everyone. In the early 2000s learning management systems such as Blackboard and Moodle were being used at an increasing rate (Picciano, 2014). Nowadays, blended learning technologies allow for more personal implementation. A wide variety of approaches, formats, and tools is present, the use of both social media and multimedia has increased, and students are dependent on portable devices for participating in course activities (Picciano, 2014).

With the popularity of the use of online technologies, tools, and environments as a replacement of parts of traditional, face-to-face education, comes the question on how effective this replacement or “blending” of different learning formats is for student achievement. This question on the influence of media on learning has been debated since the well-known Clark-Kozma debate or “media debate”. On the one side of this debate, the medium used in education is seen as merely the vehicle through which instruction is delivered, and therefore has no influence on the instruction (Clark, 1994). On the other side, Kozma (1994) argued that there is a potential relationship between media and learning, because of the interaction between the capabilities of media and the cognitive and social processes by which knowledge is constructed.

It can be valuable to look into the debated effectiveness of the use of technology and different media in the form of blended learning. In statistics subjects, students show more engagement when a variety of instructional methods is used (Biggs & Tang, 2011; Bhowmik, Meyer & Phillips, 2016). However, blended learning approaches can often be implemented without consideration of whether these approaches are truly more advantageous than the traditional format. Therefore, it is important to understand how to integrate technology effectively in statistics education (Tishkoveskaya & Lancaster, 2012).

Many studies have been conducted on the effectiveness of blended learning. Güzer and Caner (2014) categorized the development of these studies. They defined three classifications of articles on blended learning, namely the “first attempts” (1999-2001), “definition period” (2003-2006) and “popularity period” (2007-2009). Using this categorisation as a guideline, the following sections will explore the definition of blended learning and previous research on its effect on student achievement.

1.2. Defining Blended Learning

In the period of first attempts, various studies started using the term to describe different levels of integration of online learning, without giving it a clear definition. It was during the definition period that this was the topic of most articles related to blended learning. Disagreement existed over the exact definition of the term, and this still seems to be the case. Therefore, it is important to take a closer look at the different definitions and categorisations of blended learning. In their widely cited article on the subject, Osguthorpe and Graham (2003) propose that “Blended learning combines face-to-face with distance delivery systems…but it’s more than showing a page from a website on the classroom screen…those who use blended learning environments are trying to maximize the benefits of both face-to-face and online methods.” (p. 227). Garrison and Kanuka (2004) also explored the definition of blended learning and describe it to be “the thoughtful integration of classroom face-to-face learning experiences with online learning experiences” (p. 96). They point out that for this integration, a reconceptualization of both teaching and learning is necessary, and therefore, “no two blended learning designs are identical” (p. 97). This notion is in line with the argument of Oliver and Trigwell (2005) that the term “blended” should be either abandoned or radically reconceived. They analysed different “blends”, and point out that blended learning can mean anything ranging from combining e-learning with traditional learning to combining pedagogics, and criticise the term to only lead to confusion.

Because of this confounding between terms, it may prove useful to organise the different approaches of blended learning. Margulieux, McCracken and Catrambone (2016) developed a taxonomy for this purpose. The Mixed Instructional eXperience (MIX) taxonomy (Margulieux, McCracken & Catrambone, 2016) is directed towards defining courses in higher education that combine face-to-face and online learning. The taxonomy consists of two dimensions. The first is that of the delivery medium. On the one end of the dimension, delivery is given via the instructor, and on the other, via technology. The second is that of the instruction type. On the one end of this dimension, students receive the content, and on the other, they apply it. Combining the two dimensions results in the taxonomy with nine categories, as shown in Figure 1. The category an intervention falls into, depends on the ratios of delivery medium and instruction type.

The outer corners of the taxonomy are defined as the fundamentals. Courses are classified as fundamentals when they do not have a substantial portion (more than 25%) of instructional support from the other fundamentals. In an instructor-transmitted approach, an instructor delivers instructional support while students receive content. In a technology-transmitted approach, technology delivers instructional support while students receive content. In an instructor-mediated approach, an instructor delivers instructional support while students apply content. In a technology-mediated approach, technology delivers instructional support while students apply content.

Between these four fundamentals, there are three types of mixed instruction:

1. Combination: Face-to-face combination describes the pairing of instructor-transmitted and instructor-mediated instructional experiences. Online combination describes the pairing of technology-transmitted and technology-mediated instructional experiences.

2. Hybrid: Lecture hybrid describes courses in which students have instructional support for receiving content partially via an instructor and partially via technology. Practice hybrid describes courses in which students apply content with instructional support partially via an instructor and partially via technology.

3.Blended: The middle of the taxonomy is classified as the blended instructional experience, and it uses a substantial portion (at least 25%) of both delivery via an instructor and technology, and both receiving and applying content.

Both the hybrid sections and the blended section are often classified as blended learning in literature. In this research, however, the focus is on solely the middle of the taxonomy, and therefore the blended section.

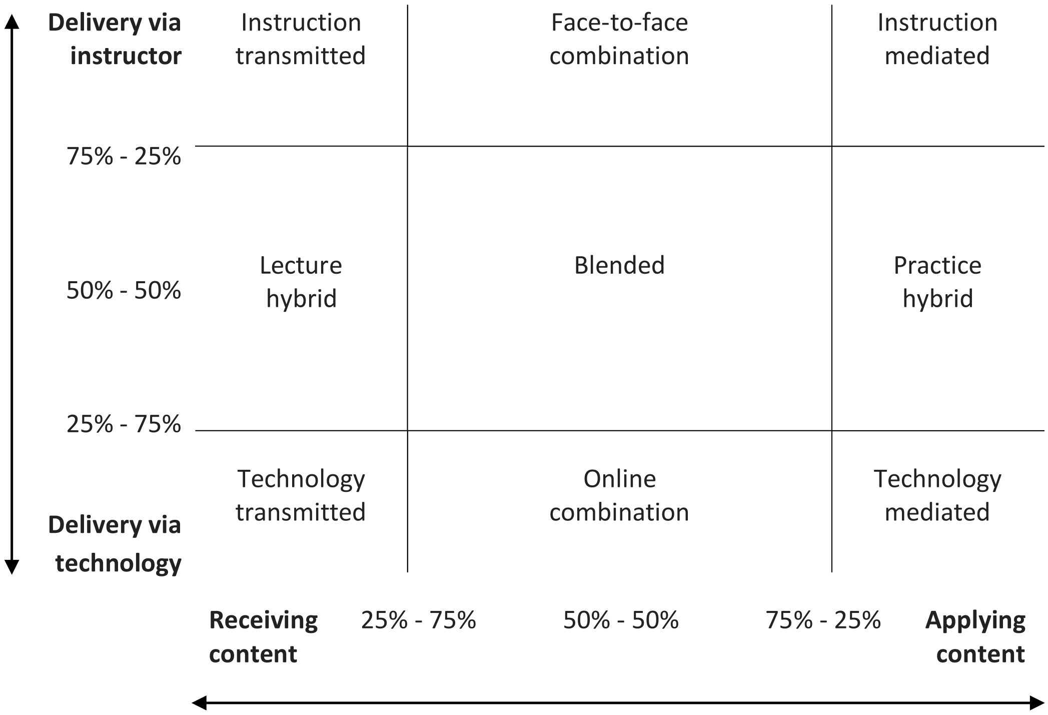


Figure 1. MIX taxonomy. Adapted from "A taxonomy to define courses that mix face-to-face and online learning.", by Margulieux, L. E., McCracken, W. M., & Catrambone, R., 2016, *Educational Research Review*, *19*, p. 110.

1.3. Related Research

In the third or "popularity period", the articles reviewed by Güzer and Caner (2014) can be categorised as studies that focus on the perceptions of participants on blended learning on the one hand, and studies that focus on the effectiveness of blended learning on the other. Since the "present" years (at the time of Güzer and Caner's study, the period of time between 2010 and 2012) blended learning as a topic continues to receive increasing attention. Studies are exploring many different areas and variables, for example outcomes such as teachers' professional development (Yeh et. al., 2011) or perceptions of academic staff (Donnely, 2010). Various research syntheses have been carried out on blended learning and the general use of online learning environments in education. A notable first mention is Hattie's (2009) *Visible Learning*, which synthesises over 800 meta-analyses on different influences on student achievement amongst school-aged students. Even though blended learning is not evaluated by Hattie (2009) as an attribute specifically, several implementations using technologies are, with effect sizes given as Cohen's *d*. For example, computer-assisted instruction was concluded to have a small effect of $d = 0.37$, web-based learning an effect of $d = 0.18$, and interactive video methods a medium effect of $d = 0.52$.

Means, Toyama, Murphy and Baki (2013) studied the effectiveness of both online and blended learning compared to traditional face-to-face instruction in the form of a meta-analysis of empirical literature. The analysis was conducted on 50 effects from 45 studies, made up of both experimental and quasi-experimental designs, with the average learner age ranging from 13 to 44 years. They found that students in online learning overall performed slightly better than those following face-to-face education, although this advantage was not significant. Comparing blended learning specifically to face-to-face education did show a significant advantage of the blended learning approach. A significant small effect size for the 23 blended versus face-to-face courses was found, with the weighted average of Hedges' g, $g^+ = 0.35$. However, a significant amount of heterogeneity was also found, and Means et al. (2013) concluded that studies using blended learning often also contain additional learning time, resources, and elements that encourage learner interaction and function as moderators. This could explain the positive outcomes of blended learning.

Bernard, Borokhovski, Schmid, Tamim, and Abrami (2014) conducted a meta-analysis of a sub-collection of comparative studies on blended learning from a larger systematic review of technology integration. The results show that in higher education, there is a significant small effect of $g^+ = 0.334$ on achievement outcomes.

McCutcheon, Lohan, Trayner and Martin (2015) conducted a systematic review on the impact of online and blended learning compared to face-to-face learning in the field of clinical skills in undergraduate nurse education. The authors identified 19 articles on the subject, 17 of which on online education and two on blended learning. The findings showed no significant difference between the effects of online learning and traditional education.

Vo, Zhu, and Diep analysed the effect of blended learning on student performance at course-level in higher education in their 2017 study. The meta-analysis was conducted using 51 effects and compared blended learning interventions to traditional classroom instruction. Blended learning had a significant small summary effect, $g^+ = 0.385$ compared to traditional teaching. A significantly higher mean effect size was found in STEM disciplines ($g^+ = 0.496$), compared to that of non-STEM disciplines ($g^+ = 0.210$). However, the weighted mean effect sizes revealed no significant differences regarding end-of-course assessment methods.

Overall, results seem to be supporting blended learning approaches with small to medium effect sizes. The differences between these and traditional approaches are not always large, and other variables might contribute to the positive effect of blended learning on student achievement. Güzer and Caner (2014) also indicate perspectives for the future of research on blended learning. They point out that in all the studies they reviewed, blended learning is perceived as useful, enjoyable, supportive and motivating. However, it is important to see that other factors should be taken into account in order to create a successful learning environment. For example, teachers should encourage student participation and collaboration, and the blending of face-to-face and online education should be planned out precisely. They conclude that in the future, studies should aim to guide teachers and administrators on how to create a blend that supports learning effectively.

1.4. Purposes of this Study

The aim of this study is to conduct a systematic review and meta-analysis on the effects of blended learning interventions on students achievement, for several reasons. Firstly, a systematic review and meta-analysis could provide a higher level of evidence by systematically analysing and combining the results of individual studies on this subject (OCEBM, 2011). Secondly, more precise conclusions could be derived when the type of blended intervention is more narrowly specified. Previous research syntheses on the subject have often been broad in defining blended learning, which could lead to inaccurate generalisations on different types of blended interventions. This study will therefore lay focus on a specific definition. In addition, some of the previous studies include a wide range of populations in their research. Therefore, this study will also be more specific in the sense that it is directed towards undergraduate education only. Thirdly, educational technology and the ways in which it is implemented develop quickly, and evaluations of older methods might have become outdated. It is important to focus on the most recent research and update the information available on interventions like blended learning. Lastly, although qualitative reviews and evaluations of blended learning practices are valuable, a systematic and quantitative review could offer more concrete insights in the effects on student achievement. This approach could therefore inform educators on the usefulness of blended learning and the implications for practice. Because the number of studies focusing specifically on the domain of statistics is limited, the synthesis will include studies on blended learning in undergraduate education in general, across all domains. If a sufficient amount of these studies is on statistics education, these will be analysed separately as well.

For the reasons mentioned above, this study will focus specifically on: (1) students in undergraduate (statistics) education, (2) blended learning interventions defined as the integration of an online learning environment requiring active participation of students, or which fall into the blended section of the MIX taxonomy (Figure 1), (3) a comparison of these interventions with completely traditional or face-to-face education, which falls into the section with more than 75% delivery via instructor of the MIX taxonomy (instruction transmitted, face-to-face combination, and instruction mediated) (Figure 1), and (4) the effects of these interventions on quantitatively measured student achievement. Putting these aspects together, the main research question to be answered is: “What is the effect of blended learning in comparison with traditional learning on the academic achievement of students in undergraduate (statistics) education?”

2. Method

2.1. Data Sources and Search Strategies

For composing the search query, it was decided that only terms for the population and the intervention would be specified. Whether the comparison was with traditional education would be evaluated with criteria during the screening process. Whether the outcome measure was one of student achievement was also not included in the search query, because this outcome has many possible definitions (e.g. grades, academic achievement, learning performance) and is not always explicitly mentioned in the title or abstract of articles. For the population, undergraduate courses were defined as courses in tertiary education in an academic direction. The main criterion for the population was thus the type of education followed, and demographic characteristics of the students themselves were not specified. The intervention needed to be defined as the integration of online learning within traditional, face-to-face education, which corresponds with the blended category of the MIX taxonomy (Figure 1).

For both the terms for the population and intervention, possible synonyms were selected based on a primary literature search. These operationalisations were then selected to be searched both as subjects headings in the respective thesauri of ERIC (Education Resources Information Center, 1966) and PsycINFO (American Psychological Association, 1967) and as keywords, abstract, and title. This was done through an iterative process of searching in the databases and consultation with a librarian. The time period in which the studies needed to be conducted, was set to 2015-2019, in order to sketch a picture of the impact of blended learning over the past 5 years. As technology uses in education are and have been quickly evolving, it was expected that this period would represent best the types of uses that are employed today. Because of limited time and resources for this study, no other registries or reference lists were searched. Due to the language restrictions of the authors, only articles in English were selected for the study. No geographical or cultural restrictions were taken. The logic grid and search query are presented in Appendix A. The search query was searched simultaneously in the databases of ERIC (Education Resources Information Center, 1966) and PsycINFO (American Psychological Association, 1967) on the Ovid platform.

2.2. Search Results

A summary of the screening process is depicted in Figure 2. The following sections describe the different screening phases: initial screening, abstract screening, full-text screening, and the risk of bias assessment.

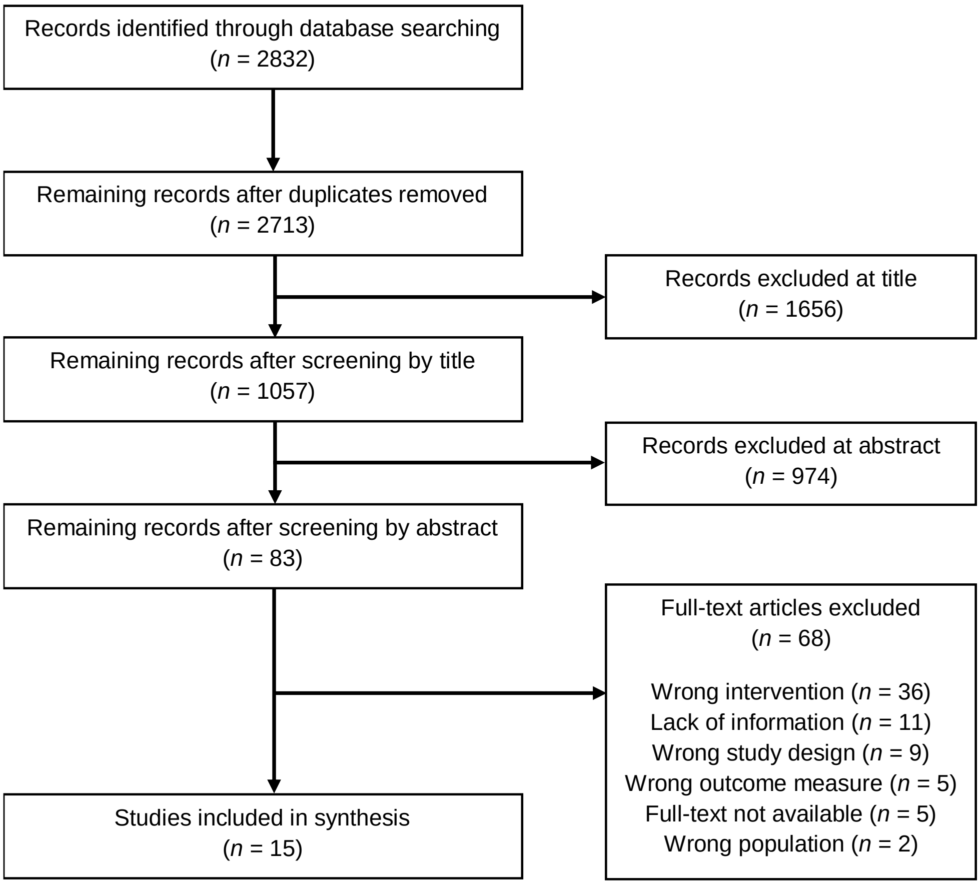

Records identified through database searching
(*n* = 2832)
Remaining records after duplicates removed
(*n* = 2713)
Records excluded at title
(*n* = 1656)
Remaining records after screening by title
(*n* = 1057)
Records excluded at abstract
(*n* = 974)
Remaining records after screening by abstract
(*n* = 83)
Full-text articles excluded
(*n* = 68)
Wrong intervention (*n* = 36)
Lack of information (*n* = 11)
Wrong study design (*n* = 9)
Wrong outcome measure (*n* = 5)
Full-text not available (*n* = 5)
Wrong population (*n* = 2)
Studies included in synthesis
(*n* = 15)

Figure 2. Flowchart of the screening process and systematic review results. Adapted from "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement", by Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., 2009, *PLoS Medicine, 6*(7), e1000097.

2.2.1. Initial screening. The initial search yielded 2832 articles combined from ERIC (Education Resources Information Center, 1966) and PsycINFO (American Psychological Association, 1967). Duplicates were removed using the deduplicating tool on the Ovid platform. More duplicates were removed in the web application Rayyan (Ouzzani, Hammady, Fedorowics, and Elmagarmid, 2016). Here, duplicates were suggested when there was a large amount of similarity between imported articles. For each of these suggestions (*n* = 246) it was manually decided by the first author which of these were duplicates. These were then removed, leaving 2713 records to be screened. All screening was carried out by the first author using Rayyan (Ouzzani et al., 2016). When uncertain about decisions, these were made in consultation with the second author. Initial screening was based on solely the titles of the articles. The author judged whether or not the article was related to blended learning in undergraduate education. Articles were excluded, for example, when the title made clear that the main intervention was not that of blended learning, when the intervention was one that was completely online (e.g. a MOOC), the article was describing a report or literature review, or that the study concerned education at a different level than the undergraduate level. This resulted in the selection of 1057 articles.

2.2.2. Abstract screening. For the second screening phase, the abstracts of the remaining articles were evaluated. Studies were eligible for inclusion if they met the following criteria:

1. The type of intervention concerned blended learning as defined before, and therefore an integration of online learning with face-to-face education. Studies on solely online education or in which blended learning functioned as a mere environment for a different intervention, were excluded.

2. The integration of online learning required the active participation of students, in accordance with the definition of blended learning in the MIX taxonomy (Figure 1). Studies in which the intervention concerned watching recorded lectures or reading texts online - without integrating practice or reflection - were excluded, as these would fall into the section of the lecture hybrid. This meant that studies on completely flipped classrooms were excluded as well, as these replace lectures by online materials and active participation takes place in face-to-face settings.

3. The study evaluated the effect of blended learning for students following undergraduate courses. Studies focusing on for example postgraduate education were excluded.

4. The intervention of blended learning was compared to a control condition of traditional, face-to-face learning. Studies in which blended learning was compared to flipped learning or online learning were excluded.

5. At least one of the outcome measures was a quantitative measure of student achievement. Studies for which the results were only related to affective variables (e.g. learning attitude or motivation), or studies using outcomes of self-reported measures, were excluded.

6. The study design was experimental or observational, and needed to include a between-groups comparison or correlation. Studies with an equal groups pre-posttest design were excluded.

When the abstract did not provide sufficient information to judge compliance with the criteria above, the article was included for the full-text screening phase. Abstract screening resulted in the selection of 83 articles.

2.2.3. Full-text screening. The records left after abstract screening were analysed in full text. Records that upon further investigation did not meet the aforementioned inclusion criteria, were excluded. In addition, the full-texts were judged against the following criterion:

7. Sufficient information was presented to calculate or derive effect size and variance, such as means, standard deviations, and sample sizes for both the treatment and control groups, or correlations.

The reasons for final exclusion are presented in Figure 2. The full-text screening left 15 records that were suitable to include in the synthesis. For the complete list of these references, see Appendix B.

2.2.4. Risk of bias assessment. The 15 records left were all assessed on their risk of bias. Non-randomised studies were assessed with the ROBINS-I tool (Sterne et al., 2016), and randomised studies were assessed with the RoB 2.0 tool (Sterne et al., 2019). The assessments were visualized using the robvis tool (McGuinness, 2019) and these visualisations are included in Appendix C. Most of the risk of bias is due to confounding or awareness of the intervention amongst outcome assessors. These problems are, however, difficult to prevent and common in educational studies like these, because the interventions are often implemented in a context where it is difficult to control for all possible confounding domains or where the educators themselves are often the researchers. Because of this, and because none of the records were classified as having a critical risk of bias, no records were excluded from the research synthesis based on these assessments.

2.3. Statistical Methods

The effect sizes and variances of the individual studies were manually calculated according to the design of the study and the information given. For studies that reported the mean for intervention (M_I), and control group (M_C), standard deviation for intervention (SD_I) and control group (SD_C), and sample size for intervention (n_I) and control group (n_C), the standardized mean difference (or Cohen's d) was calculated,

$$d = \frac{M_I - M_C}{SD_{within}}$$

$$SD_{within} = \sqrt{\frac{(n_I - 1)SD_I^2 + (n_C - 1)SD_C^2}{n_I + n_C - 2}}$$

with the variance calculated as:

$$V_d = \frac{n_I + n_C}{n_I n_C} + \frac{d^2}{2(n_I + n_C)}$$

For studies with a pre- and post-test, only the mean post-test scores were used in calculating the effect size and variance, because the standard deviations of changes were not reported in any of the studies. For one study that reported fail-pass proportions, the log odds ratio was calculated. Other studies reported on correlations. The log odds ratio and correlations were converted to Cohen's d. When a study reported on multiple relevant subgroups or outcomes, the effect sizes were combined across these. Some studies already reported the effect size, from these the variance was derived.

When all effect sizes and variances were calculated, random-effects meta-analyses were conducted using the Metafor package (Viechtbauer, 2010) in R (R Core Team, 2019). First, a meta-analysis of all 15 studies was conducted. Influence diagnostics were plotted to test for statistical outliers. Publication bias was evaluated by means of a funnel plot. In order to evaluate whether studies with a higher risk of bias and deviating influence diagnostics had a large influence on the results, a sensitivity analysis was conducted by excluding these studies. In addition, a separate meta-analysis was conducted of only the studies related to statistics education.

3. Results

3.1. Descriptive Information

An overview of the included studies and the relevant data extracted from these studies is presented in Table 1. Most studies employed a non-randomised parallel-group design (*n* = 9), other designs included were randomised (*n* = 3) or cohort studies (*n* = 3). The subjects of the courses in which the intervention took place was varied and generally speaking spread across the domains of business, psychology, mathematics, physics, chemistry, and biology. Five studies were related to statistics (and calculus) education. The aim of the individual studies was to compare either one or several blended learning or online learning interventions to a baseline. The total sample size for comparison between blended and control formats ranged from *n* = 44 to *n* = 2828. The outcome of student achievement was most often measured by means of a post-test or final exam. For each study, only the results, effect size, and variances concerning the comparison between a blended learning intervention and baseline or control group are presented in the table, as these are the ones relevant for the meta-analyses in this study.

Table 1

Overview of included studies and data-extraction.

Study	Design	Subject	Study Aim	Sample	Outcome	Results	Effect Size and Variance
Babaali & Gonzalez (2015)	Non-randomised parallel group	Pre-calculus	Assess the effectiveness of an online homework system	n = 123 I n = 122 C	Final exam score	I: M = 68.63, SD = 20.9 C: M = 53.70, SD = 24.0 Significant difference ($p < .001$)	*d for independent groups* d = 0.66364 V_d = 0.01723
Banditvilai (2016)	Non-randomised parallel group	Business English	Assess the effectiveness of e-learning lessons	n = 30 I n = 30 C	Pre- and post-test	Pre-test I: M = 27.01, SD = 5.33 Pre-test C: M = 27.53, SD = 5.65 No significant difference (p = .715) Post-test I: M = 41.43, SD = 4.7 Post-test C: M = 37.28, SD = 4.81 Significant difference (p = .001)	*d for independent groups* d = 0.18352 V_d = 0.06695

Table 1 (continued)

Study	Design	Subject	Study Aim	Sample	Outcome	Results	Effect Size and Variance
Bortnik et al. (2017)	Randomised parallel group	Analytical chemistry	Assess the effectiveness of a virtual lab-based inquiry learning environment	*n* = 25 I *n* = 25 C	Pre- and post-test	Pre-test I: *M* = 30.88, *SD* = 5.90 Pre-test C: *M* = 31.92, *SD* = 5.94 No significant difference (*p* = .5804) Post-test I: *M* = 39.2, *SD* = 5.02 Post-test C: *M* = 34.8, *SD* = 5.21 Significant difference (*p* = .011)	*d for independent groups* *d* = 0.16812 V_d = 0.08028
Botts et al. (2018)	Non-randomised parallel group	Quantitative literacy	Assess the effectiveness of blended learning	*n* = 100 I *n*= 97 C	Scores on three final exam questions (CLO_1, CLO_2, CLO_3)	I: CLO_1: *M* = 2.93, *SD* = 1.11 CLO_2: *M* = 2.67, *SD* = 1.2 CLO_3: *M* = 2.71, *SD* = 1.37 C: CLO_1: *M* = 2.84, *SD* = 1.3 CLO_2: *M* = 2.89, *SD* = 1.18 CLO_3: *M* = 2.93, *SD* = 1.44 No significant difference (*p* = .282)	*d for independent groups* CLO_1: *d* = 0.06175, V_d = 0.02047 CLO_2: *d* = -0.15530, V_d = 0.02037 CLO_3: *d* = -0.11146, V_d = 0.02034 *combine across outcomes assumed r = 1.00* $\bar{Y}$ = -0.06892 $V_{\bar{Y}}$ = 0.00777

Table 1 (continued)

Study	Design	Subject	Study Aim	Sample	Outcome	Results	Effect Size and Variance
Callahan (2016)	Cohort study	Calculus	Compare online homework vs. paper based homework	Fall *n* = 84 I *n* = 90 C Spring *n* = 62 I *n* = 68 C	Proportions of students passing the common final exam	Fall: Fail Pass I 32 51 C 22 56 Spring: Fail Pass I 22 40 C 19 49 No significant difference ($p > .05$)	*log odds ratio* Fall: *LogOddsRatio* = 0.46822 $V_{LogOddsRatio}$ = 0.11417 Spring: *LogOddsRatio* = 0.34954 $V_{LogOddsRatio}$ = 0.14349 *combine across subgroups* M = 0.41564 V_M = 0.06358 *convert log odds ratio to d* d = 0.22915 V_d = 0.01933
Dry et al. (2018)	Non-randomised parallel group	Psychology	Assess the effectiveness of an online adaptive learning tool	Group 1A *n* = 254 I *n* = 205 C Group 1B *n* = 425 I *n* = 90 C	Final exam score	I: 1A: M = 67.47, SD = 12.14 1B: M = 64.67, SD = 14.96 C: 1A: M = 60.68, SD = 13.38 1B: M = 51.20, SD = 14.67 Significant difference ($p < .01$)	*d for independent groups* 1A: d = 0.53429, V_d = 0.00913 1B: d = 0.90341, V_d = 0.01426 *combine across subgroups* M = 0.87658 V_M = 0.00556

Table 1 (continued)

Study	Design	Subject	Study Aim	Sample	Outcome	Results	Effect Size and Variance
Förster et al. (2018)	Cohort study	Advanced statistics	Assess the effectiveness of electronic quizzes	*n* = 762	Final course grade	$R^2 = 0.19$	*correlation* $r = 0.43589$ $V_r = 0.00086$ *convert r to d* $d = 0.96864$ $V_d = 0.00648$
Goette et al. (2017)	Non-randomised parallel group	Abnormal psychology	Assess the effectiveness of blended learning	*n* = 49 I *n* = 65 C	Final exam score	I: *M* = 78 *SD* = 21.0 C: *M* = 75 , *SD* = 24.7	*d for independent groups* $d = 0.12938$ $V_d = 0.03587$
Jonsdottir et al. (2017)	Repeated randomised crossover	Introductory statistics	Assess the effectiveness of web-based homework	Not given	Final exam score	$r = 0.416$	*correlation* $r = 0.416$ (given) $V_r = 0.02496$ *convert to r to d* $d = 0.914925$ $V_d = 0.17658$

Table 1 (continued)

Study	Design	Subject	Study aim	Sample	Outcome	Results	Effect size and variance
Molnar (2017)	Non-randomised parallel group	Computer applications in business	Assess the effectiveness of a flipped classroom	*n* = 58 I *n* = 59 C	Final exam score	I: *M* = 85.741, *SD* = 7.6767 C: *M* = 86.729, *SD* = 7.8275 No significant difference (*p* = .492)	*d for independent groups* *d* = -0.12743 V_d = 0.03426
Potter (2015)	Non-randomised parallel group	Management	Assess the effectiveness of hybrid teaching (40% online)	*n* = 50 I *n* = 50 C	Final exam score	I: *M* = 0.831, *SD* = 0.095 C: *M* = 0.782, *SD* = 0.097	*d for independent groups* *d* = 0.51039 V_d = 0.04130
Powers et al. (2016)	Non-randomised parallel group	Introductory psychology	Assess the effectiveness of an online learning environment	*n* = 291 I *n* = 439 C	Final exam score	I: *M* = 78.1, *SD* = 15.7 C: *M* = 82.9, *SD* = 12.7	*d for independent groups* *d* = -0.34353 V_d = 0.00580

Table 1 (continued)

Study	Design	Subject	Study aim	Sample	Outcome	Results	Effect size and variance
Reyneke et al. (2018)	Cohort study	Introductory business statistics	Compare three methods: (I_1) Phase 1: online homework system (I_2) Phase 2: online homework system in combination with flipped classroom (C) Baseline	*n* = 1343 I_1 *n* = 1466 I_2 *n* = 1485 C	Final exam score	I_1: *M* = 59.24, *SD* = 13.93 C: *M* = 58.35, *SD* = 14.36 No significant difference	*d for independent groups* *d* = 0.08531 V_d = 0.00142
Yapici (2016)	Non-randomised parallel group	Biology	Assess the effectiveness of a blended cooperative learning environment	*n* = 30 I *n* = 31 C	Pre- and post-test	Pre-test I: *M* = 9.73, *SD* = 1.50 Pre-test C: *M* =9.40, *SD* = 1.38 No significant difference (*p* = .87) Post-test I: *M* = 18.33, *SD* = 1.37 Post-test C: *M* = 13.50, *SD* = 0.9 Significant difference (*p* < .05) *d* = 0.75	*d* = 0.75 (given) V_d = 0.07020

Note. I = Intervention, C = Control. Pre-test results are shown for the purpose of comparison between intervention and control. Change scores between pre-test and post-test were not used in the calculation of effect size and variance. Instead, only the difference between intervention and control on the post-test was used.

3.2. Undergraduate Education

A random-effects meta-analysis was conducted on the effect sizes for student achievement. The forest plot depicted in Figure 3 shows the estimated effect sizes and 95% confidence intervals for each study, and the summary effect and corresponding 95% confidence interval. The results show a significant summary effect, with $d = 0.33$ [0.11; 0.55], and $p < .01$. However, the results also show a significant amount of heterogeneity, $Q(df = 14) = 265.54$, $p < .0001$ and $I^2 = 86.80\%$.

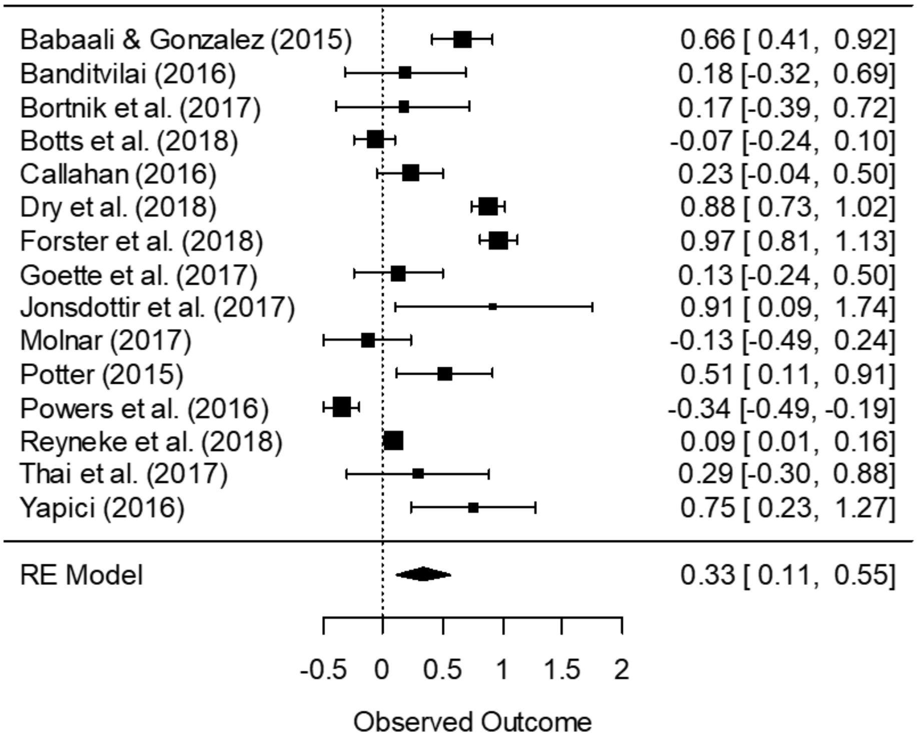


Figure 3. Forest plot of the estimated effect sizes of all studies.

3.2.1. Evaluation of publication bias. The presence of publication bias was evaluated by means of the funnel plot shown in Figure 4. In accordance with what would be expected, the studies are distributed symmetrically around the mean effect size. Contrary to what would be expected, the studies with larger sample sizes show more dispersion in effect sizes than the studies with smaller sample sizes, and are spread across the top of the graph. This causes the results to deviate from the funnel shape. Therefore, it is possible that publication bias is present. However, according to Zwetsloot et al. (2017), funnel plots of the standardised mean difference plotted against the standard error are susceptible to distortion, especially when the primary studies have a small sample size and when an intervention effect is present. This means that the existence and extent of publication bias can be overestimated when interpreting the funnel plot (Zwetsloot et al., 2017). It is also possible that studies with smaller sample sizes that did not find significant results have not been published. This would cause studies with both negative and positive results to be missing across the bottom of the graph. This would not necessarily mean that publication bias is present, as this is often the case when mainly smaller studies with negative effects are missing. However, it would mean that the current funnel plot presents an inaccurate image of the distribution of the results. Overall, it is difficult to interpret whether publication bias is present.

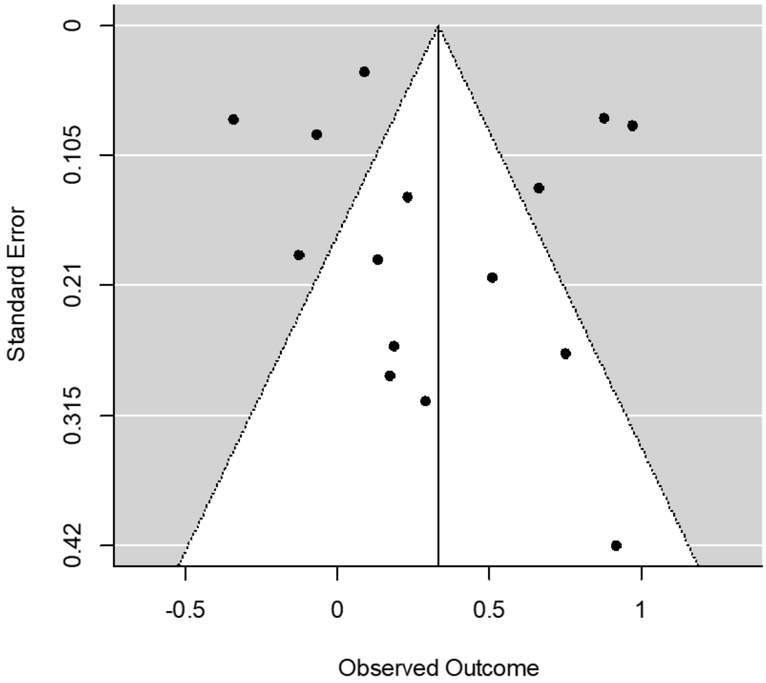


Figure 4. Funnel plot of all studies.

3.2.2. Sensitivity analysis. To evaluate whether studies with a higher risk of bias and deviating scores on the influence diagnostics had a large influence on the results, a sensitivity analysis was conducted excluding the following studies: Powers et al. (2016), Förster et al. (2018), Goette et al. (2017), and Molnar (2017). These studies were chosen, because they all scored a serious risk of bias. In addition, both Powers et al. (2016) and Förster et al. (2018) both show more deviation in the influence diagnostics (Appendix D). It is also noticeable that Powers et al. (2016) is the only study reporting a significant negative result. Possible explanations for this result given by the author, are that students in the intervention only completed 66% of their required homework in the online learning environment, and that both instructors and students reported criticism on the technology used, mainly related to time management and technical difficulties (Powers, 2016). The sensitivity analysis, however, showed no relevant difference in the results. The results of the analysis are included in Appendix E.

3.2.3. Post-hoc analyses. When I^2 is large, it is advised to try to explain the heterogeneity (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 119). In order to do this, several subgroup analyses were conducted based on different characteristics of the studies, such as the ratio of time spent studying between control and intervention, or whether participation in the intervention was mandatory or optional. However, none of these characteristics could explain the heterogeneity, as the results of these analyses did not show lower amounts of heterogeneity than the amount found in the main analysis. Besides this, it was not always possible to categorise the studies into subgroups because of a lack of reported information.

In an additional attempt to explain the heterogeneity, a post-hoc analysis was conducted through an exploratory approach. It was notable that the studies that contributed to the heterogeneity the most, were those with larger sample sizes. In an educational context, the larger sample sizes correspond with larger course sizes and, possibly, class sizes. Therefore, in a post-hoc analysis, the studies with larger sample sizes (and deviating from the funnel shape) were excluded: Babaali and Gonzalez (2015), Botts et al. (2018), Dry et al. (2018), Förster et al. (2018), Molnar (2017), Powers et al. (2016), and Reyneke et al. (2018). The results show a significant summary effect, with d = 0.32 [0.16; 0.47], and $p < .0001$. The test for heterogeneity was not significant, with $Q(df = 7) = 7.53$, $p = .3757$ and $I^2 = 0.00\%$. The forest plot is shown in Figure 5. The funnel plot is shown in Figure 6 for the purpose of comparison with the funnel plot for all studies. It has not been used to evaluate publication bias, as this should only be done when at least ten studies are included in the meta-analysis (Higgins & Green, 2011). Because these eight studies show more homogeneous results, it is possible that the studies on smaller courses form a subgroup. The seven excluded studies, however, did not form a homogeneous subgroup. For Powers et al. (2016), Molnar (2017), and Förster et al. (2018) exclusion can be explained by the reasons mentioned in the first sensitivity analysis. An additional explanation for exclusion and heterogeneity amongst the excluded studies, could be that the effects of blended learning differ more strongly in larger courses, possibly because moderators have more influence here.

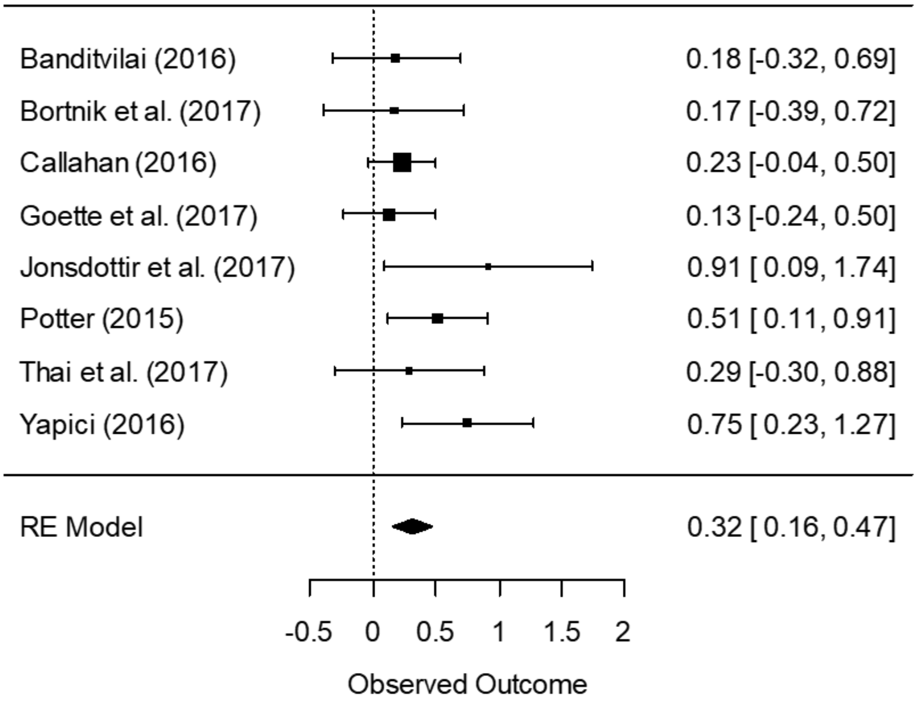


Figure 5. Forest plot of the estimated effect sizes in the post-hoc analysis.

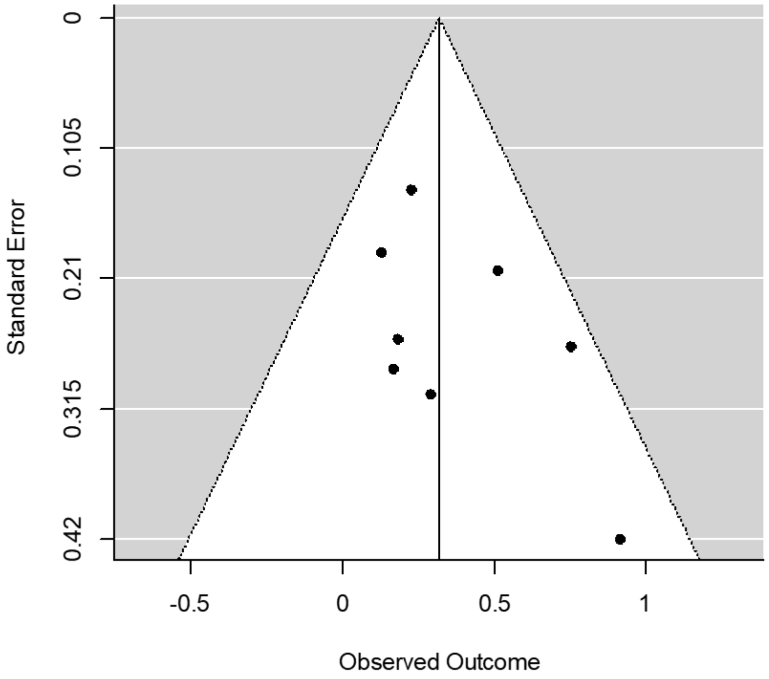


Figure 6. Funnel plot for the post-hoc analysis.

3.3 Statistics Education

A separate meta-analysis of the studies related to statistics education was conducted. The forest plot depicted in Figure 7 shows the estimated effect sizes and 95% confidence intervals of each study, and the summary effect and corresponding 95% confidence interval. The results show a significant summary effect, with $d = 0.53$ [0.16; 0.90], and $p < .01$. However, the results also show a significant amount of heterogeneity, $Q(df = 4) = 110.78$, $p < .0001$ and $I^2 = 94.12\%$. With only five studies, a funnel plot was not evaluated and it was not feasible to find an explanation for the heterogeneity through sensitivity or subgroup analyses.

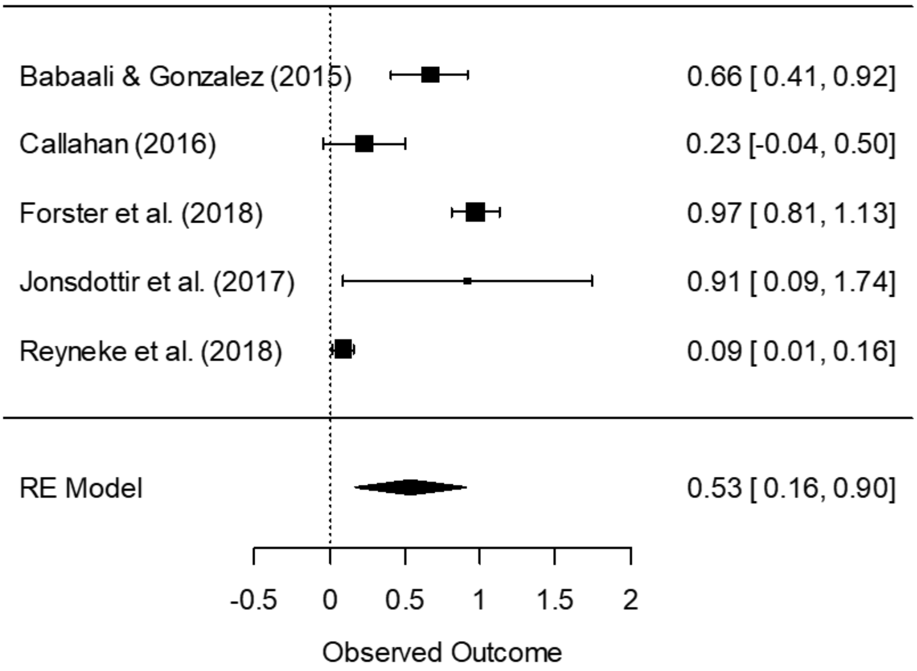


Figure 7. Forest plot of the estimated effect sizes of the studies related to statistics education.

4. Discussion

4.1. Main Findings

This study aimed to synthesise the results of studies on the effect of blended learning in both general undergraduate education and undergraduate statistics education. First, the results for general undergraduate education found a significant summary effect of *d* = 0.33. However, a large amount of heterogeneity was found. Therefore, the summary effect should not be interpreted as meaningful (Hak, Van Rhee, & Suurmond, 2016).

In order to find possible explanations for the large amount of heterogeneity found, exploratory post-hoc analyses were carried out. It was notable that the studies that contributed to the heterogeneity the most, were those with larger sample sizes. These correspond with larger course sizes and, possibly, class sizes. Therefore, in one post-hoc analysis, the studies with large sample sizes were excluded. The results show a small summary effect of *d* = 0.32 with minimal heterogeneity. In an educational context, guidelines to interpret effect sizes given by Hattie (2009) follow four categories of effects. His guidelines are based on the claim that influences in education are relative, and that the success of an innovation should be judged relative to *d* = 0.40, as opposed to *d* = 0.0. The reason for this is that *d* = 0.40 summarizes the typical effect of all possible influences and alternative innovations in education (Hattie, 2009). Accordingly, he points out that an effect should not be compared to a situation where there is no effect (*d* = 0.0), but to *d* = 0.40, because we want to learn whether an innovation is more successful than others. Average effects between -0.20 and 0.00 are reverse effects, which decrease achievement. Those between 0.00 and 0.15 are developmental effects, which describe what students could probably achieve if there was no schooling. Those between 0.15 and 0.40 are teacher effects, which means that influences are similar to what teachers can accomplish in a typical year of schooling. Finally, those higher than 0.40 are considered to be in the zone of desired effects, meaning that attributes with these effects have the greatest impact on student achievement outcomes, and that effects from the innovation are better for students than what they would achieve if they had received alternative innovations. The summary effect found in this study falls in the range of teacher effects. This would mean that blended learning interventions in undergraduate education do not have a relevant effect on student achievement, but do not decrease achievement either. Furthermore, Hattie mentions that "…it is not as simple as saying that all effects below *d* = 0.40 are not worth having (it depends on costs, interaction effects, and so on)" (2009, p. 16).

It is important to look at the possible reasons exclusion of the larger studies led to a more homogeneous set of results. This selection of eight studies showing a homogeneous result, have small sample sizes in common. In an educational context, these small sample sizes correspond with smaller courses, and possibly, smaller class sizes. The studies included in the selection had control or intervention groups with no more than 90 students, whereas the studies excluded from the selection had at least one control or intervention group with a minimum of 100 students in a course. In the studies with smaller sample sizes, often class sizes of between 20 and 30 students are reported. In the studies with larger sample sizes, class sizes often seem to be larger. This suggests that blended learning has a more consistent effect in courses with smaller class sizes. This can be supported by the negative effects found of larger class sizes in higher education on student performance (De Paola, Ponzo, & Scoppa, 2013; Jepsen, 2015) and student satisfaction (Monks & Schmidt, 2011; Gannaway, Green, &

Mertova, 2018). It is imaginable that students in smaller groups are more inclined to follow the blended learning interventions as intended, thus resulting in the benefits of the blended intervention to show. For example, larger classes in higher education have lower classmate supportiveness, student preparedness, and class participation compared with smaller classes (Bai & Chang, 2016). However, the larger, excluded studies show both negative and positive effects. The heterogeneity might indicate that these negative effects are controlled for in some of the larger studies, for example because of incentives to complete homework. For example, Powers et al. (2016) likely showed a significant negative result, because students did not complete all of the required online homework – and were not incentivised to do so. It is, however, not always possible to determine what class sizes were from the information given. It is not always reported whether the courses were divided into sections, and if so, whether these sections were always separate or followed instruction together as well. If the sample sizes do not correspond with class sizes, it can at least be said that they correspond with course sizes. It is possible that it is easier to implement blended interventions in smaller courses in a more controlled manner, from an organisational perspective.

Finally, the summary effect of d = 0.53 for undergraduate statistics education found in this study, would fall into the zone of desired effects. However, a large amount of heterogeneity was found here as well, and therefore this effect can also not be interpreted as meaningful. However, implications derived from the other results found in this study, can apply to undergraduate statistics courses as well, as is discussed in section 4.3.

4.2. Limitations

This research has several limitations. First of all, only one author conducted the screening process, because of limited resources. This caused the screening not to be checked for reliability, which entails that it is possible that studies that were suitable were unjustly excluded from the final selection. Second of all, publication bias could be present among the studies included in the research synthesis. This may partially be explained by the fact that only two databases were used to search for literature. No grey literature, unpublished studies, or other literature was used. It is also worth noting that some studies that initially seemed eligible, were excluded because of a lack of information presented in the articles. Third of all, the large amount of heterogeneity within the final selection seems to point out that the studies are not all comparable. Despite the set definition and strict criteria, it seems that there are still large differences between the studies – at least between those with larger sample sizes. However, the relatively small number of studies also makes it more difficult to explain this heterogeneity.

4.3. Implications for Theory and Practice

To a large extent, the findings of this study are in line with previous research. Results of other meta-analyses on blended learning often show small to medium effect sizes. Moreover, the heterogeneity found compares to the results of Means et al. (2013), Bernard et al. (2014), and Vo et al. (2017). Means et al. (2013) point out that the presence of heterogeneity and moderators likely mean that the advantage of treatment conditions stems from aspects of these conditions other than the use of the Internet for delivery in itself. Bernard et al. (2014) support this notion by explaining that meta-analysis may not be able “to deal with confounds among substantive moderator variables, since there is no latitude for control as there is in true experiments" (p.116). They add that even though it is difficult to control for these moderators, blended learning can be said to have a modest but significant effect, and that this positive effect supports the further investigation of blended learning as a possibly superior option to face-to-face learning. They also question whether all potential moderators should be controlled for, which would lead to risking results that are not replicable or applicable in practice. Instead, it might be better to "consider technology and instructional method as an inseparable dyad that are used together to achieve the goals of education and use experimental replication as a means of determining which combinations work best" (Bernard et al., 2014, p.116) - as was proposed before by Ross and Morrison (1989).

On the one hand, if the responsibility for the positive effect of blended learning lies mainly in moderators, this would support Clark's (1994) position that blended approaches merely use a different medium as a vehicle for delivery, and that there is no relationship between media and learning. On the other hand, it could be interpreted that blended learning does show a relationship between media and learning, because of the interaction between its capabilities and learning. This would be in support of Kozma's (1994) position. At the least, blended learning interventions do not seem to do any harm, when implemented correctly. Therefore the choice to employ them may not lay in their effectiveness, but instead in their efficiency. Blending instruction can offer a more cost-effective and less time-consuming approach, enabling students to study when and where they prefer, while maintaining the social interaction and accountability of face-to-face instruction. The presence of moderators implies that factors such as motivation of the students, accessible technology use for both teachers and students, and incentive to follow the intervention correctly, need to be considered by educators before implementing blended learning environments in their courses. It is also important to consider whether such an intervention is relevant in the context: Is there, for example, a significant part of the course that students would benefit from doing in their own time, such as practice and homework?

In exploring both benefits and challenges in blended education, it is often pointed out that the model provides students with more flexibility, by facilitating self-paced learning and offering an individual path of learning (Vaughan, 2007; Castro, 2019), but this means that they initially need to adapt to time management, taking greater responsibility for their own learning, and using sophisticated technologies (Vaughan, 2007). From an organisational perspective, reducing operating costs is a main benefit, with which comes the challenge of aligning the interventions with institutional goals and priorities (Vaughan, 2007). First, by providing access to more students and facilitating self-paced online learning activities. Second, by offering an individual path of learning for each student, thus improving out-of-class activities and feedback. More recently, Medina (2018) pointed out different implications of similar interventions from different perspectives. From the student perspective, benefits are flexibility and personalisation. In addition, it is important to incentivise attendance and participation, and to properly employ assessment strategies, with the goal of engaging the less intrinsically motivated students. From the teacher perspective, the same flexibility is a benefit, as lecture time can be reduced. Teachers should take this flexibility as an opportunity to accommodate learner differences. Furthermore, teachers should ensure that blended learning systems are implemented correctly by appropriately integrating online and face-to-face education and deciding what can be put online and what not. Internal expertise and support can be valuable in implementing this. This is in agreement with the suggestion that implementing blended learning while controlling for moderators is more difficult in larger courses. In larger courses, it might be more challenging to make sure that blended learning is implemented correctly, because either larger classes or more sections need to be taught, and the organisation might be more complex. Especially in this case, internal expertise and support are important. Some of the courses in this research synthesis might have employed these, whereas some might have not. Finally, from an institutional perspective, blended learning could reduce costs and enhance an institution's reputation. Institutions need to plan and ensure academic quality in blended courses (Medina, 2018; Serrano, Dea-Ayuela, Gonzalez-Burgos, Serrano-Gil & Lalatsa, 2019).

For larger courses, this study does not show whether implementing blended learning is useful or not. However, it could be derived from the results that moderators might be especially important in large-scale courses, both in general undergraduate education and statistics education. When blended approaches are implemented, it is important to look at which conditions decrease the effects of blended learning, and which conditions increase the effects. Like this, blended learning can be implemented in a way that makes the most of its benefits. Taking together the results found in this study and those from previous research, the conclusion is that blended learning does not offer a convenient solution to increasing student achievement in undergraduate education in and of itself. However, it may offer a small positive effect in terms of its efficiency and other capabilities. This gives reason to implement it in undergraduate courses, albeit with careful consideration of context and contributing factors.

4.4. Guidelines for Future Research

For future research, it is important to classify the exact blended learning intervention of which the effects need to be synthesized, and to investigate the optimal conditions to maximise the benefits of blended learning. Moderators to be investigated are, for example, the type of learning environment, the ratio of online to face-to-face learning, or the technological literacy of the users. As seen in the results from this study, course and class size might also be a contributing factor, which in turn might relate to the motivation of students and accountability for homework. It is also important for individual studies to document and report these characteristics accurately, so that they can be used in research syntheses. Moreover, the search query needs to be carefully designed, and in addition, the criteria need to be specific. It is useful to employ a taxonomy such as the one shown in Figure 1, but even within this taxonomy, the interventions should be more precisely specified. The subject investigated should not be blended learning in general, but more precisely determined terminology, e.g. "learning environments with automated feedback". It would also be valuable to look at the qualitative data in terms of these interventions, as these could further explain and support the quantitative results and explain the presence of moderators or what attributes of blended learning approaches are conducive to learning. Following these guidelines, the research would be better informed, and studies included in the synthesis would be more comparable, which would lead to more valuable and relevant evidence for practice.

References

American Psychological Association. (1967). *PsycINFO* [database].

Bai, Y., & Chang, T. S. (2016). Effects of class size and attendance policy on university classroom interaction in Taiwan. *Innovations in Education and Teaching International*, *53*(3), 316-328. https://doi-org.proxy.library.uu.nl/10.1080/14703297.2014.997776

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. doi: 10.1007/s12528-013-9077-3

Bhowmik, J., Meyer, D., & Phillips, B. (2016). Blended Learning in Postgraduate Applied Statistics Programs. Proceedings: OZCOTS 2016, Canberra, Australia. Retrieved from: https://iase-web.org/Publications.php?p=Regional_Publications

Biggs, J. & Tang, C. (2011). Teaching for quality learning at university (4 th ed.). Maidenhead, UK: Open University Press.

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

Castro, R. (2019). Blended learning in higher education: Trends and capabilities. *Education and Information Technologies*, *24*(4), 2523-2546. https://doi.org/10.1007/s10639-019-09886-3

Clark, R. E. (1994). Media will never influence learning. *Educational technology research and development*, *42*(2), 21-29. Retrieved from: http://www.jstor.com/stable/30218684

De Paola, M., Ponzo, M., & Scoppa, V. (2013). Class size effects on student achievement: heterogeneity across abilities and fields. *Education Economics*, *21*(2), 135-153. https://doi-org.proxy.library.uu.nl/10.1080/09645292.2010.511811

Education Resources Information Center. (1966). *ERIC* [database].

Gannaway, D., Green, T., & Mertova, P. (2018) So how big is big? Investigating the impact of class size on ratings in student evaluation. *Assessment & Evaluation in Higher Education, 4*(2), 175-184. doi:10.1080/02602938.2017.1317327

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

Güzer, B., & Caner, H. (2014). The past, present and future of blended learning: an in depth analysis of literature. *Procedia-social and behavioral sciences*, *116*, 4596-4603. doi:10.1016/j.sbspro.2014.01.992

Hak, T., Van Rhee, H. J., & Suurmond, R. (2016). How to interpret results of meta-analysis. (Version 1.3) Rotterdam, The Netherlands: Erasmus Rotterdam Institute of Management. Retrieved from: www.erim.eur.nl/researchsupport/meta-essentials/downloads

Higgins, J. P. T., Green, S. (editors). *Cochrane Handbook for Systematic Reviews of Interventions.* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

Jepsen, C. (2015). Class size: does it matter for student achievement?. *IZA World of Labor*.

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. *Educational technology research and development, 42*(2), 7-19. Retrieved from: http://www.jstor.com/stable/30218683

Margulieux, L. E., McCracken, W. M., & Catrambone, R. (2016). A taxonomy to define courses that mix face-to-face and online learning. *Educational Research Review, 19*, 104-118. http://dx.doi.org/10.1016/j.edurev.2016.07.001

McCutcheon, K., Lohan, M., Traynor, M., & Martin, D. (2015). A systematic review evaluating the impact of online or blended learning vs. face-to-face learning of clinical skills in undergraduate nurse education. *Journal of advanced nursing, 71*(2), 255-270. https://doi-org.proxy.library.uu.nl/10.1111/jan.12509

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.

Medina, L. C. (2018). Blended learning: Deficits and prospects in higher education. *Australasian Journal of Educational Technology*, *34*(1). doi:10.14742/ajet.3100

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Monks, J., & Schmidt, R. M. (2011). The impact of class size on outcomes in higher education. *The BE Journal of Economic Analysis & Policy, 11*(1). doi:10.2202/1935-1682.2803

Oliver, M., & Trigwell, K. (2005). Can ‘blended learning’be redeemed? *E-learning and Digital Media*, *2*(1), 17-26. Retrieved from: https://journals.sagepub.com/doi/pdf/10.2304/elea.2005.2.1.17

Osguthorpe, R. T., & Graham, C. R. (2003). Blended learning environments: Definitions and directions. *Quarterly review of distance education*, *4*(3), 227-33.

Picciano, A. G. (2014). A critical reflection of the current research in online and blended learning. *European Lifelong learning magazine*. Retrieved from: http://www.elmmagazine.eu/articles/a-critical-reflection-of-the-current-research-in-online-a nd-blended-learning

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL http://www.R-project.org/.

Ross, S. M., & Morrison, G. R. (1989). In search of a happy medium in instructional technology research: Issues concerning external validity, media replications, and learner control. *Educational Technology Research and Development, 37*(1), 19–33. doi:10.1007/BF02299043.

Vaughan, N. (2007). Perspectives on blended learning in higher education. *International Journal on E- learning*, *6*(1), 81-94.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. UsRL: http://www.jstatsoft.org/v36/i03/

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*, 17-28. http://dx.doi.org/10.1016/j.stueduc.2017.01.002

McGuinness, L. A. (2019). *robvis: An R package and web application for visualising risk-of-bias assessments.* Retrieved from: https://github.com/mcguinlu/robvis

OCEBM (2011.) The Oxford 2011 Levels of Evidence. *Oxford Centre for Evidence-Based Medicine.* Retrieved from: http://www.cebm.net/index.aspx?o=5653

Serrano, D. R., Dea-Ayuela, M. A., Gonzalez-Burgos, E., Serrano-Gil, A., & Lalatsa, A. (2019). Technology-enhanced learning in higher education: How to enhance student engagement through blended learning. *European Journal of Education*, *54*(2), 273-286. doi:10.1111/ejed.12330

Sterne, J.A.C., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I., Carpenter, J.R., Chan, A.W., Churchill, R., Deeks, J.J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F., Higgins, J.P.T. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *The BMJ, 355,* [i4919]. doi: 10.1136/bmj.i4919.

Sterne, J. A. C., Savovi, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H. Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., Reeves, B. C., Shepperd, S., Shrier, I., Stewart, L. A., Tilling, K., White, I. R., Whiting, P. F., & Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *The BMJ, 366,* [l4898]. https://doi.org/10.1136/bmj.l4898

Tishkoveskaya, S., & Lancaster, G.A. (2012). Statistical Education in the 21st Century: a Review of Challenges, Teaching Innovations and Strategies for Reform. *Journal of Statistics Education, 20*(2), 1-56. doi:10.1080/10691898.2012.11889641

Zwetsloot, P. P., Van Der Naald, M., Sena, E. S., Howells, D. W., IntHout, J., De Groot, J. A., Chamuleau, S. A. J., MacLeod, M. R., & Wever, K. E. (2017). Standardized mean differences cause funnel plot distortion in publication bias assessments. *Elife, 6*, e24260. https://doi.org/10.7554/eLife.24260.001

Appendix A
Search Query and Logic Grid

The search query used is the following:

(undergraduate education/ or undergraduate study/ or undergraduate students/ or college freshmen/ or college students/ or (undergraduate).ti,ab,id.) AND (blended learning/or electronic learning/ or technology integration/ or web based instruction/ or virtual classrooms/ or (blended learning or electronic learning or technology integration or web based instruction or virtual classrooms or flipped classroom or online learning environment).ti,ab,id.)

This query was set up according to the logic grid shown in Table A1.

Table A1.

Logic grid of the terms used in the search query.

	Used terms	
	Students in undergraduate courses	Blended learning
Subject headings	undergraduate education/ or undergraduate study/ or undergraduate students/ or college freshmen/ or college students/	blended learning/or electronic learning/ or technology integration/ or web based instruction/ or virtual classrooms/
Title, abstract, keywords	(undergraduate).ti,ab,id.	(blended learning or electronic learning or technology integration or web based instruction or virtual classrooms or flipped classroom or online learning environment).ti,ab,id.)

Appendix B
List of References Used in the Research Synthesis

Babaali, P., & Gonzalez, L. (2015). A quantitative analysis of the relationship between an online homework system and student achievement in pre-calculus. *International Journal of Mathematical Education in Science and Technology*, *46*(5), 687-699. http://dx.doi.org/10.1080/0020739X.2014.997318

Banditvilai, C. (2016). Enhancing Students' Language Skills through Blended Learning. *Electronic Journal of e-Learning*, *14*(3), 220-229. Retrieved from: https://files.eric.ed.gov/fulltext/EJ1107134.pdf

Bortnik, B., Stozhko, N., Pervukhina, I., Tchernysheva, A., & Belysheva, G. (2017). Effect of virtual analytical chemistry laboratory on enhancing student research skills and practices. *Research in Learning Technology*, *25*. http://dx.doi.org/10.25304/rlt.v25.1968

Botts, R. T., Carter, L., & Crockett, C. (2018). Using the blended learning approach in a quantitative literacy course. *PRIMUS*, *28*(3), 236-265. doi:10.1080/10511970.2017.1371264

Callahan, J. T. (2016). Assessing online homework in first-semester calculus. *PRIMUS*, *26*(6), 545-556. doi:10.1080/10511970.2015.1128501

Dry, M. J., Due, C., Powell, C., Chur-Hansen, A., & Burns, N. R. (2018). Assessing the Utility of an Online Adaptive Learning Tool in a Large Undergraduate Psychology Course. *Psychology Teaching Review*, *24*(2), 24-37. Retrieved from: https://files.eric.ed.gov/fulltext/EJ1196465.pdf

Förster, M., Weiser, C., & Maur, A. (2018). How feedback provided by voluntary electronic quizzes affects learning outcomes of university students in large classes. *Computers & Education*, *121*, 100-114. https://doi.org/10.1016/j.compedu.2018.02.012

Goette, W. F., Delello, J. A., Schmitt, A. L., Sullivan, J. R., & Rangel, A. (2017). Comparing delivery approaches to teaching abnormal psychology: investigating student perceptions and learning outcomes. *Psychology Learning & Teaching*, *16*(3), 336-352. https://doi-org.proxy.library.uu.nl/10.1177%2F1475725717716624

Jonsdottir, A. H., Bjornsdottir, A., & Stefansson, G. (2017). Difference in learning among students doing pen-and-paper homework compared to web-based homework in an introductory statistics course. *Journal of Statistics Education*, *25*(1), 12-20. doi:10.1080/10691898.2017.1291289

Molnar, K. K. (2017). What effect does flipping the classroom have on undergraduate student perceptions and grades?. *Education and Information Technologies*, *22*(6), 2741-2765. doi:10.1007/s10639-016-9568-8

Potter, J. (2015). Applying a Hybrid Model: Can It Enhance Student Learning Outcomes?. *Journal of Instructional Pedagogies*, *17*. Retrieved from: https://files.eric.ed.gov/fulltext/EJ1102855.pdf

Powers, K. L., Brooks, P. J., Galazyn, M., & Donnelly, S. (2016). Testing the efficacy of MyPsychLab to replace traditional instruction in a hybrid course. *Psychology Learning & Teaching*, *15*(1), 6-30. https://doi-org.proxy.library.uu.nl/10.1177%2F1475725716636514

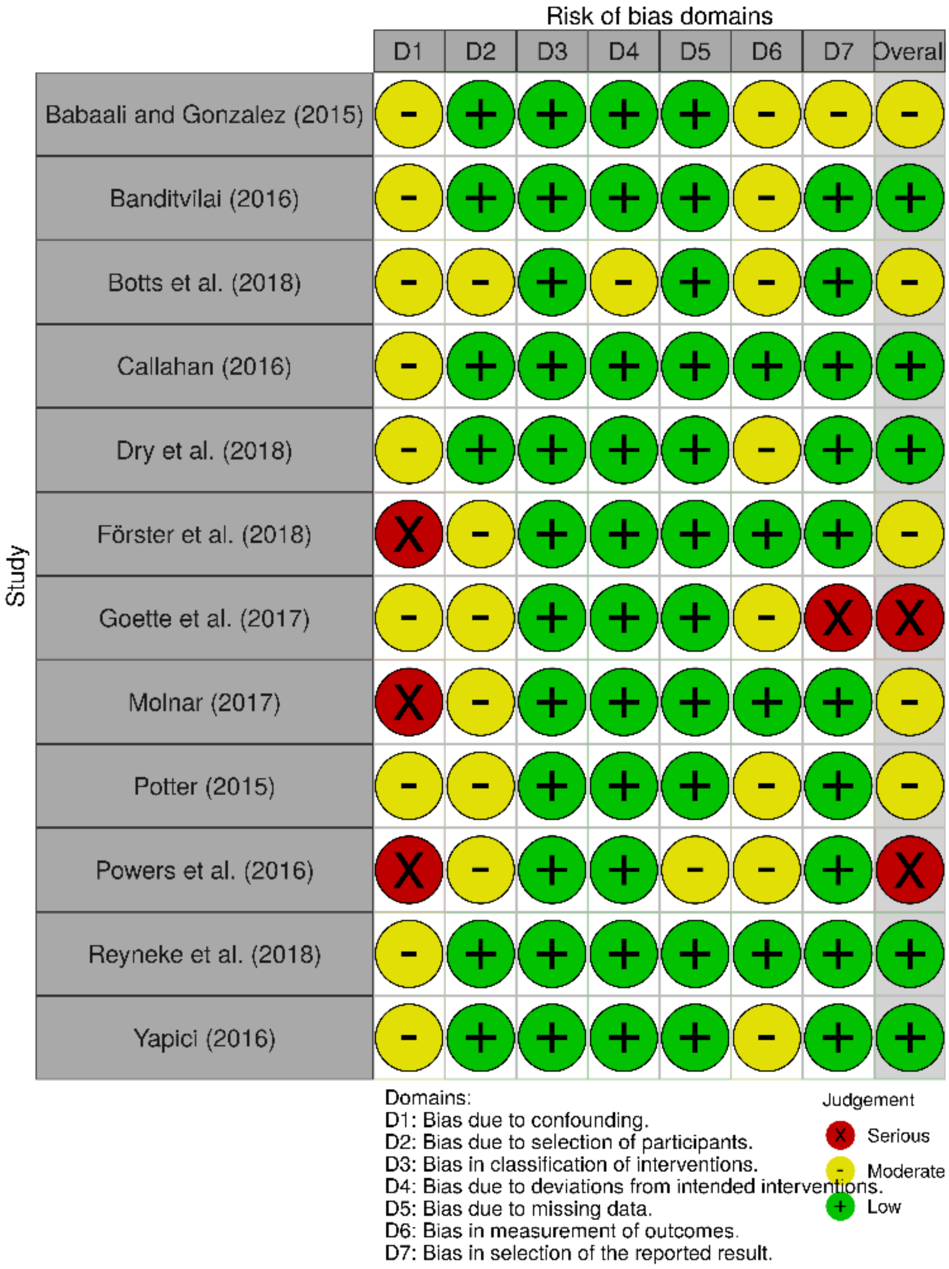
Reyneke, F., Fletcher, L., & Harding, A. (2018). The effect of technology-based interventions on the performance of first year university statistics students. *African Journal of Research in Mathematics, Science and Technology Education*, *22*(2), 231-242. doi:10.1080/18117295.2018.1477557

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. http://dx.doi.org/10.1016/j.compedu.2017.01.003

Yapici, İ. Ü. (2016). Effectiveness of Blended Cooperative Learning Environment in Biology Teaching: Classroom Community Sense, Academic Achievement and Satisfaction. *Journal of Education and Training Studies*, *4*(4), 269-280. doi:10.11114/jets.v4i4.1372

Appendix C
Risk of Bias Assessment

The visualisations of the risk of bias assessments are shown below. Figure C1 presents the assessments for the non-randomised studies. Figure C2 presents the assessments for the randomised studies.

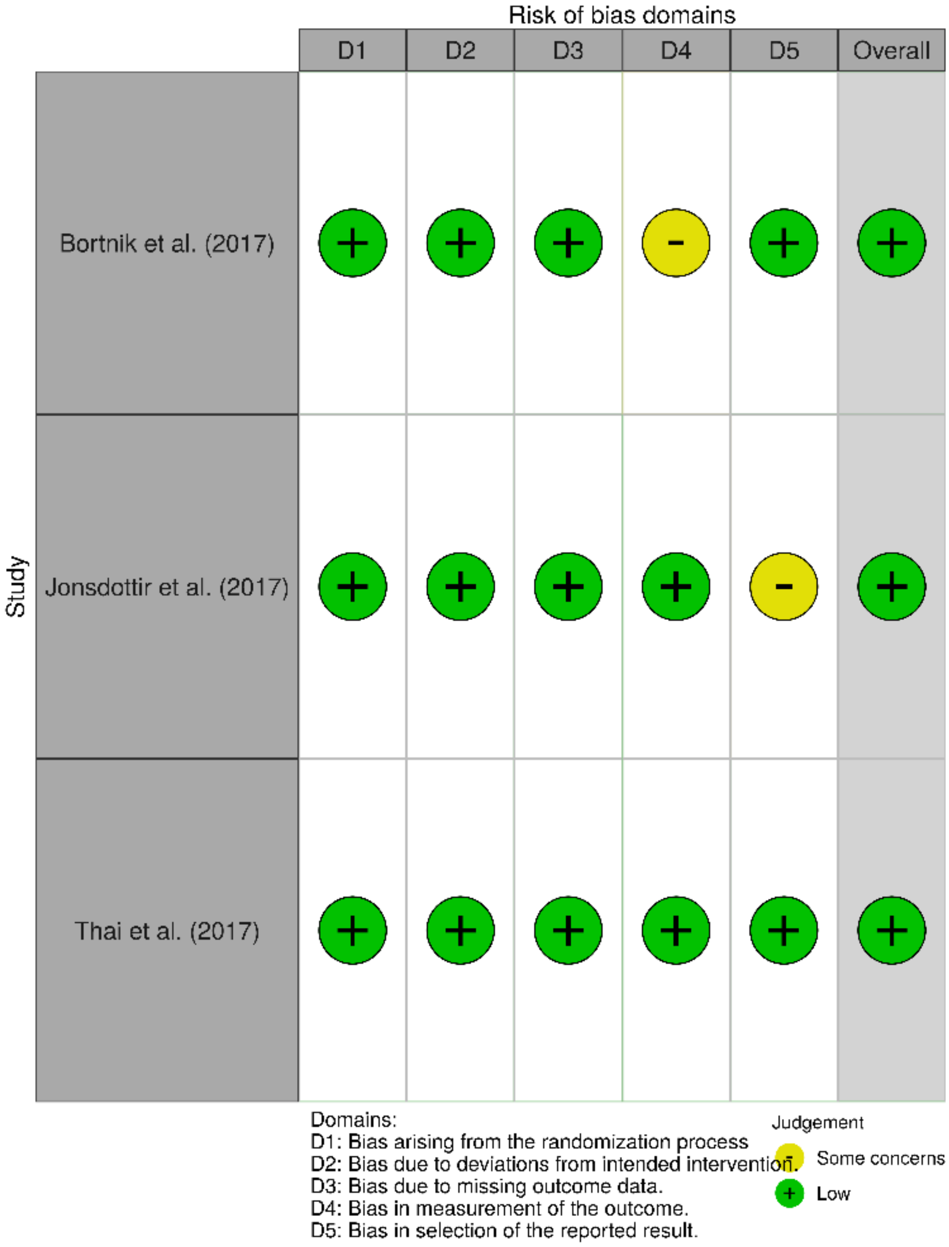


	Risk of bias domains							
Study	D1	D2	D3	D4	D5	D6	D7	Overal
Babaali and Gonzalez (2015)	-	+	+	+	+	-	-	-
Banditvilai (2016)	-	+	+	+	+	-	+	+
Botts et al. (2018)	-	-	+	-	+	-	+	-
Callahan (2016)	-	+	+	+	+	+	+	+
Dry et al. (2018)	-	+	+	+	+	-	+	+
Förster et al. (2018)	X	-	+	+	+	+	+	-
Goette et al. (2017)	-	-	+	+	+	-	X	X
Molnar (2017)	X	-	+	+	+	+	+	-
Potter (2015)	-	-	+	+	+	-	+	-
Powers et al. (2016)	X	-	+	+	-	-	+	X
Reyneke et al. (2018)	-	+	+	+	+	+	+	+
Yapici (2016)	-	+	+	+	+	-	+	+

Domains:
D1: Bias due to confounding.
D2: Bias due to selection of participants.
D3: Bias in classification of interventions.
D4: Bias due to deviations from intended interventions.
D5: Bias due to missing data.
D6: Bias in measurement of outcomes.
D7: Bias in selection of the reported result.

Judgement
X Serious
- Moderate
+ Low

Figure C1. Risk of bias assessments for the non-randomised studies, using the ROBINS-I tool.



Risk of bias domains

Study	D1	D2	D3	D4	D5	Overall
Bortnik et al. (2017)	+	+	+	-	+	+
Jonsdottir et al. (2017)	+	+	+	+	-	+
Thai et al. (2017)	+	+	+	+	+	+

Domains:
D1: Bias arising from the randomization process
D2: Bias due to deviations from intended intervention.
D3: Bias due to missing outcome data.
D4: Bias in measurement of the outcome.
D5: Bias in selection of the reported result.

Judgement
- Some concerns
+ Low

Figure C2. Risk of bias assessments for the randomised studies, using the RoB 2.0 tool.

Appendix D
Influence Diagnostics

In order to inspect outliers, the influence diagnostics available in the Metafor package (Viechtbauer, 2010) were plotted (Figure D1). The studies 7 Förster et al. (2018), and 12 (Powers et al., 2016) show slightly more deviation on some of the diagnostics.

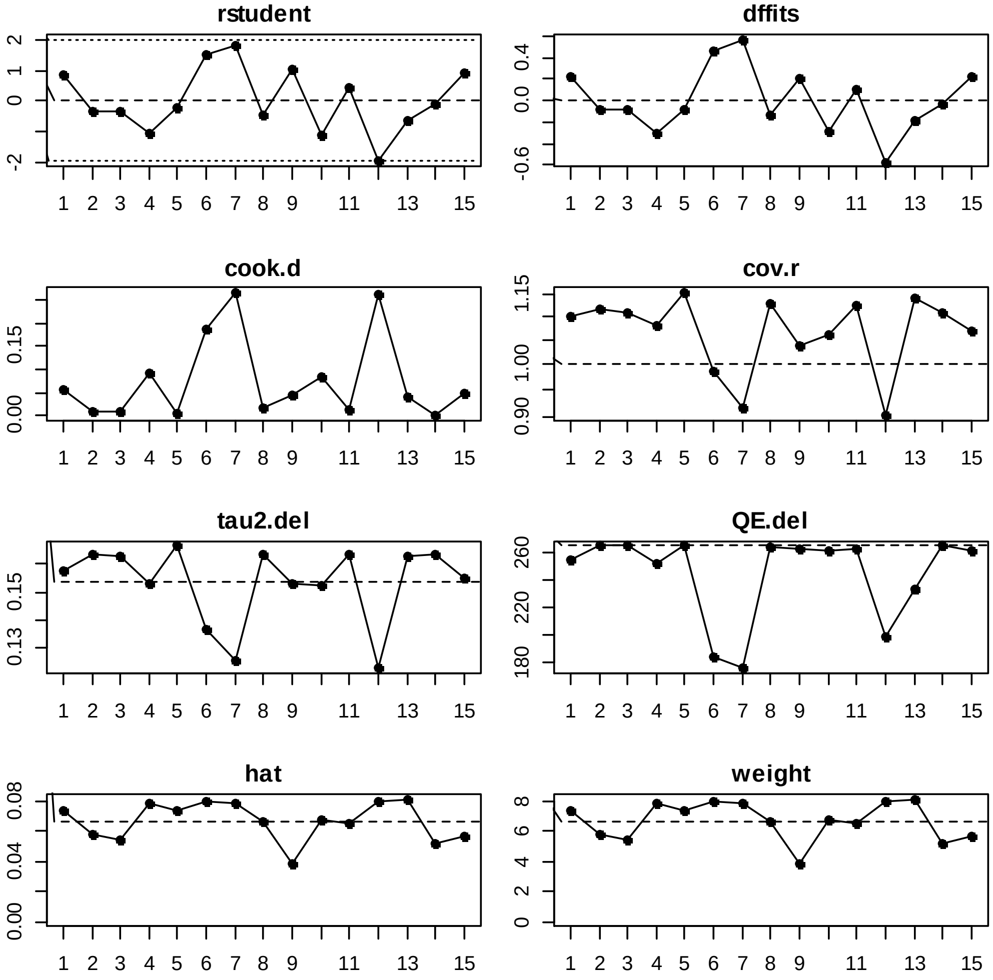


Figure D1. Influence diagnostics plotted for all studies included in the meta-analysis.

Appendix E
Sensitivity Analysis

In the sensitivity analysis, Powers et al. (2016), Förster et al. (2018), Goette et al. (2017), and Molnar (2017) were excluded. The forest plot is shown in Figure E1. The results show a significant summary effect, with *d* = 0.39 [0.18; 0.61], and *p* < .001. However, the results also show a significant amount of heterogeneity, *Q*(*df* = 10) = 120.60, *p* < .0001 and I^2 = 87.66%. The funnel plot is shown in Figure E2.

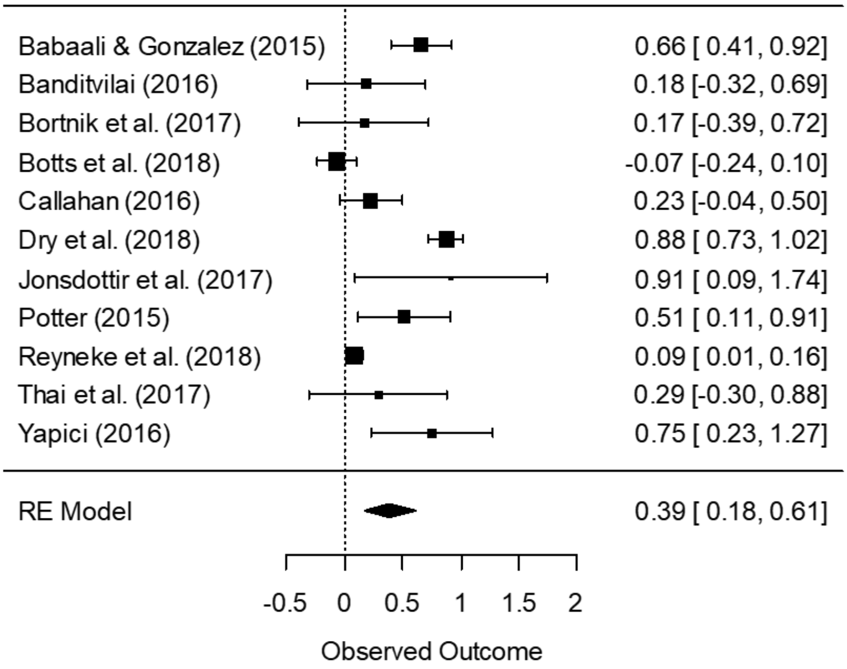


Figure E1. Forest plot of the estimated effect sizes in the sensitivity analysis.

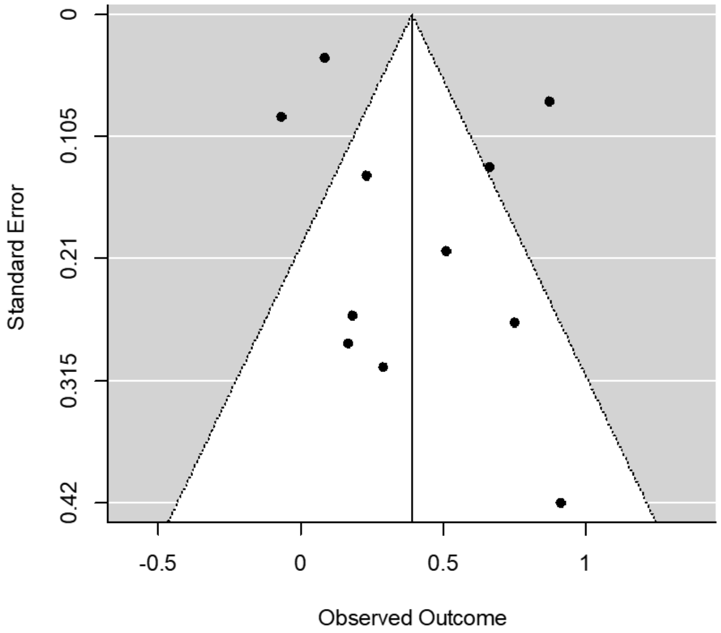


Figure E2. Funnel plot for the sensitivity analysis.