Flipped Classrooms versus Traditional Classrooms:

A systematic review and meta-analysis of student achievement in higher education

Carol Nancy Sparkes

A Thesis

In the Department

of

Education

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Educational Technology) at

Concordia University

Montreal, Quebec, Canada

February 2019

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Carol Nancy Sparkes

Entitled: Flipped Classrooms Versus Traditional Classrooms: A Systematic Review and Meta-Anaylsis of Student Achievement in Higher Education

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy (Educational Technology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Helena Osana

_____ External Examiner
Dr. Ron Owston

_____ External to Program
Dr. Norman Segalowitz

_____ Examiner
Dr. Richard Schmid

_____ Examiner
Dr. David Waddington

_____ Thesis Supervisor
Dr. Robert Bernard

Approved by

Dr. Steven Shaw, Graduate Program Director

April 10, 2019

_____
Dr. Andre Roy, Dean
Faculty of Arts and Science

ABSTRACT

Flipped Classrooms versus Traditional Classrooms:

A systematic review and meta-analysis on student achievement in higher education

Carol Nancy Sparkes, PhD

Concordia University, 2019

In an attempt to understand what makes blended learning (BL) more effective than Classroom Instruction (CI), this research looked more closely at the Flipped Classroom (FC) model of BL. The FC takes a relatively consistent approach to course design by flipping what is traditionally done in the classroom (i.e., lecture) with what is traditionally done as homework (i.e., application).

Numerous studies have been conducted comparing FC with the CI on student achievement in higher education without conclusive results. To synthesize the literature, this dissertation implemented a systematic review and meta-analysis to measure the average effect size and the direction of the impact and to determine the conditions under which students learn more effectively. To ensure a transparent process the potential for bias in each step of a meta-analysis was acknowledged and addressed.

Through a systematic review of the literature from 2000 to 2017, 114 studies were included and 125 effect sizes were calculated. Using meta-analysis these effect sizes created a weighted mean effect-size of +0.30, which was statistically significant at $p < 0.05$ and educationally significant.

Study features were analyzed to determine if there were any attributes that made a difference but none were found to be significant. The use of quizzes, however, showed an interesting pattern and near significant difference ($p = .058$) when the effect sizes were grouped by STEM, non-STEM and Health-related disciplines. No publication bias was found, no outliers were found from the sensitivity analysis, and there was no significant difference between the effects from quasi-experimental and experimental designs.

While the FC significantly outperformed CI it was not to a greater extent than general BL outperformed CI. Future research is encouraged between levels of treatments, instead of between FC and CI, in order to provide more nuanced results about how to improve instructional design in future courses.

ACKNOWLEDGEMENTS

I would like to acknowledge those who not only made this research possible but who nurtured my intellectual curiosity along the way.

To my supervisor Dr. Robert Bernard, thank you for your encouraging, wise, and experienced counsel that guided me through this incredible learning journey. It was an honour to be your student.

To Dr. Richard Schmid thank you for your feedback when I needed it most. Like concepts maps, feedback may be uncomfortable at the time, but important to learning.

To Dr. David Waddington thank you for introducing me to the many educational philosophers upon whose shoulders we stand. Your teachings enriched my learning deeply.

To Dr. Evgueni Borokhovski thank you for sharing your wealth of experience and knowledge with meta-analysis. Your patience is admirable.

To David Pickup thank you for searching the literature and then helping to those code studies. Like Sisyphus, in Greek mythology, your work seemed never ending, yet your positive attitude never wavered.

Thank you to Nadine Wright for your smile while helping me to navigate the many regulations and deadlines. Where would we be without you?

From a personal perspective, I would like to thank my loving parents, Mary and Tom Sparkes, who encouraged my love of learning and rewarded my achievements. From you, I learned about tenacity, and the many transferable skills needed in nursing, farming, and life in general.

Thank you to my sisters, Claire and Ruth, my earliest peer-reviewers, who taught me to deal with criticism, a valuable skill indeed.

Thank you to my Aunt Anne Coddington, PhD, whose own academic pursuits encouraged my own. Your love and friendship have been wonderful sources of comfort. Your bookshelf was a never-ending source of inspiration.

To my husband Don, thank you for your enduring love and support.

Thank you to the many people who encouraged me along the way, and thank you even to the few who doubted me, as both helped me to dig deeper and lean in.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW 1

What is the Flipped Classroom? 3

Active Learning's Central Role in the Flipped Classroom 7

The Effectiveness of DE, OL, BL and the Flipped Classroom: An Examination of Meta-Analyses 13

The Methodology of Systematic Review and Meta-Analysis 20

The Purpose of this Study 25

CHAPTER 2: METHOD 28

Research Questions, Terms and Definitions, and Inclusion/Exclusion Criteria 29

Literature Search Strategies and Search Outcomes 32

Selecting Studies, Extracting Effect Sizes, and Coding Study Features 33

Statistical methods 37

Conclusion 39

CHAPTER 3: RESULTS 40

Overview of Included Studies and Average Effect Size 40

Publication Bias 44

Test of Moderator Variables 49

CHAPTER 4: DISCUSSION 56

Major Findings and Interpretation 56

Generalizability of the Conclusions 59

Implications for Theory and Practice 61

Limitations 63

Suggestions for Future Research 65

Overall Summary 67

REFERENCES 68

APPENDIX A. CATEGORIES, NUMBERS, AND % OF EXCLUDED FULL-TEXT STUDIES 95

APPENDIX B. CODEBOOK 97

APPENDIX C: FOREST PLOT FOR THE FULL SET OF 125 EFFECT SIZES 100

LIST OF TABLES

1. Summary of meta-analyses conducted comparing DE, OL, BL, and FC to CI (adapted from Bernard 2017, OLC presentation) 15
2. Overall results 41
3. Sensitivity analysis 43
4. Duval and Tweedie's trim and fill (zero studies trimmed) 46
5. Other forms of potential bias 48
6. Demographic variables 50
7. Educationally relevant moderator variables 54
8. Meta-analyses comparing Blended Learning/Flipped Classroom versus Classroom Instruction 57

LIST OF FIGURES

Figure 1. Connection between Classroom Instruction and FC to Bloom's Taxonomy (adapted from Lopes & Soares, 2018, p. 3847). 4

Figure 2. Evolution of the Flipped Classroom (FC) from Classroom Instruction (CI)....6

Figure 3. A Venn diagram of active learning (Bishop and Verleger, 2013, p. 6) 8

Figure 4. Blended-learning taxonomy (Staker and Horn, 2012, p. 2)....11

Figure 5. Flipped Classroom concepts in relation to the research. 27

Figure 6. PRISMA flow diagram 33

Figure 7. Funnel plot with effect sizes (horizontal axis) and standard errors (vertical axis) for the 125 effect sizes (hollow dots)....46

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

As a new instructional approach is developed, researched and introduced into educational practice, its ability to improve student learning is often compared with that of the traditional lecture approach (e.g., Bernard, Abrami, Lou, Borokhovski, Wade, Wozney, Wallet, Fiset & Huang, 2004; Means, Toyama, Murphy & Bakia, 2013). The traditional lecture has been traced back to the 13th century, when books were so rare that the professor would read to the students from the one copy available, yet the lecture continues still as the main form of educational delivery in post-secondary education at the undergraduate level, even with the abundance of information on the internet (Bates, 2015), ease of access to it, and higher literacy for the general population to read it. Although academic achievement is only one measure of success of an instructional approach, an alternative to lecturing would be difficult to be recommended unless it was also at least as effective in student learning as the lecture.

One of the more recent instructional approaches that has been inserted into this quest to improve student learning is the Flipped Classroom (FC). The FC is used in both K-12 and postsecondary courses, but research is more abundant in the latter. The focus of this study is on postsecondary, partly because this is where the FC began (Lage, Pratt & Treglia, 2000) but also because of major differences between how the FC is implemented with students in postsecondary classrooms and K-12 classrooms (Staker & Horn, 2012).

The FC can be traced back to 1996 known then as the *inverted classroom,* when Maureen Lage and two fellow economics professors at Miami University in the US were trying to provide alternatives to lectures and a more inclusive environment that appealed to "all types of learners" (Lage et al., 2000, p. 32). Lage et al. recorded lectures on VHS videotapes, and PowerPoint files with audio for students with a computer to watch at home or in a lab. Class time was reserved for

students to participate in activities that gave them the opportunity to see the economic principles in action. An example of such an activity is auctioning off a can of cola in class and charting the resulting supply and demand curves. Lage et al. found that the inverted classroom approach was more effective for female students' achievement. As the field of economics was male dominated at that time, and when inclusivity was a goal, this was considered a significant result.

In 2007, a similar approach was used by Bergmann and Sams, two Colorado (US) high school chemistry teachers, to accommodate students who needed to miss classes to participate in school sporting events. As Internet bandwidth and computer access had improved by 2007, the video lectures were provided through the Internet as downloadable podcasts (audio) or vodcasts (video). Students were asked to listen to the lecture before class, so that class time was "reserved exclusively for lab activities, demonstrations, one-to-one assistance, and small group tutoring" (Bergmann & Sams, 2008, p. 22). Bergmann and Sams referred to this instructional approach as the flipped classroom. Students continued at their own pace achieving mastery at each stage before moving to the next level of the course material, harking back to the mastery learning literature (e.g., Bloom, 1968; Guskey, 2007) that generated considerable interest and debate from the 1970s to the end of the 1980s (Guskey, 1987). Mastery instruction, however, is not considered a requirement in more recent FC literature.

Bergmann and Sams posted their instructional videos online for anyone to use, which likely helped the idea of the flipped classroom approach spread quickly. In 2012, they shared their experiences of flipping their classroom in a book, entitled *Flip Your Classroom: Reach Every Student in Every Class Every Day* (Bergmann & Sams, 2012). The title promised what teachers were already striving to do, that is to reach every student in every class, every day. Since then, more primary studies comparing the effectiveness of the flipped classroom with the traditional lecture-based classroom instruction (CI) have been conducted and appeared in the

literature each year. The corpus of these studies has now reached a point that a comprehensive meta-analysis of the empirical literature is warranted.

This dissertation uses the methodology of meta-analysis to summarize the literature from 2000 to 2017 in an attempt to determine if FCs live up to their hype and to search for common instructional features that might moderate the overall effect. In the following sections, there is more about what the FC is, active learning's central role in the FC, FC as a form of BL, the effectiveness of DE, OL, BL and the FC by examining various meta-analyses, and the methodology of systematic review and meta-analysis.

What is the Flipped Classroom?

The FC is a form of blended learning (BL), meaning that part of a course is conducted online and part in the classroom, that flips or reverses what is traditionally done in the classroom (i.e., lecture instruction) with what is traditionally done as homework (i.e., active application of theory to problems). Figure 1 shows the Classroom Instruction Model on the left of Anderson and Krathwohl's (2001) modification of Bloom's Taxonomy Framework (Bloom, 1956; Bloom 1968) and the Flipped Classroom Model on the right indicating its reversed/flipped connection to the levels. In the Classroom Instruction Model the bottom three levels of Bloom's taxonomy (i.e., remembering, understanding and applying) are addressed in the classroom, while in the Flipped Classroom Model they are "flipped" and addressed at home through the student watching video lectures and completing worksheets or quizzes. In the Classroom Instruction Model the top three levels of Bloom's Taxonomy (i.e., analyzing, evaluating and creating) are addressed at home alone, while in the Flipped Classroom Model they are "flipped" and addressed in the classroom with the additional support of other students and the instructor(s) (Lopes & Soares, 2018).

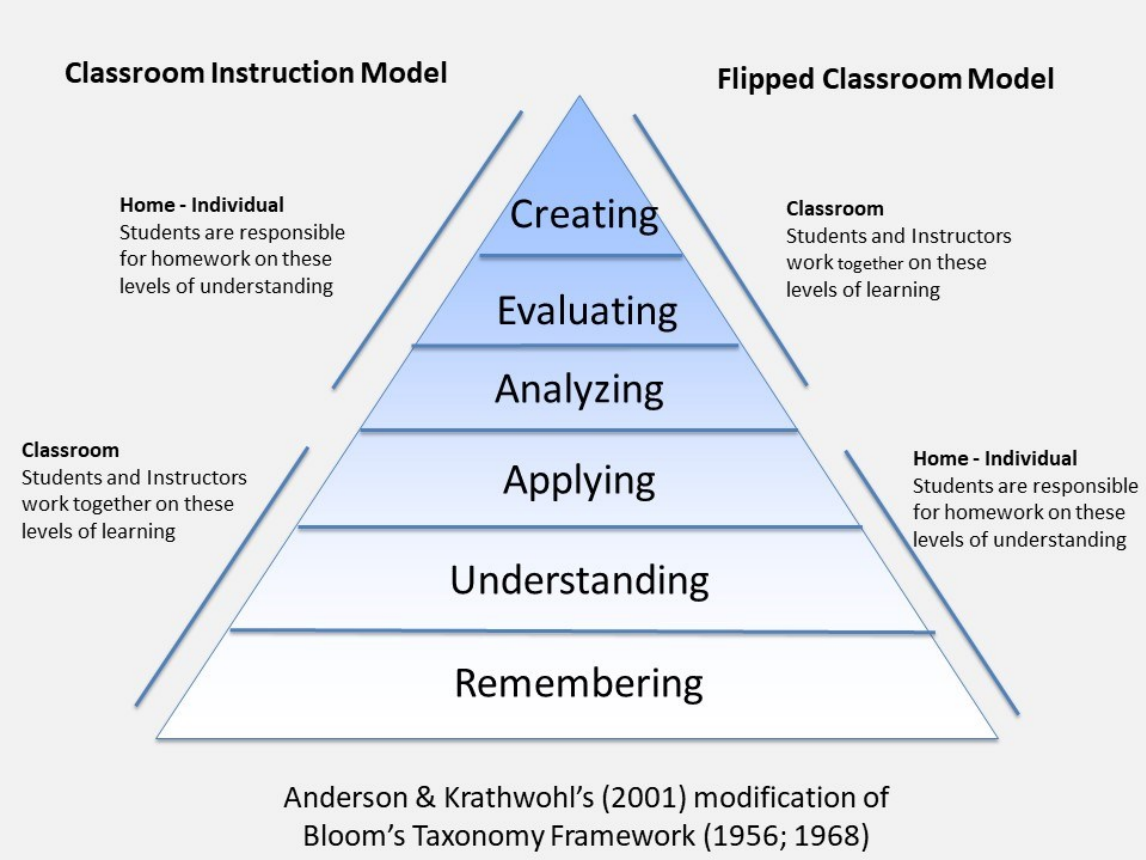


Figure 1. Connection between Classroom Instruction and FC to Bloom's Taxonomy (adapted from Lopes & Soares, 2018, p. 3847).

The FC is also known as the *inverted classroom* and *reversed instruction* (Lage et al., 2000; Bergmann & Sams, 2012; Ruddick, 2012; Baepler, Walker & Driessen, 2014). Bishop and Verleger (2013) described the FC as "an educational technique that consists of two parts: interactive group learning activities inside the classroom, and direct computer-based individual instruction outside the classroom" (p. 5). *Direct computer-based individual instruction* in this case refers to lectures that are recorded as video for the students to watch at home as their first introduction to the material and to prepare them with the pre-requisite knowledge needed to participate in active learning in the classroom. This *direct computer-based individual instruction* referred to by Bishop and Verleger is not to be confused with the *Direct Instruction (referred to as DI)* of Zigfred Engelman who developed a scripted model of instruction to teach at-risk

children in the 1960s, on which the National Institute for Direct Instruction was founded (https://www.nifdi.org/). A meta-analysis was recently published on DI (Stockard, Wood, Coughlin, & Rasplica Khoury, 2018) reporting its effectiveness. Another form of *direct instruction* (referred to as *small di*) was introduced by Rosenshine in 1976 in his teacher effectiveness research (e.g., daily review, presenting new material, guiding student practice, providing feedback and corrections, conducting independent practice, and weekly and monthly review may apply). Rosenshine laments that some authors refer to direct instruction as any instruction that is led by the teacher no matter how systematic or unsystematic it is (e.g., Kuhn, 2007; Rosenshine, 2009). Even though the flipped classroom video lecture does not necessarily follow Engelman's (large DI) or Rosenshine's (small di) systematic form of instruction, they are all a form of teacher led explicit instruction.

To encourage students to watch the video lectures of the FC carefully, pre-class or beginning of class quizzes based on the video lectures are commonly used in the FC. Sometimes worksheets for students to complete prior to class are used to ensure that students watch the videos and are prepared for the class activities (Lage et al., 2000). The results of these online assignments provide just-in-time feedback to inform the teacher of concepts that need clarification during the class session.

Figure 2 shows the evolution of the FC model from blended learning (i.e., face-to-face and online learning) but with a specific pedagogical approach to be used in the classroom (i.e., active learning) and online (i.e., direct instruction in form of video lectures). Blended learning evolved from face-to-face classroom instruction by including the online environment for learning as well. With the increased accessibility of the Internet and personal computers, OL (i.e., students working completely on a computer on the Internet) evolved from distance education (i.e., primarily paper, radio, television, VHS tape, DVD based).

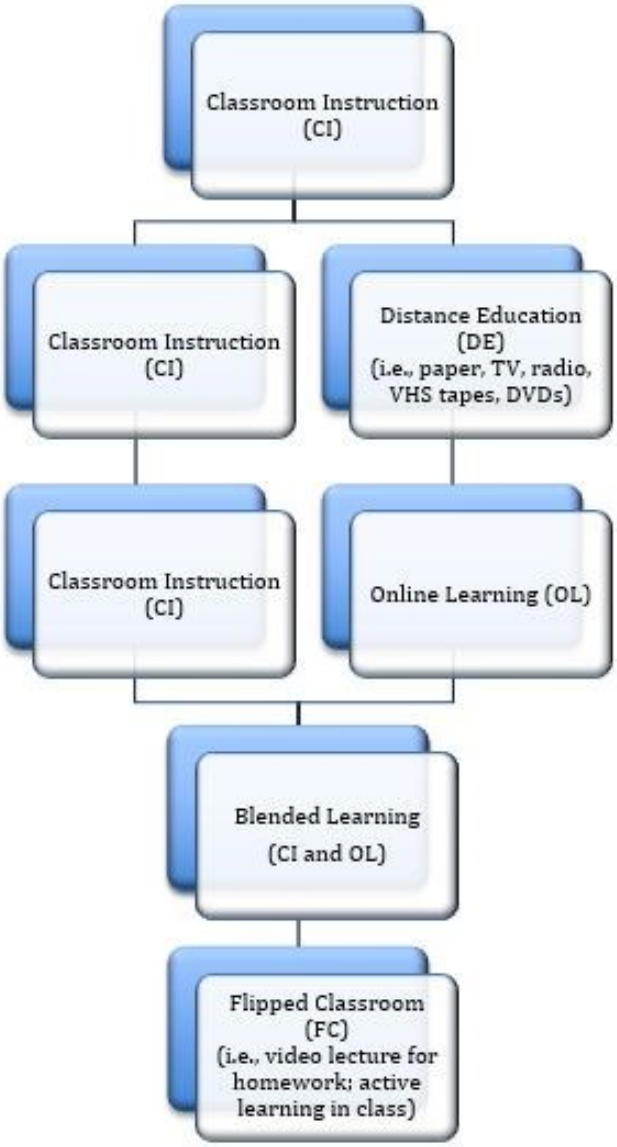


Figure 2. Evolution of the Flipped Classroom (FC) from Classroom Instruction (CI)

The FC comes with opportunities and challenges as noted in two qualitative reviews (Halili & Zainuddin, 2015; Karabulut-Ilgu, Jaramillo Cherrez & Jahren, 2018). Students valued the flexibility to watch and re-watch lecture videos at their own pace, and the access to the instructor for help with active learning and complex problem solving during class time. Instructors appreciated opportunities to interact with students to better understand their difficulties, and the ability to respond quickly with the necessary personalized corrective

feedback. Through this regular interaction the instructor has opportunities to get to know the students and their interests thus helping to customize more effective responses. Sometimes challenges arose, however, when students were reluctant to prepare sufficiently for class, did not participate fully during class, and complained of technical issues. Students did not always buy into the idea of taking responsibility for their own learning. Instructors felt overwhelmed sometimes by the greater workload; preparing the recorded lectures ahead of time and facilitating active learning classes instead of lecturing as they were used to doing. It must be noted, however, that these assessments were based on qualitative reviews.

Active Learning's Central Role in the Flipped Classroom

The idea of active learning has a long history including ideas from Dewey, Piaget, Vygotsky, and Bruner. John Dewey, an American philosopher and psychologist in the early 1900s, was a pragmatist who encouraged moving school activities from "drill, recitation, rote memorization, lecturing" to "broad-scale and open-ended group projects that involved activities such as carpentry, weaving, cooking, and candle making" (Dewey & Jackson, 1990, p. xxxiii). Jean Piaget, a Swiss philosopher who was known as a cognitive constructivist, argued that acquiring knowledge was "a process of continuous self-construction" (Driscoll, 2005, p. 191). Lev Vygotsky, known for social constructivism, and Jerome Bruner, known for constructivist theory, both agree that, "individual development could not be understood without reference to the social and cultural context within which such development is embedded" (Driscoll, 2005, p. 247). Each of these people remind us that learners are not just empty vessels waiting to be filled or blank slates waiting to be "written on" by the teacher's words yet "much of U.S. schooling has been based on this premise" (Wilson & Peterson, 2006, p. 2).

Active learning is central to the FC. Prince (2004), when reviewing the research, generally defines active learning as "any instructional method that engages students in the learning process" (p. 1) but clarifies that having students engage in meaningful learning activities and thinking about what they are doing refers to classroom activities as opposed to homework. Prince contrasts the term *active* learning with the *passive* listening done in a traditional lecture.

Active learning is described in Figure 3 from Bishop and Verleger (2013, p. 6). According to this description, problem-based learning (PBL) and peer-assisted learning, partially overlap indicating that PBL can be used individually or in a group. The two key components of peer-assisted learning (i.e., collaborative learning and peer tutoring) also overlap each other. *Collaborative* learning is a broad term for instructional methods where students work together in small groups, usually to achieve a common goal. *Cooperative* learning is a form of group work where students pursue common goals but are assessed individually. At the core of cooperative learning is a perception of interdependence between the individual and the group so that the individual's success is not possible without the group's success (Johnson, Johnson & Smith, 2014).

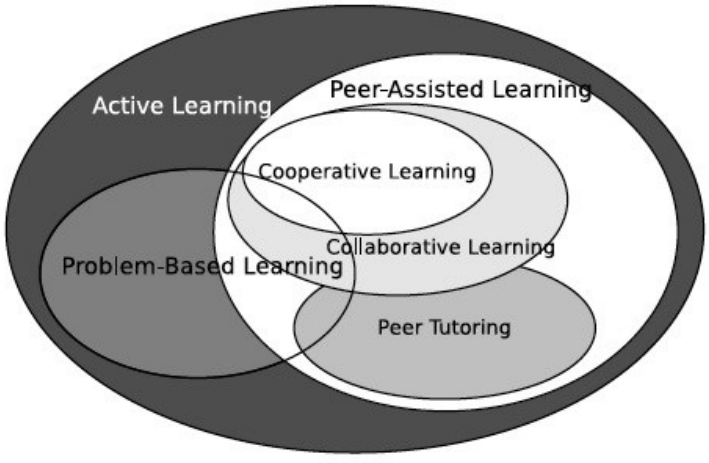


Figure 3. A Venn diagram of active learning (Bishop and Verleger, 2013, p. 6)

Johnson, Johnson, and Smith (2014) synthesized the results of 168 university level studies in a meta-analysis and found that cooperative learning was significantly more effective at promoting higher individual achievement than competitive learning. They also found that cooperative learning was significantly more effective than individualistic learning with average effect sizes of +0.50. These effects are large given that the comparison is education. Most educationally significant outcomes are recognized when the effect size is greater than 0.30 (Lipsey, Puzio, Yun, Hebert, Steinka-Fry, Cole, Roberts, Anthony & Busick, 2012).

Also in support of the effectiveness of active learning is the Freeman, Eddy, McDonough, Smith, Okoroafor, Jordt, & Wenderoth (2014) meta-analysis on university-level studies. In their study, "the active learning interventions varied widely in intensity and implementation, and included approaches as diverse as occasional group problem-solving, worksheets or tutorials completed during class, use of personal response systems with or without peer instruction, and studio or workshop course designs" (p. 1). The Freeman et al. meta-analysis examined 225 studies involving active learning but specifically undergraduate courses in science, technology, engineering and math (STEM), and concluded that active learning significantly increased student performance according to exam scores and reduced failure rates. They found active learning, in comparison to traditional lecturing, to be significantly more effective across all STEM disciplines and all class sizes although it was greatest in smaller classes that had fifty or fewer students.

The results of these two meta-analyses on active learning and cooperative learning respectively (i.e., Freeman et al., 2014; Johnson et al., 2014) could indicate that improvement in the FC may be at least partially due to an emphasis on active learning in general or even cooperative learning (Jensen, Kummer, & Godoy, 2015).

It is important to note that active learning can be used within a lecture as well or to replace a lecture. A simple example of active learning during a lecture is think-pair-share (i.e., asking students to take two minutes to clarify their notes with a partner two or three times during a one-hour lecture) (Prince, 2004). On the FC's video lectures the pause and rewind buttons may serve the same purpose as a two-minute activity by providing the opportunity for a student to stop, reflect on what was said, confirm one's understanding, and prepare oneself to take in the information in the next 15 minutes of lecture. A question embedded in the FC video lectures is sometimes used to encourage students to stop and reflect on or actively engage with the video content. However, the video lecture is primarily direct instruction and not active learning.

Flipped Classroom as a Form of Blended Learning

The FC is a form of blended learning (BL). In the FC the lectures are recorded on video usually available online through the Internet, and students meet face-to-face for active learning in class. BL can be seen as evolving from online learning as it provides some course time online and some face-to-face. Online learning (OL), ranging from computer based training (CBT) to asynchronous online discussions to synchronous virtual classroom, was a natural evolution from the early manifestations of distance education (i.e., which just meant full-time learning at a distance) after the Internet and personal computers became more widely available. Distance education took many forms prior to OL from paper-based correspondence courses, to radio and television broadcasts and included VHS and DVD delivery (Bernard et al., 2004).

BL is considered the *thoughtful integration* of online and face-to-face instruction (Osguthorpe & Graham, 2003) and about "rethinking and redesigning the teaching and learning relationship" (Garrison & Kanuka, 2004, p. 99). Given that BL "combines face-to-face instruction with computer-mediated instruction" (Graham, 2006) the current review defines the

blended FC as including video lectures for homework and active learning during class time, and is a more specific variant of BL.

A BL course has also been referred to as a *hybrid* course although the term *blended learning* (*BL*) has become the most commonly used term (Spring & Graham, 2017). *Hybrid* was used as early as 2002 by the University of Wisconsin-Milwaukee and is still used in some circles.

According to Staker and Horn (2012), who focuses on K-12 education, there are four models of BL with the FC model being a form of the Rotation model. In the K-12 world, students are not sent home to watch the videos but instead rotate through stations in the classroom to watch the videos, solve problems, or receive individual support.

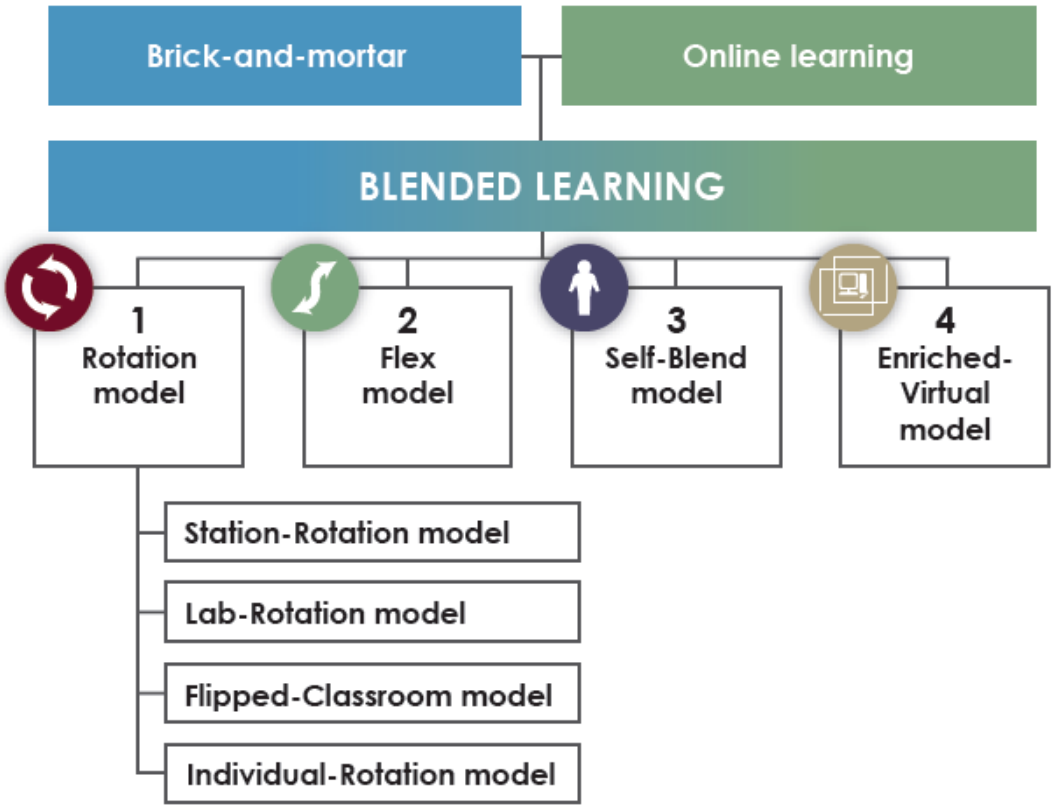


Figure 4. Blended-learning taxonomy (Staker and Horn, 2012, p. 2)

BL has also been defined by the replacement of class time with online time (Garnham & Kaleta, 2002; Owston, York & Murtha, 2013). For those courses that did not reduce class time to compensate for the online time and just added the online time on, BL has been referred to as a *course and a half* (Garrison & Vaughan, 2008). The extra time students spent on such a BL course has been considered as a possible reason some students do better in these courses (Means, Toyama, Murphy & Baki, 2013). In primary studies, it is really just a confound to the design and should be investigated as such.

In order to distinguish BL from online learning and web-assisted courses the Online Learning Consortium (OLC; previously Sloan-C) estimated the percentage of time spent online or face-to-face. For example, OLC has defined BL as having anywhere from 30 to 70 percent of a course online, while Allen and Seaman (2013) estimated between 30 and 80 percent of the course content was offered online. BL is recognized and researched internationally in places such as Africa, Asia, Europe, Latin America, Middle East, North America, and Oceania (Spring & Graham, 2017) so the percentages are discussed globally.

BL itself does not distinguish which aspects of teaching and learning take place online or face-to-face in the classroom. For example, there could be active learning on discussion forums and lectures face-to-face or vice versa or some other configuration. In reaction to such a large range of pedagogical approaches and the resulting difficulty to make comparisons with BL as a cohesive approach, Margulieux, Bujak, McCracken, and Majerich (2014) developed a taxonomy to organize BL approaches. They used this new taxonomy to categorize course design based on the type of instruction (i.e., lecturing content or giving feedback on activities) and how the content was delivered (i.e., via technology or via the instructor). Margulieux, McCraken and Catrambone (2015) recognized the FC pattern as “a flipped blend” which delivered content via

technology and provided feedback through the instructor (p. 220.) This taxonomy provided a framework in which to look at BL courses through a new lens.

The Effectiveness of DE, OL, BL and the Flipped Classroom: An Examination of Meta-Analyses

Since early in 2000, meta-analysis has been used to synthesize the literatures of distance education (DE), online learning (OL), and Blended Learning (BL). See Table 1 below for a summary of these reviews (Bernard, 2017).

Impact of DE. Between 2000 and 2006 seven meta-analyses were conducted on the impact of DE compared with CI (i.e., Machtmes & Asher, 2000; Cavanaugh, Gillan, Kromrey, Hess & Blomeyer, 2001; Shacher & Neumann, 2003; Allen, Mabry, Mattry, Bourhis, Titsworth & Burrell, 2004; Bernard, Abrami, Lou, Borokhovski, Wade, Wozney et al., 2004; Zhao, Lei, Yan, Lai, Tan, 2005; Williams, 2006) creating eight average effects. Seven out of eight average effect sizes were between -0.10 and +0.15 indicating that there was little difference between DE and CI. The category for a small effect starts at 0.20 (Cohen, 1988). Shacker and Neumann (2003) with an average effect of +0.37 was an anomaly, possibly due to their inclusion of only published studies, and a selection bias from the inability of the authors to find 68 (26 percent) full text studies that were to be reviewed for inclusion. Also they chose to include some very high effect sizes that when removed resulted in the average dropping substantially.

Impact of OL. Between 2004 and 2013 five meta-analyses were conducted on the impact of OL compared with CI (i.e., Cavanaugh, Gillan, Kromrey, Hess & Blomeyer 2004; Sitzmann, Kraiger, Stewart, & Wisher, 2006; Jahng, Krug & Zhang, 2007; Cook Levinson, Garside, Dupras, Erwin, & Montori, 2008; Means, et al., 2013. The range of the five average effect sizes

+0.02 to +0.15 indicated little difference between OL and CI. This is not surprising since DE and OL are both full-time, off-campus study arrangements.

Table 1

Summary of meta-analyses conducted comparing DE, OL, BL, and FC to CI (adapted from Bernard 2017, OLC presentation)

Author(s) Publication Year	**Inclusive Years**	**Learner Population**	**DE Context**	**Number of *Effect Sizes/Studies***	**Mean/Sig. (*$p \leq .05$)**
Gillette et al. (2018).	2000-2017	Student pharmacists	FC	5	ns
Cheng, Ritzhaupt, & Antonenko, (2018)	2000-2016	All	FC	55	0.193*
Hu, Gao, Ye, Ni, Jiang, & Jiang (2018)	2015-2017	Nursing students in China	FC	8	1.06*
Tan, Yue, Fu (2017)	unknown	Nursing students in China	FC	16	1.13*
Cirak Kurt et al. (2018)	2010-2016	All in Turkey	BL	32	3.114
Vo, Zhu, & Diep (2017)	2001 +	Higher Education (HE)	BL	51	0.39*
Spanjers et al. (2015)	unknown	All	BL	24 studies	0.34*
Bernard et al. (2014)	2000-2010	HE	BL	117	0.33*
Means et al. (2013)	1996-2008	HE	OL	27	0.05
			BL	23	0.35*
Cook et al. (2008)	1990-2007	Health Workers	OL	63 studies	0.12*
Jahng et al. (2007)	1995-2004	HE	OL	20	0.02
Sitzmann et al. (2006)	1996-2005	Adults	Web-based (OL)	71	0.15
Williams (2006)	1990-2003	Health Workers	All DE	34	0.15
Zhoa et al. (2005)	1966-2002	HE	All DE	98	0.10
Cavanaugh et al. (2004)	1999-2004	K-12	Web-based (OL)	116	0.03
Bernard et al. (2004)	1985-2002	All Learners	Asynchronous DE	174	0.05*
			Synchronous DE	92	–0.10*

Allen et al. (2004)	unknown	unknown	All DE	39	0.10
Shachar & Neumann. (2003)	1990-2002	unknown	All DE	86	0.37*
Cavanaugh (2001)	1980-1998	K-12	DE	19	0.15
Machtmes & Asher (2000)	1943-1997	HE	Tele-Courses	19	–0.01

Impact of BL. Between 2009 and 2017 four meta-analyses compared the impact of BL to CI on achievement and found small but significant effects ranging from +0.33 to +0.39 (i.e., Means et al., 2013; Bernard, Borokhovski, Schmid, Tamim, & Abrami, 2014; Spanjers, Könings, Leppink, Verstegen, de Jong, Czabanowska, & Van Merriënboer 2015; Vo, Zhu & Diep, 2017). The two study features attributed with making the difference were quizzes, (i.e., quizzes in the blended condition when there were no quizzes in the CI condition) (Spanjers et al., 2015), and discipline area, (i.e., the effect was significantly greater for STEM courses than for non-STEM courses) (Vo et al., 2017). In these BL meta-analyses, the FC as defined in this study was not address separately, however, the effect of the use of technology was studied in Bernard et al. (2014). Cirak Kurt et al. (2018) meta-analysis was written in Turkish, and from the English abstract it is impossible to tell why the average effect size is so unusually large, which brings the findings into question.

Impact of the FC. Three meta-analyses were conducted on the impact of the FC as compared to CI. Two of the meta-analyses had a specific focus of nursing education in China (i.e., Tan, Yue & Fu 2017; Hu, Gao, Ye, Ni, Jiang & Jiang, 2018) and the third meta-analysis included kindergarten to postsecondary (K-20) students (Cheng, Ritzhaupt & Antonenko, 2018). Both of the nursing education meta-analyses found the FC significantly outperformed CI with large average effect sizes of greater than 1.00 (e.g., Hu et al., 2018, $\overline{d}$ = 1.06, k = 8; Tan et al.

2017, $\overline{d}$ = 1.13, k = 16). These effect sizes were unexpectedly large in comparison to meta-analyses done on BL where average effects were closer to +0.33 to +0.35. Cohen (1988) is often cited as broadly categorizing effect sizes of 0.20 as small and 0.50 as medium, while Tallmadge (1977) had indicated 0.25 as the marker of educationally significant outcomes, while more recently Lipsey et al. (2012) noted that effect sizes in education rarely are as large as 0.30. Either way, these effect sizes of greater than 1.00 are anomalous and indicate a need to critically question the outcomes.

On closer inspection, Hu et al. (2018) was based on a small number of studies in which anomalously large effect sizes have a greater impact on the resulting average effect size (e.g., the two largest effect sizes were 1.59 and 1.68). There were only eleven studies from 2015 to 2017 in the entire meta-analysis, however, after sensitivity analysis the knowledge scores were based on only eight studies. Hu et al. also included only randomized control studies (RCT), excluding 194 studies because they were not RCTs. Even though Hu et al. claimed there was no publication bias according to the visual examination of the funnel plot, they were clear that the unpublished literature had not been searched, and that they excluded any conference abstracts that might otherwise have been included. The small number of studies may have prevented the funnel plot view from showing the inherent publication bias from not searching or including unpublished studies and conference papers. Publication bias creates a higher average effect size because studies with significant results are more likely to be published in academic journals, and studies are "more likely to be statistically significant if the effect size is larger" (Borenstein, Hedges, Higgins & Rothstein, 2011, Ch. 30, p. 283) yet RCT generally create lower average effect sizes than quasi-experimental studies. Perhaps the large effect sizes were due to some limitations of the studies that were included. Hu et al acknowledged low methodological quality in regard to

randomization methods and lack of blinding of assessors resulting in potential selection and detection bias. When an assessor grades a paper that they know belongs to the treatment group, they may want them to do better and thereby unintentionally grade it more leniently. This bias could result in higher effect sizes than normal.

In 2017, Tan et al. also included only peer-reviewed randomized controlled trials of nursing studies in China so publication bias might have inflated the average effect size yet again the high quality of RCT studies would have reduced the effect. Tan et al. found the FC created significant academic improvements in knowledge ($\overline{d}$ = 1.13) compared to the traditional CI based 16 studies. Selection and detection bias may have inflated the results in this study as well.

Perhaps the overarching issue in these two meta-analyses (i.e., Tan et al., 2017; Hu et al., 2018) is the small number of studies included (i.e., 16 and 8 respectively). Given both meta-analyses are working with the random-effects model, the dispersion in effects is assumed to be real as opposed the fixed-effect model where the dispersion in effects is assumed to be a result of sampling error (Borenstein et al., 2011). When a meta-analysis is "based on a small number of studies, the estimate of between-studies variance (T^2) may be substantially in error" (Borenstein, et al., 2011, Ch. 40). Borenstein et al (2011) indicate that because the standard error of the average effect size is based on this between-study variance (T^2) the resulting average effect size and the confidence interval may be wrong. They also note that with few studies we cannot tell if the dispersion effect is consistent or varies across studies. For this reason, Borenstein notes it maybe better not to summarize studies when the number is small as the results may be misleading. With this in mind, the data from the studies included in the Tan et al. and Hu et al. meta-analyses should be viewed individually and not in the form of an average effect-size.

The third FC meta-analysis by Cheng et al. (2018) found an average effect size of +0.193 in favour of the flipped classroom based on 55 combined effect sizes of K-20 students. This resulting average effect size was lower than the average effect size found for BL but closer in proximity than that of Tan et al. (2017) and Hu et al. (2018). Cheng et al. indicated that they did not code for study quality, and they averaged effect sizes from 115 assessments taken throughout the courses to create 55 effect sizes as opposed to having taken the most cumulative assessment such as the final exam mark. There were thirty-nine studies based on undergraduate students, four based on graduate students, and twelve based on K-12 students.

There have been a number of reviews, other than meta-analyses, on the FC. Margulieux, McCracken & Catrambone (2015) conducted a vote count and found that 17 of the 21 flipped courses added instruction during application and reported improved learning outcomes. In 2015, O'Flaherty and Phillips' scoping review found that the FC literature was lacking any conclusive evidence that it was more effective than the traditional CI approach yet they recognized the importance of the pre-class quiz results to the instructor's ability to address students' misconceptions. Three qualitative reviews were conducted in nursing education (i.e., Betihavas, Bridgman, Kornhaber, & Cross, 2016; Presti, 2016, Njie-Carr, Ludeman, Lee, Dordunoo, Trocky, Jenkins, 2017). Betihavas et al. (2016) included nine studies, Presti (2016) reviewed 13 studies, and Njie-Carr et al (2017) reviewed 13 nursing studies, and they all found neutral or positive results.

Two FC reviews were conducted from an engineering perspective. Bishop & Verleger (2013) conducted a survey of the research and Karabulut-Ilgu et al., (2018) conducted a qualitative review and a vote count. Bishop and Vergeler found that most research as of 2013 was focused on student perceptions. By 2018, Karabulut-Ilgu et al. found 30 studies that directly compared student achievement in traditional and FC but the results were tabulated as a vote

count, and were inconclusive. Thirteen studies clearly indicated that FC was more effective but only seven of these were statistically significant; four studies had mixed results, two showed the FC underperforming the traditional CI, while eight indicated no difference. Karabulut-Ilgu et al.'s systematic review reported that there is a rapidly increasing interest in the flipped classroom approach in engineering and a meta-analysis would be ideal to make a definitive statement about whether the flipped classroom's approach has any advantages over the traditional one.

The Methodology of Systematic Review and Meta-Analysis

The literature review on meta-analysis introduces a brief history of this research method, briefly discusses the potential bias that can exist in each of the seven steps involved in a meta-analysis, with a more in-depth look at three (i.e., publication bias, outliers, and extreme effect sizes from large samples).

Brief history. Gene Glass developed meta-analysis in 1976 as a way to quantify the standardized size of the effect that a treatment has on an outcome in comparison to a control condition. Prior to meta-analysis, the options for a synthesis of the literature were either a qualitative narrative review of the literature, resulting in a compilation of descriptions of studies, or a vote count approach that gives one vote for or against the treatment over the control condition based on the significance level reported in the study, regardless of how many participants were involved in the study. While any of the forms of review can qualify as a systematic review, described by the Cochrane Collaboration as "an appraisal and synthesis of primary research papers using a rigorous and clearly documented methodology in both the search strategy and the selection of studies. This minimizes bias in the results. The clear documentation of the process and the decisions made allow the review to be reproduced and updated." (Cochrane

Handbook for Systematic Reviews of Interventions, 2011, https://libguides.library.qut.edu.au/systematicreviews).

A meta-analysis specifies: (1) quantitative data; (2) effect sizes that are defined by a difference between groups, a correlation, or a 2 X 2 frequency table from which an odds-ratio or risk ratio can be calculated; and (3) Statistical methods are then used to synthesize the data.

Seven steps. Meta-analysis follows seven steps (Cooper, 2017) and each one has the potential for bias (Bernard et al., 2014). The steps and a brief reference to potential biases follow.

1. **Formulating the problem.** In order to formulate the problem, one should refer to at least one past meta-analysis and show that studies available for review are relevant to the topic.
2. **Searching the literature.** Searching the literature involves searching for all relevant studies that are published or unpublished. Searching the literature would involve determining key words and systematically searching online databases as well as manually searching through reference sections of relevant studies. There is a potential for publication bias if the researcher only sought published journal studies. Selection bias may appear when deciding on search terms. If a keyword is left out, a search may systematically miss studies that should be included. (Kugley, Wade, Thomas, Mahood, Jørgensen, Hammerstrøm & Sathe, 2016).
3. **Formulating criteria for including and excluding studies.** When formulating criteria for including and excluding studies, a researcher should consider key aspects of the control and treatment group being studied (e.g., the flipped classroom required video lectures as homework), relevant years of interest, and data required to calculate an effect size.

4. **Extracting effect sizes and moderator effects.** To extract the effect sizes for each study, data such as means and standard deviations for the control and treatment groups are determined. A code book is created to indicate which study features are to be coded as moderator variables and the levels for each. Ideally, two independent reviewers are trained to agree on studies to include/exclude, to extract the effect sizes, and to code the moderator variables. The inter-rater reliability (Cohen's *K*) should be presented in the meta-analysis to show the lack of bias and the reliability of the coders. If the meta-analysis only has one coder, this person could be unknowingly biased.
5. **Assessing the quality of the studies.** While extracting effect sizes and coding moderator variables, the researcher has to assess whether a study fits the quality needed. For example, a researcher should never combine outcomes that are not the same category of outcome measure, such as attitude and achievement. Small sample sizes should be converted from Cohen's *d* to Hedges' *g*. Samples should be independent (i.e., not using the same participants more than once) if taking multiple effect sizes from one study. Any of the threats to validity (Campbell & Stanley, 1966; Shadish, Cook & Campbell, 2002) could make for a poor quality study. Randomized Control Trials (RCTs) are the "gold standard," high-quality Quasi-experimental Designs (QEDs) are acceptable, but pre-experimental studies are generally to be unacceptable except in certain instances (e.g., *N*-of-one studies).
6. **Evaluating the research outcomes.** Outliers and extreme effect sizes need to be determined. CMA's one-study-removed and the funnel plot can assist in this process. These are discussed in more detail in the following section.
7. **Interpreting the results.** The researcher should note the limitations of the study, and ensure the results are those of interest to the reader and not just the researcher.

Publishing the paper is a way of sharing the results with a community of people who are interested. In order to assure that important information is included there are some standards developed for reporting meta-analysis: PRISMA (Preferred Reporting Items for Systematic Review and Meta-analyses; Moher, Liberati, Tetzlaff, Altman, The PRISMA Group, 2009) and MARS (Meta-Analysis Reporting Standards; APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) being two. There should be a code of ethics to guide the author, the reviewer, and the publisher on how to avoid bias in meta-analysis (Bernard et al. 2014).

Two sources of bias in Meta-analysis. Two sources of bias addressed below are publication bias during the literature search, and extreme influence effect sizes from large samples in combination with using the correct model for analysis (e.g., fixed-effect versus random-effects model).

Publication bias. When searching the literature, there is potential for publication bias as a result of only searching within the published literature (e.g., peer-reviewed journals) for studies to include in the meta-analysis. Journals are known to more likely publish large studies and those with significant effects (Borenstein et al., 2011). If the researcher only searches journals, online or paper, then they may be missing all the other studies that could provide different results. A meta-analysis is meant to include all studies on the topic, not just a sample, so it is important to look further. The solution to this bias is to search, not only in journal databases, but also in what is called the *grey literature* to find the remaining studies that have not been published (Bernard et al., 2014). Some ways to find grey literature include searching through the reference section of studies that were previously found; searching prominent journals that are not online, or back issues of journals that are only online after a certain year; conference papers and proceedings on the topic; and dissertations and theses.

To explore whether the meta-analysis has publication bias, the researcher should use Comprehensive Meta-analysis™ software or other dedicated meta-analysis software (e.g., RevMan) to create a **funnel plot** diagram which indicates the effect size on the x-axis (with zero in the middle, negative effect sizes on the left and positive effect sizes on the right), and a reflection of the sample size, and usually the standard error on the y-axis (zero on the top, so that relatively large studies appear at the top of the plot (relatively low standard errors) and small studies with larger standard errors at the bottom of the plot. The effect sizes show a visual assessment of potential bias. The most common area missing studies is at the bottom left of the plot as that is where small samples and negative effect sizes would be located (Borenstein et al., 2011). To conduct a statistical analysis of this funnel plot, the researcher should use the *Fill and Trim* method (Duval & Tweedie, 2000) through Comprehensive Meta-analysis to report if there is any publication bias showing and what would happen to the average effect size if these studies were added or trimmed (trim and fill is linked with the funnel plot). *Orwin's Fail-safe N* can also be used to assess potential publication bias by noting how many studies with an effect size of zero it would take to reduce the average effect size to zero. Orwin's *Fail-safe N* (as opposed to Rosenthal's fail safe N) will also give the options to choose the average effect size that would be considered negligible (e.g., $\overline{g}$ = 0.05) as well as an alternative effect size (i.e., the minimum expected) to zero of any missing studies (Borenstein et al., 2011).

Extreme influence effect sizes. When evaluating the research outcomes there is the potential for extreme influence effect sizes from very large samples especially on the margins (+ or -) of the effect size distribution. Studies like this can skew the average effect size, either positively or negatively. Using the random-effects model, as contrasted with the fixed effect model, can reduce the influence of these kinds of effect sizes. If, for example, most of the studies

include samples of 100 and one study has 5,000 participants, this large study could have a much greater influence on the average effect size, especially if the effect size is very large or small. The effect size may not be large but it will have a disproportionally large influence if the sample size is large. The influence would be greater if the fixed-effect model results were used instead of the random-effects model results. The fixed-effect model uses a different weighting system than the random effects. The fixed-effect model weight is 1/variance $_{within}$ while the random-effects model uses 1/ variance $_{within}$ + average variance$_{between}$ groups (tau^2). Because the random weights are smaller than fixed weights, the random-effects model allows less influence from studies with high/low and relatively large sample sizes. Such a study could also be an outlier in which case there is the option to remove it or truncate it (i.e., Winsorizing). Comprehensive Meta-analysis ™ software has a program called *One-Study Removed* that indicates what the average effect size and standard error would be with each individual study removed in turn. The average effect size may change dramatically when an outlier is removed, especially if it is a high influence effect size. The funnel plot can also provide a visual depiction of potential outliers.

Solution**:** The social sciences should use the random-effects model because not all studies have the same kind of treatment, outcome measure, sample, etc. and there is not one true effect to be found (Borenstein et al., 2011).

Goals and objectives. Systematic review and meta-analysis is used in this dissertation to answer the question "Is there a difference between FC and CI on achievement outcomes in higher education?"

The Purpose of this Study

The purpose of this study is to compare the impact of the Flipped Classroom on higher education student achievement in comparison with Classroom Instruction (i.e., primarily lecture-

based). Study features were coded and analyzed to determine if they explain a significant amount of the difference. Figure 5 shows the position of this research in relation to the research on blended learning (a combination of Classroom Instruction and Online Learning as shown by the overlapping circles in the top part), active learning, collaborative learning and cooperative learning, as well as the two moderators that were found to be significant in the BL literature (i.e., subject matter and the use of quizzes).

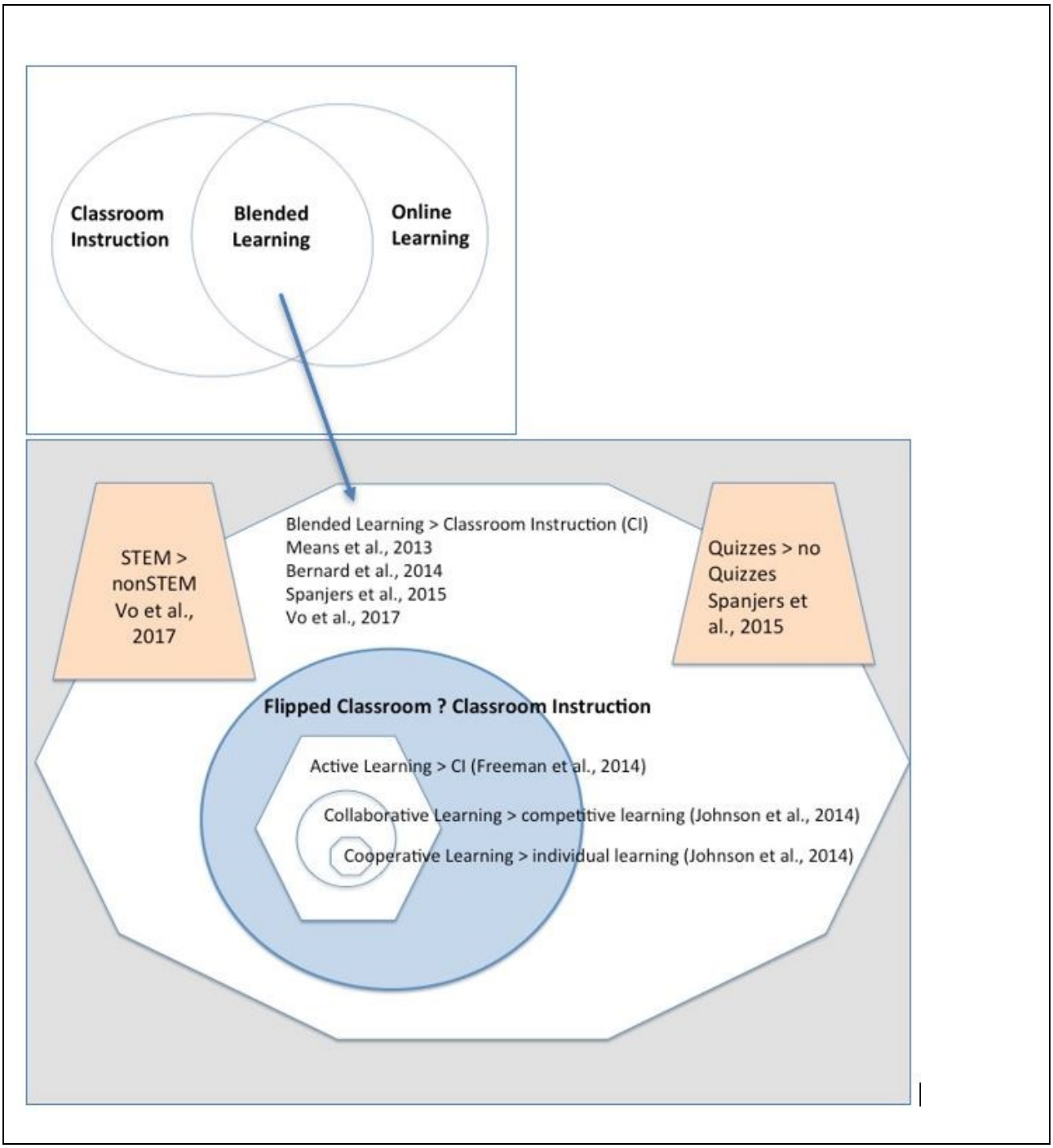


Figure 5. Flipped Classroom concepts in relation to the research.

Chapter 2 (Method) describes the methods that were used at each stage of the meta-analysis, Chapter 3 (Results) outlines the results, and Chapter 4 (Discussion) discusses the findings as they relate to the literature.

CHAPTER 2: METHOD

"Research syntheses focus on empirical studies and seek to summarize past research by drawing overall conclusions from many separate investigations that address related or identical hypotheses."

(Cooper, 2017, p. 4)

As the above quote from Harris Cooper, a recognized authority on meta-analysis, foretells, this study is a research synthesis that summarizes empirical studies on the same question. In this case, the question of interest is, "What is the impact of the Flipped Classroom (FC) compared to Classroom Instruction (CI) on achievement in higher education?" Data from the studies selected to address this question were analyzed using a set of statistical procedures, collectively referred to as *meta-analysis*, that were first introduced by Gene Glass in 1976. Meta-analysis normally proceeds from a systematic review (Cooper, 2017), whereby a research question is identified, definitions of terms and inclusion/exclusion criteria are specified, and an iterative process of search, retrieval, and selection of studies that meet the inclusion/exclusion specified is followed. The first section of this chapter describes the research synthesis and meta-analytic procedures that were implemented according to the following key areas: (1) literature search strategies; (2) selecting studies, extracting effect sizes and coding study features; and (3) statistical methods. The research questions, terms with definitions and inclusion criteria have been added at the beginning for context.

Research Questions, Terms and Definitions, and Inclusion/Exclusion Criteria

Research questions. Three qualitative surveys of the FC literature (Bishop & Verleger, 2013; Margulieux et al., 2015; O'Flaherty & Phillips, 2015) indicated that there is an interest in and need for a study comparing student achievement in the FC (or 'inverted classroom') with the traditional CI, featuring mostly lecture-based teaching. This meta-analysis attempts to address this need by answering the following research questions:

- Does an analysis of methodological factors associated with the distribution of effect sizes extracted from studies (i.e., publication bias analysis, sensitivity analysis and research design analysis) suggest that a meta-analysis of the literature is advised?
- What is the impact of the FC (i.e., courses where video lectures are watched at home and class time is spent on student-centered activities) compared to their traditional CI counterparts on the learning achievement of higher education students in formal educational settings?
- How do course demographic study features (e.g., course subject matter) moderate the overall average effect size?
- How do pedagogical factors, (e.g., regular quizzes in the FC condition) moderate this effect?

Terms and definitions. The following terms and definitions are used in this meta-analysis:

- *Flipped Classroom (FC)*: Students use video lectures for homework followed by active learning in the classroom. The flipped classroom is also known as the *inverted classroom or reversed instruction* in some studies. The flipped classroom in this study is a form of blended learning.

- *Classroom Instruction (CI)*: Classroom time is primarily a lecture-based teaching approach. Classroom Instruction is also known as the traditional classroom). This lecture-based approach could be delivered in person for a campus-based course or online for a distance course. Homework is usually reading and active learning exercises and applications.
- *Student Achievement*: Learning achievement is measured by performance on final exams, midterms and other tests. The final exam score is the first-choice indicator of student achievement. Alternatively, a midterm that assessed the intervention is the second option. The third option is a final grade in the course, but because this measure is usually a composite of many forms of evaluation (e.g., attendance) it is not considered the ideal measure of achievement effects. The number of students who pass and fail or withdraw is used as a last option. The last two options are used only if no other information on student achievement is available.
- *Active learning*: Active learning takes the form of designed activities using problem-based learning, peer-assisted learning, peer tutoring, collaborative learning, and cooperative learning (Bishop & Verleger, 2013). In some studies active learning primarily took the form of students completing work in class that was also assigned to the control group as homework.

Inclusion and exclusion criteria. Included studies met the following criteria:

- Published and unpublished studies in English were available through a systematic search as outlined later in the Search Strategies section.
- Studies were published or otherwise available from 2000 through 2017.
- Students were in higher education courses (e.g., university or college)

- FC treatment group used video lectures for homework and active learning in the classroom.
- The control group received primarily lectures in class or online.
- Study design was either two-group experimental or quasi-experimental (i.e., with some method of determining prior group equivalence).
- Data sufficient to calculate an effect size were available.

The year 2000 was used as a starting year because that was when Lage et al. (2000) coined the term *inverted classroom* while teaching an undergraduate economics course even though it was Bergmann and Sams (2008) who popularized the term *flipped classroom*.

The included studies took place in institutions of higher education. All other forms of adult education, including employer run courses, were not part of this review. The definition of FC required video lectures for homework, which meant that studies that used printed reading homework alone were not included. The FC in-class participation was required to be predominantly active for the student (i.e., the in-class component could not be focused on lectures). Some micro-lectures were permitted in class to correct student understanding. Measures of learning outcomes in both treatment and control conditions needed to be equivalent. Sufficient data available to calculate an effect size with independent samples was required. The sample size was a necessary statistic for retaining a study because it is needed to calculate the standard error. To be included, the study design needed to be *experimental* (i.e., random assignment of participants to both the control and the treatment groups), or *quasi-experimental* (i.e., the experimental and control groups were shown to be equivalent from the beginning). Excluded were one-group pre-test post-test design studies, as were two group *pre-experimental* studies where experimental and control pre-experimental group equivalence could not been ascertained.

Literature Search Strategies and Search Outcomes

Searching the literature involved a search strategy including keywords and places to search as outlined below:

Search strategy. Keywords searched included variations on the following depending on the options best suited to the various online databases: (flip* OR invert*) in the title AND (undergraduate* OR postsecondary OR university OR college OR higher education OR tertiary) AND (outcomes OR achievement) AND (comparison OR experiment OR quasi-experiment OR evaluation).

The databases searched included the following: ERIC (EBSCO), ABI InformGlobal (ProQuest), Academic Search Complete (EBSCO), CBCA Education (ProQuest), Communication Abstracts (CSA), EdLib or AACE Digital Library, Education Source (EBSCO), Education: A SAGE Full-text Collection, Francis (CSA), Medline (PubMed), ProQuest Dissertation and Theses, PsycINFO (EBSCO), Australian Policy Online apo.org.au, and Social Science Information Gateway.

The search strategies outlined by the Campbell Collaboration Policy Brief on Searches (Kugley et al., 2016). guided the electronic and manual searches. Grey literature was sought through web searches, branching of the qualitative reviews, subject indexes, and manual searches of key journals and conference programs.

Search Outcomes. In total, the search produced 1,442 abstracts from the database search and 259 from branching of which 1281 remained after duplicates were removed. Of these 1,047 were excluded due to being the wrong topic, the wrong population, the wrong language or the wrong year. The remaining 234 full text articles were reviewed for eligibility. One hundred and twenty studies were excluded for various reasons including study design quality. The remaining

114 studies met all the inclusion criteria, and produced 125 effect sizes that were included in this meta-analysis. See the PRISMA Flow Diagram in Figure 6 for a graphical view of this data.

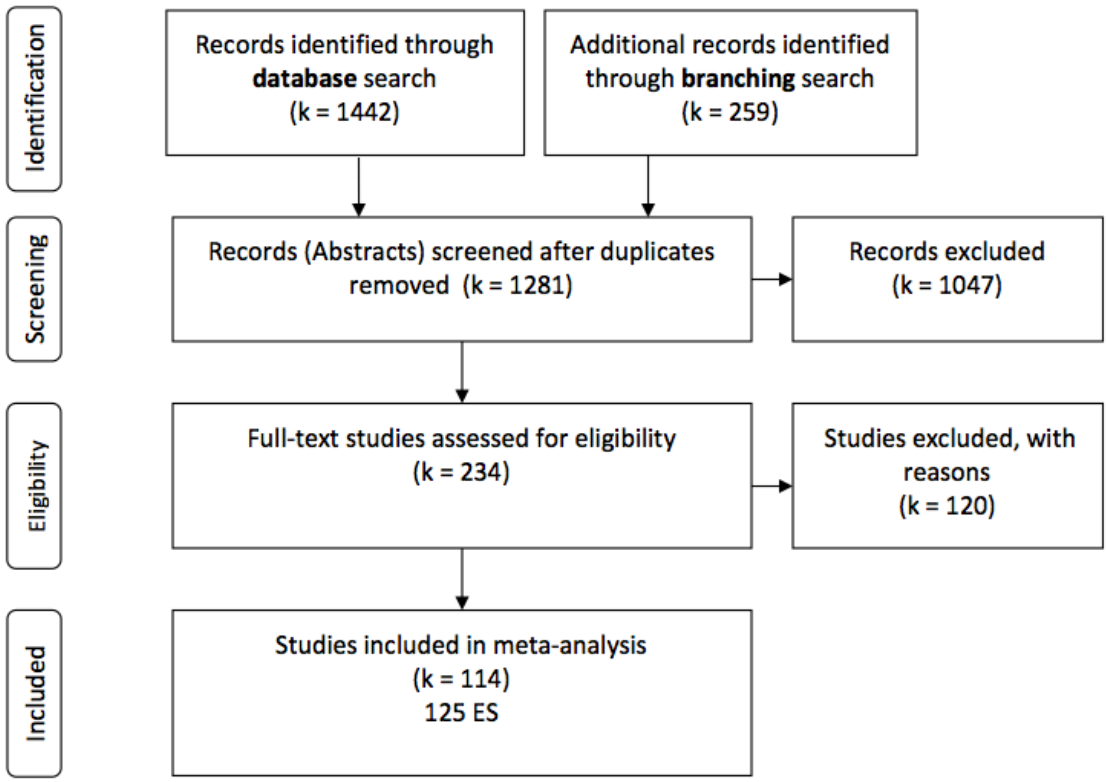


Figure 6. PRISMA flow diagram

For more about the reasons why studies were excluded see Appendix A.

Selecting Studies, Extracting Effect Sizes, and Coding Study Features

This section involves four steps: (1) selecting the studies for inclusion, (2) identifying the number of effect sizes, (3) extracting the effect sizes and (4) coding the study features (moderator variables). Details about each step follow.

Selecting Studies for Inclusion. An Information Specialist retrieved the abstracts for all the studies found. *Two* coders worked independently on randomly selected 25% of the *abstracts* to help reduce (or identify) coding bias. Any discrepancies between the two coders were

discussed until a consensus was reached. After establishing an acceptable level of inter-coder reliability, the author continued to make the decisions regarding the remaining studies.

The information specialist then retrieved the *full text* for the studies determined to fit the inclusion criteria. The two coders again worked independently on randomly selected 25% sample of the studies to extract effect sizes from the full text, and code moderator variables. The inter-rater reliability was 90.9 (or Cohen's Kappa of 0.82), which ensured the coders were working under the same assumptions and basically making the same coding decisions. Any discrepancies between the two coders were discussed until a consensus was reached. The author continued to code the remaining studies.

Identifying the number of effect sizes. When more than one independent subgroup was compared in a study, more than one effect size was extracted. Each subgroup contributed independent information so they were treated as if they were independent studies. The control group needed to be independent for each comparison so that no participants were counted twice. If there was only one control group and two comparison groups there was a choice to be made. The options included (1) randomly select one group to include, or (2) choose the more representative group to include. A type II error is the potential result if participants are repeatedly used in calculating more than one effect size. A Type II error would occur if the results indicated there was a significant difference but in reality there was not one.

Extracting effect sizes. Cohen’s d was used to calculate the effect size by subtracting the mean of the control (C) from the mean of the treatment/experimental group (E), and dividing the difference by the pooled standard deviation (SD) of the two groups as seen in the equation below.

$$d = \frac{\bar{X}_E - \bar{X}_C}{SD_{Pooled}}$$

The SD_{pooled} was calculated by applying the following formula:

$$SD_{Pooled} = \sqrt{\frac{(n_E-1)SD_E^2+(n_C-1)SD_C^2}{(n_E-1)+(n_C-1)}}$$

To overcome the problem of small-group bias (i.e., those groups with fewer than 20 individuals), the Hedges' *g* multiplier was applied to all studies. Hedges' *g* corrects small study bias while not affecting larger studies. The correction factor called *J* is used to convert Cohen's *d* to Hedges' *g* (Borenstein et al., 2011).

$$J = (1 - \frac{3}{4df-1})$$

Also seen as:

$$J = (1 - \frac{3}{4N-9})$$

then, $g = d \times J$

or

$$g \approx d(1 - \frac{3}{4N-9})$$

The variance of Cohen's *d* is calculated as follows:

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

The standard error is the square root of the variance of Cohen's *d*,

$$SE_d = \sqrt{V_d}$$

The variance of Hedges' g was derived from J squared times the variance of Cohen's d:

$$V_g = J^2 \, x \, V_d$$

$$SE_g = \sqrt{V}_g$$

To determine if the average effect size was educationally significant and not just statistically significant this study followed Lipsey et al. (2012) who set the threshold for the average effect size to pass at 0.30. Lipsey et al.'s approach is more tailored to educational studies than Cohen's (1988) general guidelines (i.e., small average effects are where *d* starts at 0.20; medium starts at 0.50; and large starts at 0.80.

Evaluating the quality of studies. Evaluating the quality of studies involved (1) assessing the methodological quality, (2) assessing the publication bias, and (3) performing a sensitivity analysis.

To determine if there was a difference in methodological quality between research designs, studies were coded as experimental or quasi-experimental and analyzed to see if the average effect sizes were significantly different. Given there was no significant difference between the two groups, the quality of both research designs were considered equivalent.

Publication bias analysis was used to determine if a theoretically large number of studies were missed or not present when searching the literature, and to what extent those hypothetical missing studies would have changed the resulting average effect size. A funnel plot, Duval and Tweedie's (2000) Trim and Fill, and classic fail-safe analysis procedures were used to analyze the data for publication bias. These procedures were performed in *Comprehensive Meta-Analysis*™. The funnel plot graphically showed where each study's effect size is along the range

of standard errors (i.e., an indicator of sample size) so the reader can visually see how the studies were distributed. Duval and Tweedie's trim and fill indicates where studies' effects sizes on the funnel plot would need to be removed (trimmed) or added (filled) to make the funnel plot symmetrical around the mean. The ***classic fail-safe*** analysis indicates how many studies of null effect would need to be found to make the effect size insignificant = α.

Sensitivity analysis helps to determine if there are outliers skewing the results. Comprehensive Meta-Analysis ™ was used to perform the One-Study-Removed –Analysis where the average effect size and relevant statistical data were re-calculated as each study was removed in turn.

Coding Study Features (Moderator Effects). Studies were coded for the following features: research design, outcome measure, size of experimental and control groups, effect size, and direction of the effect, publication year, publication type, same or different instructor, same or different semester, graduate or undergraduate course, discipline (e.g., STEM, Health-related, or non-STEM), elements of STEM, instructor experience, whether the control was strictly lecture or it included some active learning, as well as, whether pre-class quizzes were used in the FC. See Appendix B for the codebook including levels of each study feature.

Statistical methods

Comprehensive Meta-Analysis ™, version 3.3.070 (Borenstein et al., 2014) was used to calculate the overall weighted mean estimate of the treatment effect (i.e., Hedges' $\overline{g}$) from the effect sizes. The Q_{Total} statistic (Hedges & Olkin, 1985) was used to test the assumption of heterogeneity of effect size. The descriptive statistics I^2 (i.e., percentage of true variation beyond sampling error) and tau-squared (τ^2), a measure of average heterogeneity present, were also interpreted.

Meta-analyses in education and the social sciences in general, typically find the effect size distributions to be heterogeneous because experiments in these areas can vary widely in many ways. For instance, primary studies of education are often conducted by different researchers/teachers, using different research designs, measured with different instruments and on students of various levels. The *random-effects model* is the most appropriate approach for systematic review analysis in areas such as education where the true effects are known to be heterogeneous, and this model should be chosen *a priori* before any analyses or tests of heterogeneity are conducted (Borenstein et al., 2011). As a result of using the random-effects model, less weight is given to large studies and more weight to smaller studies than the fixed-effect model, because each individual study does not represent one true effect, but represents an average effect of a hypothetical population of like studies.

The *fixed-effect model* is typically used in areas where the treatment and experimental conditions are very consistent, such as pharmaceutical trials, because it assumes there is a single true effect average size and that the variation among studies is limited to sampling error (i.e., between-study variability is low).

With a heterogeneous result, Q_{Total} and I^2 show how much variability there is in the group of studies. This extra variability indicates that there is potential for moderator variables to explain some of the variation. The *mixed-effects model* was used for moderator analysis. Categorical and continuous moderator variables are analyzed differently to determine if they account for between study variability; categorical variables (e.g., subject matter) use a method equivalent of *ANOVA in that it tests differences among groups*, while continuous variables (e.g., publication year) use random effects weighted meta-regression.

Interpreting the evidence. Interpreting the evidence involved discussing the results and drawing conclusions. This step provided the opportunity to review the research questions and

discuss how the data helped to address them. Any practical, theoretical and/or conceptual implications were considered in context with the literature as well. In this case the FC was compared to traditional CI for its effect on achievement in higher education. Moderator variables were tested to determine if any produced a significant difference between/among categories.

Conclusion

This chapter reviewed the method used to conduct this systematic review, and meta-analysis. The method section began with the statement of research questions, and definitions of the terms flipped classroom, traditional classroom, student achievement, and active learning. This chapter then described the meta-analysis procedures according to: (1) search strategies, (2) selecting studies, extracting effect sizes, and coding study features, and (3) statistical methods. The results and the discussion sections follow.

CHAPTER 3: RESULTS

In this chapter results are presented in three sections, the first of which presents the overall average effect size of the random-effects models. The second section addresses publication bias including the funnel plot. The third section provides the test of the moderator variables, (e.g., methodological, demographic, and pedagogical).

Overview of Included Studies and Average Effect Size

Overview of included studies. This meta-analysis of the flipped classroom included 114 studies and 125 effect sizes. Nine studies had multiple independent effect sizes where each treatment group had its own independent control group (i.e., Gillispie, 2016; Haughton & Kelly, 2015; Horton, Craig, Campbell, Gries & Zingaro, 2014; Hu, Montefort & Tsang, 2017; Lape, Levy, Yong, Haushalter, Eddy & Hankel, 2014; Margoniner, 2014; Prescott, Woodruff, Prescott, Albanese, Bernhardi & Doloresco, 2016; Quint, 2015; Ruddick, 2012;). All included studies were quasi-experimental or experimental.

Average effect size. The weighted effect sizes were analyzed using two different models, the random-effects model, and the fixed-effect model. The random model was used to form an overall average for the collection ($\overline{g}$). This analysis is shown in the upper part of Table 2. The fixed effect model was used to estimate within-group heterogeneity in the collection. Table 2 shows that the overall weighted average effect size is ($\overline{g}$ = 0.300, k = 125, $p < 0.001$) using the random-effects model.

Table 2

Overall results

Model	**Effect size and 95th Confidence Interval**					**Test of null**	
Random Effects	*k*	$\overline{g}$	*SE*	Lower 95th	Upper 95th	*z*-value	*p*-value
Total Collection	125	0.300	0.041	0.220	0.380	7.365	0.000

Model	**Heterogeneity**				
Fixed Effect	*Q*-value	*df*	*p*-value	I^2	Tau^2
Total Collection	1487.391	124	0.000	91.663	0.160

In this meta-analysis, as in most social science meta-analyses, the random-effects model was the most appropriate model to use to report the average overall effect size because the studies differed in terms of sample size, subject matter, and research design (Borenstein et al., 2011). Statistically these differences are confirmed by the significant heterogeneity shown in the Q-value in Table 2.

A forest plot helps to provide the context in which to interpret the statistics as it "shows if the overall effect is based on many studies or a few, on studies that are precise or imprecise; whether the effects for all studies tend to line up, or whether they vary substantially from one study to the next" (Borenstein et al, 2011, Ch. 41). The forest plot shows each study and the summary effect as a point estimate with its boundaries in the form of a confidence interval. Appendix C: Forest Plot shows the 125 effect sizes on which the average random effect size is

based. Although most studies have a positive point estimate, many confidence intervals cross the zero line into the negative range indicating that they are not significantly different than zero. Some studies have shorter confidence intervals showing the results were more precise than those with longer confidence intervals. Notice the point estimates do not all line up but instead range from +1.500 at the top and -0.656 at the bottom reflecting the heterogeneity as noted above in the Q-value.

The ***One Study Removed*** analysis, as seen in Table 3 (shows only the top six and the bottom six effect sizes), indicated that there was no meaningful difference to the average effect size even when each study was systematically removed, and hence no study effect sizes were considered outliers.

Table 3

Sensitivity analysis

Study Names $K = 124$	Actual g	One Study Removed						Relative Weight
		$\overline{g}$	*SE*	Lower 95th	Upper 95th	*z*-Value	*p*-value	
Wong2014	1.50	0.29	0.04	0.21	0.37	7.15	0.00	0.87
Reza2015	1.46	0.29	0.04	0.21	0.37	7.17	0.00	0.68
Turan2016	1.04	0.29	0.03	0.22	0.35	9.21	0.00	1.04
Prescott2016b	0.98	0.29	0.04	0.21	0.37	7.20	0.00	0.93
Pereira2007	0.94	0.29	0.04	0.21	0.38	7.21	0.00	0.85
Li & Dan2015	-0.31	0.31	0.04	0.23	0.39	7.49	0.00	0.92
Cobb2016	-0.43	0.31	0.04	0.23	0.39	7.47	0.00	0.70
Witman Cobb2013	-0.43	0.31	0.04	0.23	0.39	7.48	0.00	0.73
Kang2015	-0.51	0.31	0.04	0.23	0.39	7.50	0.00	0.74
Hu2017d	-0.57	0.31	0.04	0.23	0.39	7.56	0.00	0.87
Moffett2014	-0.66	0.31	0.04	0.23	0.39	7.61	0.00	0.90
Overall (k = 125)	0.30	—	0.04	0.22	0.38	7.36	0.00	100.00

Explanation: The studies selected and displayed in this table are those that lower (above the middle line) the overall average effect size below $\overline{g}$ = 0.30 when they are individually removed (i.e., one study removed) and ones (below the middle line) that raise the average effect size of $\overline{g}$ = 0.30 when they are removed. All of the other effect sizes in the distribution neither lower nor raise the one study removed effect size beyond the overall average effect size of $\overline{g}$ = 0.30 and *SE* = 0.04. Outliers (i.e., high influence effect sizes) are a function of two variables either working separately or in combination: (1) the magnitude of the study's effect size (either higher or lower effect sizes); and (2) the relative weight that studies are given in the synthesis of studies. Large studies (relatively speaking) have smaller standard errors and larger weights. So large *N* studies with a smaller SE have a smaller *CI* (and thus are more likely to be significant) and small N studies with a larger SE have a wider *CI* that is more likely to cross 0 (and not be significant).

Another aspect of the table to look at is the standard error of the one study removed group and the overall average standard error. Only in one case do they vary, and then by only 0.01. Similarly, the upper and lower boundaries of the 95th confidence interval (based on the average effect size and the standard error [Lower Limit = ($\overline{g}$) - (1.96 x SE) and Upper Limit = ($\overline{g}$) + (1.96 x SE)]. In this case, the standard errors are all the same (SE = 0.04) and the confidence intervals vary no more than ±0.03. The overall conclusion from this analysis is that there are no outliers needing removal or adjustment.

Publication Bias

Publication bias is due to the tendency of journals to publish positive studies and large studies (Thornton & Lee, 2000). Given that published studies tend to report higher effect sizes

than unpublished studies, the potential bias would result in an inflated effect size if the review only included published studies.

The following three tests of publication bias: (1) Funnel Plot, (2) Duval and Tweedie's Trim and Fill and (3) Classic fail-safe N, found no sign of publication bias according to the Comprehensive Meta-analysis™ report.

The funnel plot locates large studies near the top of the graph and near the mean, while smaller studies are located near the bottom and more dispersed given the greater sampling error (CMA report). If there was publication bias the smaller studies at the bottom of the funnel would be concentrated on one-side of the mean, indicating that it likely was their larger effect sizes that resulted in them being published and easier to find. A funnel plot indicating no publication bias would display the effect sizes symmetrically.

The **funnel plot** in Figure 7 shows the calculated effect sizes (by the hollow dots) in relation to the standard error, a representation of sample size. The effect sizes (hollow dots) appear to fall symmetrically on either side of the mean.

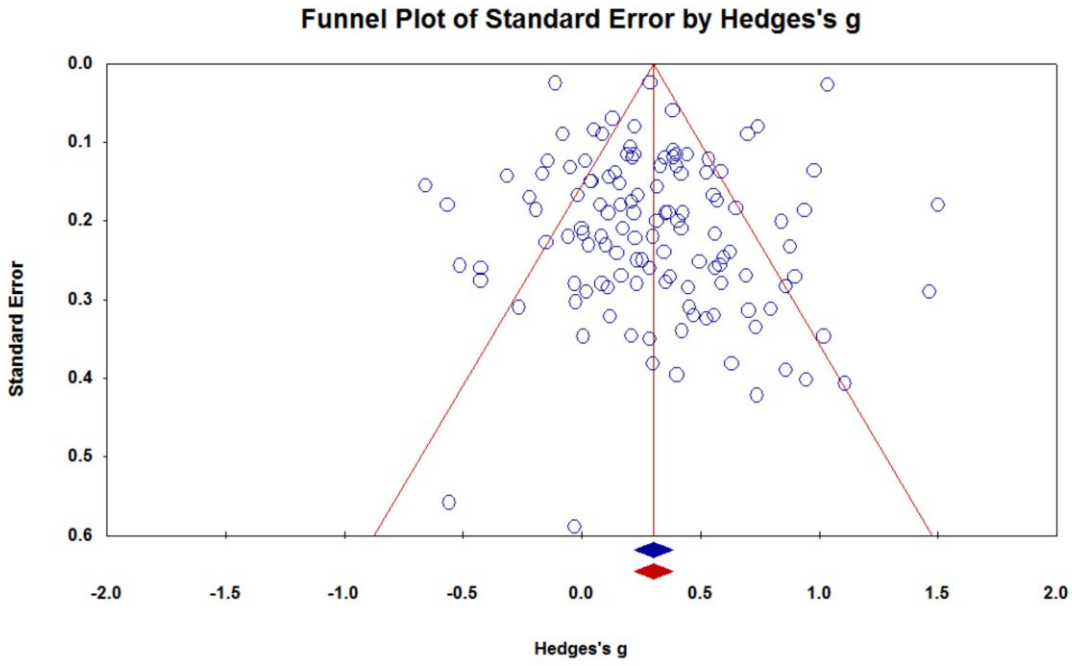


Figure 7. Funnel plot with effect sizes (horizontal axis) and standard errors (vertical axis) for the 125 effect sizes (hollow dots).

Duval and Tweedie's Trim and Fill analysis, as seen in Table 4, indicates that zero studies need to be trimmed in order to make the funnel plot symmetrical, and thereby supports the claim that this meta-analysis has no publication bias.

Table 4

Duval and Tweedie's trim and fill (zero studies trimmed)

		Random Effects			*Q* Value
	Studies Trimmed	Point Estimate	Lower Limit	Upper Limit	
Observed values		0.30	0.22	0.38	1482.32
Adjusted values	0	0.30	0.22	0.38	1482.32

Classic Fail-Safe N, named by Harris Cooper, addresses whether the entire observed effect is a result of bias due to missing studies that were not included (CMA Report). These missing studies would be likely unpublished and still possibly sitting in 'file-drawers'. In this meta-analysis the Classic Fail-Safe N indicates we would need to find 14,716 new studies with null effect sizes to change the result to not statistically significant (2-tailed *p*-value greater than 0.050). This number of 14,716 is far too large to suggest that the observed effect is a result of bias. If the number was small then there might be reason for concern.

Other Forms of Potential Bias. Three other forms of potential bias include research design, method of effect size calculation, and publication source.

For research design, the effect sizes are categorized based on their design (quasi-experimental versus experiments) showed no significant difference, nor did the effect sizes based on calculation format (i.e., whether the effect size was calculated from one final exam, from an average of tests, or a combined form of assessment such as the course grade.

The only evidence of bias can be seen in Table 5-c where the difference in effect size for type of publication was statistically significant ($p < 0.01$). The average effect size for published literature was $\overline{g} = 0.339$ ($k = 98$) while unpublished grey literature such as conference papers and theses/dissertations had an average effect size of $\overline{g} = 0.140$ ($k = 27$), $p < 0.05$. This significant difference is to be expected because it is in line with the theory on which publication bias is based. A meta-analysis would likely have even more publication bias if it only included published studies. Twenty-seven of 125 (or 22%) of the effect sizes in this review are from unpublished sources so again this study does not appear to have biased results. Given that a well-seasoned professional information retrieval specialist was used, it is very likely that all studies in English available from 2000 to 2017 that met the inclusion criteria were retrieved.

Table 5

Other forms of potential bias

Codes	*k*	$\overline{g}$	*SE*	Lower 95th	Upper 95th	*z*-value	*p*-value	*Q*-B	*df*	*p*-value
a) Research design										
QED	119	0.303	0.045	0.214	0.392	6.701	0.000			
RCT	6	0.266	0.157	-0.042	0.574	1.693	0.091			
Total between								0.050	1	0.822
b) Effect Size Calculation										
Average	11	0.416	0.079	0.260	0.572	5.233	0.000			
Combined	11	0.333	0.105	0.126	0.539	3.156	0.002			
One	103	0.276	0.051	0.176	0.375	5.446	0.000			
Total between								2.233	2	0.327
c) Publication Source										
Published	98	0.339	0.048	0.246	0.432	7.142	0.000			
Unpublished	27	0.140	0.047	0.047	0.233	2.941	0.003			
Total between								8.835	1	0.003

Borenstein et al. (2011) suggests that instead of asking if there is *any* bias, we should ask how much impact the bias would have (trivial, modest or substantial). The impact of any bias would be trivial as the funnel plot, and trim and fill indicate no bias. The average effect size of $\overline{g}$ = 0.30 is right on the mark of educationally significant (Lipsey et al., 2012).

Test of Moderator Variables

Demographic study features. The *demographic study features* included four areas: (1) year of publication, (2) educational level (i.e., graduate/ undergraduate), (3) broad subject matter (i.e., STEM/ non-STEM/ health-related) and, (4) detailed subject matter (i.e., science/ technology/ engineering/ math/ non-STEM). See Table 6 for details.

- Year of publication. Even though this synthesis covers 18 years, 108 of the 125 effect sizes (or 86%) came from last three years 2014- 2017. Table 6-a shows the interest in 2014 with 24 effect sizes and that interest peaked in 2016 with 36 effect sizes. There was no significant difference between the time periods. The effect size and the number of effects were as follows: Years 2000 - 2011 ($\overline{g}$ = 0.551, k = 5); 2012 ($\overline{g}$ = 0.435, k = 4); 2013 ($\overline{g}$ = 0.293, k = 8); 2014 ($\overline{g}$ = 0.204, k = 24); 2015 ($\overline{g}$ = 0.260, k = 32); 2016 ($\overline{g}$ = 0.338, k = 36); 2017-18 ($\overline{g}$ 0.283, k = 16) p = .427.
- Educational Level. Graduate courses comprised 10 of 115 (or 9%) of all courses and had a higher average effect size ($\overline{g}$ = +0.46) however it was not significantly different than that of undergraduate courses. See Table 6-b where effect sizes from graduate courses ($\overline{g}$ = 0.46, k = 10) are compared with those of undergraduate courses ($\overline{g}$ =0.286, k = 115, p = .134).
- Broad Subject Matter. The FC appears to be of most interest to instructors of STEM courses as STEM courses account for 61.6% of the calculated effect sizes (i.e., 77 of 125), and health-related subjects that require STEM knowledge such as medicine, pharmacy and nursing accounted for another 16.8% (i.e., 21 of 125) of the effect sizes. There were only 21.6% (i.e. 27 of 125) of the effect sizes that were calculated from non-STEM courses. The impact of the FC was not significantly different between STEM ($\overline{g}$ = 0.273,

k = 77), non-STEM ($\overline{g}$ = 0.277, k = 27), and health-related ($\overline{g}$ = 0.434, k = 21) courses (p = 0.149). See Table 6-c about broad subject matter.

- Specific subject matter. The impact of the FC was not significantly different between Engineering ($\overline{g}$ = 0.210, k = 17); science ($\overline{g}$ = 0.219, k = 30); math $\overline{g}$ = 0.268, k = 23); non-STEM $\overline{g}$ = 0.322, k = 26; health-related (medicine/ nursing/ pharmacy) ($\overline{g}$ = 0.376, k = 24); and technology ($\overline{g}$ = 0.525, k = 5), p = .456. Even though technology had the highest average effect size ($\overline{g}$ = +0.525) there were only five effect sizes so the lack of power from this sample may have caused the no significant difference outcome. See Table 6-d for the range of average effects.

Table 6

Demographic variables

Levels	*k*	$\overline{g}$	*SE*	Lower 95th	Upper 95th	*z*-value	*p*-value	*Q*-Bet.	*df*	*p*-value
a) Year of Publication										
2000-2011	5	0.551	0.137	0.283	0.819	4.024	0.000			
2012	4	0.435	0.216	0.012	0.859	2.015	0.044			
2013	8	0.293	0.080	0.137	0.450	3.666	0.000			
2014	24	0.204	0.072	0.063	0.344	2.838	0.005			
2015	32	0.260	0.064	0.134	0.385	4.048	0.000			
2016	36	0.338	0.104	0.133	0.543	3.237	0.001			
2017-2018	16	0.283	0.066	0.154	0.412	4.300	0.000			
Total between								5.964	6	0.427

b) Educational Level										
Graduate	10	0.460	0.108	0.249	0.671	4.268	0.000			
Under	115	0.286	0.043	0.202	0.370	6.648	0.000			
Total between								2.246	1	0.134
c) Broad Subject Matter (Health-Related vs. Non-STEM vs. STEM)										
Health-R	21	0.434	0.064	0.307	0.560	0.307	0.560			
Non-STEM	27	0.277	0.089	0.103	0.452	0.103	0.452			
STEM	77	0.273	0.060	0.155	0.392	0.155	0.392			
Total between								3.808	2	0.149
d) Specific Subject Matter										
Engineering	17	0.210	0.083	0.047	0.373	2.523	0.012			
Math	23	0.268	0.049	0.172	0.364	5.474	0.000			
Health-R	24	0.376	0.066	0.246	0.506	5.670	0.000			
Non-Stem	26	0.322	0.081	0.164	0.480	3.996	0.000			
Science	30	0.219	0.064	0.094	0.344	3.442	0.001			
Technology	5	0.525	0.315	-0.091	1.142	1.669	0.095			
Total between								4.678	5	0.456

The educationally relevant moderator variables. The *educationally relevant moderator variables* included: (a) instructor (i.e., same/different), (b) semester (same/different), (c) quizzes were used in the FC condition, (d) broad subject matter with and without quizzes, (e) control condition (lecture vs active), and (f) whether the instructor of the FC had prior experience teaching in an active classroom. See Table 7 for details.

a) Instructor. Table 7-a shows there was no significant difference between effect sizes where the control and FC conditions had the same instructor ($\overline{g}$ = 0.259, k = 69) versus different instructor ($\overline{g}$ = 0.302, k = 35) p = .634.

b) Semester. Table 7-b shows there was no significant difference between effect sizes where the control and FC conditions took place during the same semester ($\overline{g}$ = 0.247, k = 58) versus different semesters ($\overline{g}$ = 0.319, k = 60) p = .341.

c) Quizzes. As seen in Table 7-c, including quizzes ($\overline{g}$ = +0.30, k = 78) compared to not including quizzes ($\overline{g}$ = +0.236, k = 26) in the FC condition was not significantly different (p = .478).

d) Subject categories with and without quizzes. Table 7-d shows there was a range of average effects when testing to see if quizzes made a difference when grouped by broad subject matter. The difference in average effect sizes was not significant (p = .058) although it was very close. The subgroup of health-related studies average effects sizes were very close showing little to no difference ($\overline{g}$ = +0.45; $\overline{g}$ = +0.46) when quizzes were used or not used. However, the STEM and nonSTEM subgroups seemed to follow a pattern of their own (i.e., with no quizzes the average effect sizes were low at $\overline{g}$ = + 0.10 and $\overline{g}$ = + 0.14, but with quizzes the average effect sizes for both STEM and nonSTEM were much higher at $\overline{g}$ = + 0.30 and reflective of the overall average effect size). The vast majority of effect sizes did use quizzes in the FC condition yet there were enough that did not, to show this pattern.

e) Control condition. Table 7-e shows that when the control condition was all lecture ($\overline{g}$ = +0.267, k = 78) or it included some active learning ($\overline{g}$ = +0.348, k = 39) the difference was not significant (p = .336).

f) FC Instructor prior experience teaching actively. Not all studies reported whether the teacher had experience teaching actively but for those that did Table 7-f shows the difference was not significant (i.e., not experienced ($\overline{g}$ = 0.289, k = 18) versus yes experienced ($\overline{g}$ = +0.404, k = 12) p = .374.)

Table 7

Educationally relevant moderator variables

Levels	*k*	$\overline{g}$	*SE*	Lower 95th	Upper 95th	*z*-value	*p*-value	*Q*-Bet.	*df*	*p*-value
a) Instructor (different/same)										
Different	35	0.302	0.058	0.189	0.415	5.235	0.000			
Same	69	0.259	0.068	0.127	0.392	3.830	0.000			
Total between								0.226	1	0.634
b) Semester (different/same)										
Different	60	0.319	0.058	0.205	0.433	5.504	0.000			
Same	58	0.247	0.049	0.150	0.343	5.017	0.000			
Total between								0.905	1	0.341
c) Quiz in FC										
No	26	0.236	0.071	0.098	0.375	3.353	0.001			
Yes	78	0.300	0.055	0.193	0.407	5.477	0.000			
Total between								0.503	1	0.478
d) Subject Categories with and without quizzes										
Health-R/No	5	0.445	0.147	0.156	0.733	3.017	0.003			
Health-R/Yes	12	0.460	0.095	0.273	0.646	4.817	0.000			
N-Stem/No	6	0.101	0.184	-0.258	0.461	0.553	0.580			
N-Stem/Yes	14	0.297	0.152	-0.001	0.594	1.955	0.051			
STEM/No	14	0.141	0.061	0.022	0.259	2.323	0.020			
STEM/Yes	52	0.278	0.078	0.125	0.432	3.551	0.000			

Total between								10.678	5	0.058
e) Control Condition (lecture vs. active)										
Active	39	0.348	0.066	0.219	0.477	5.295	0.000			
Lecture	78	0.267	0.052	0.165	0.369	5.139	0.000			
Total between								0.927	1	0.336
f) FC Instructor Prior Experience										
No	18	0.289	0.080	0.133	0.446	3.624	0.000			
Yes	12	0.404	0.102	0.205	0.604	3.972	0.000			
Total between								0.789	1	0.374

This chapter reported on the results of the overall average effect size, the publication bias indicators from CMA (Comprehensive Meta-analysis) and the moderator variables. The discussion of these findings follow in the next chapter.

CHAPTER 4: DISCUSSION

The main question addressed in this review consists of the impact of the flipped classroom (FC) on achievement in higher education as compared to the traditional classroom instruction (CI). This chapter is organized in three sections. The first section interprets the major findings including the overall effect size and the moderator variables. The second section addresses the generalizability of the conclusions and general limitations. The third section covers the implications for theory, policy, and practice along with suggestions for future research.

Major Findings and Interpretation

Overall effect sizes. The overall weighted average random-effects of the FC over the CI is $\overline{g}$ = +0.30, k = 125, p < .01. The random-effects model is the appropriate model to use with social science studies, such as these in education, where there is variation across a number of experimental conditions, including the treatment, unlike many pharmaceutical trials for example (Borenstein et al., 2011) where there is a more standardized treatment and random assignment of participants to groups.

Interpretation of effect size. According to Cohen (1988) this effect size of $\overline{g}$ = 0.30 would be considered small as it is between 0.20 and 0.50. However, it may be better to compare this effect size with others in education to attain a relative comparison (Lipsey et al., 2012) where 0.30 is considered the threshold beyond which is considered educationally significant. As this is one of the first extensive meta-analysis comparing the FC with CI, the most appropriate comparison effect sizes would be the four meta-analyses that compared BL in general with CI: Means et al. (2013) who found an average effect size of $\overline{g}$ = +0.35 (k = 23), Bernard et al. (2014) who found an average effect size of $\overline{g}$ = +0.33 (k = 117), Spanjers et al. (2015) who found an average effect size of $\overline{g}$ = +0.34 (k = 24), and Vo et al. (2017) who found an average effect size

of $\overline{g}$ = +0.39 (k = 51). As such, when comparing this study's effect size between the FC versus traditional CI (+0.30), and with previous BL versus CI studies, there is almost no difference, as seen in Table 8.

Table 8

Meta-analyses comparing Blended Learning/Flipped Classroom versus Classroom Instruction

Authors of Meta-analyses	Comparison	Average ES	k (# of ESs)
Means et al. (2013)	BL vs. CI	+0.35	23
Bernard et al. (2014)	BL vs. CI	+0.33	117
Spanjers et al. (2015)	BL vs. CI	+0.34	24
Vo et al. (2017)	BL vs. CI	+0.385	51
Sparkes (2019)	FC vs. CI	+0.30	125

Notwithstanding the overall effect size, it is important to note that the average effect size is highly heterogeneous (Q-value 1487.391, $p < .001$). As such, it is important to be careful generalizing the effect to the general population. Some studies had positive impact while other studies showed negative effects. The heterogeneity analysis thus suggests the need to examine moderator variables to explain some of the variation in effect sizes.

Moderator variables findings. Estimating the overall average effect size is only one objective of the meta-analysis. In order to more thoroughly answer the research questions, the focus now turns to the moderator variables in an attempt to understand the variability of the effect sizes and explain why the average effect size for FC was significantly larger than CI (i.e., what are the important factors that made the difference). Returning to the previous literature, of particular interest were the two moderator variables found significant in the BL meta-analyses

(i.e., subject matter of STEM versus non-STEM (Vo et al., 2017) and whether quizzes were included or not (Spanjers et al., 2015) as seen in Figure 5. A discussion follows on possible reasons why this meta-analysis did not find these study features to be significant in the FC studies.

Subject Matter. The first moderator variable is subject matter. The FC is more commonly researched in STEM courses, so naturally there is interest in whether the FC is more effective in STEM areas. From this meta-analysis there was no significant difference found when comparing the effect of the FC in STEM courses versus non-STEM courses. This finding is different from that of Vo, Zhu and Diep (2017) who found that BL improved learning more in STEM courses than non-STEM courses. There were two choices that Vo et al. made that were different from the current study. The first choice was how to categorize studies as STEM or non-STEM. In particular, for health-related courses Vo et al. did not appear to have an obvious rationale for categorizing a course as STEM or non-STEM (e.g., nursing (electrocardiography) was coded as non-STEM, while a General Health course was coded as STEM. The second choice was about which data to use to calculate an effect size (e.g., a study coded as STEM that used grading criteria in the BL condition different from the CI condition yet Vo et al. used the course grade as the data for effect size calculation resulting in a large effect size of +2.87). Because the Vo et al. meta-analysis was based on BL it included some FC studies. This common study, when coded in this FC meta-analysis, was coded as non-STEM and the effect size used a common assessment to both groups resulting in a lower effect size of +0.65. Because of the difficulty of categorizing health-related discipline courses into either STEM or non-STEM, this meta-analysis created a third category called “health-related” to include courses specific to nursing, pharmacy, and medicine. This third category was found to be useful in the second moderator variable analysis discussed next.

Quizzes. The second moderator is quizzes. For coding purposes the FC quizzes only included graded quizzes, but not formative quizzes, worksheets or questions embedded in video. Spanjers et al. (2015, p. 68) found quizzes to be a significant moderator of the BL success. The FC courses with quizzes did not significantly outperform the FC course comparisons without quizzes. However, when the effect sizes were sub-categorized into STEM, non-STEM, and health-related, then the STEM and non-STEM categories showed noticeably higher effect sizes for those using quizzes, and lower effect sizes for those not using quizzes. The health-related subgroup showed little difference between conditions using quizzes and those not using quizzes.

Generalizability of the Conclusions

When addressing the generalizability of the conclusions the following are considered: the participants, variations in the predictor and outcome variables, and research designs.

Participants. The majority of the population was comprised of English speaking, undergraduate students in higher education studying STEM subjects and so the outcome is generalizable to this group. Ninety two percent of the effect sizes were from studies of undergraduate students. The majority of effect sizes (62%) were from studies based on STEM courses, followed by non-STEM courses (21%), and then by health-related courses (17%). One can thus conclude that, overall, the FC model will result in better results than CI, while noting the heterogeneous variability indicated above.

Variations in predictor variables (FC) and outcome variables. The FC is comprised of active learning in the classroom and video lectures to be watched at home. As simple as this sounds there was a large range of variations in the types and amounts of active learning and the type and amount of video lectures. Some studies included active learning in the form of students working individually on their homework with the instructor available for questions similar to

office hours. However, other studies included active learning in the first minutes of class in the form of short quizzes about the material covered in the video lectures, followed by structured collaborative and cooperative activities, including discussions, problem solving, debates, presentations, and role plays. Video lectures ranged in time from short video clips to full hour-long lectures, and in content from the instructor lecturing about the topic to third party material professionally produced. The benefit of this range in the predictor variables is that the results of this meta-analysis are representative of many variations of FC.

Achievement was represented in a number of forms: the final exam score, the average of tests, and a combination of scores to form a course mark. Over eighty-two percent of the effect sizes were based on the standard representation of achievement, (i.e., one exam score). If only studies that recorded final exam scores were included, there would be fewer effect sizes and less information about study features in this meta-analysis. Statistically there was no difference between the average effect sizes among these different forms of achievement. As equivalent measures were always used to calculate the effect size, the results were representative of effect sizes using a final exam. As a result, this meta-analysis benefited from the additional effect sizes being calculated and study features being coded.

Research designs. Research designs included experimental (i.e., random assignment of participants to control and treatment groups), and quasi-experimental (i.e., assignment of equivalent groups to the control and treatment group). The two-group pre-experimental group design studies were excluded, as there were no bases to determine if the comparison group was equivalent to the treatment group at the beginning (Shadish, Cook & Campbell, 2002).

Within these research designs, variations of the control and treatment were coded, such as the timing of the semester (concurrent or consecutive) and the consistency of the instructor (same

or different). Forty-six percent of studies used concurrent semesters for the control and treatment groups, while the remaining used consecutive semesters. Ideally studies would use concurrent semesters, however excluding studies that used consecutive semesters would greatly reduce the number of studies included. As there was no statistical difference found between the average effect size of these studies, the results were representative of comparisons in either concurrent or consecutive semesters. Similarly, fifty-five percent of the comparison studies used the same instructor while the remaining studies used different instructors to teach the treatment and control groups. Again, while using the same instructor reduces the extraneous variables that come with different instructors, there was no significant difference between the effect size of these groups. Including studies that use different instructors for the control and treatment conditions benefits the meta-analysis by increasing the number of studies included and providing results that are representative of both situations.

The research designs of the studies used in this meta-analysis asked the same question that the meta-analysis was trying to answer (i.e., Is the FC at least as effective as CI in regards to achievement in higher education?). The answer is yes, the FC is even more effective that CI on average. Therefore, the research designs were in alignment with the meta-analysis purpose and support the findings' generalizability.

Implications for Theory and Practice

Theoretical Significance. Researchers may be interested in the results of this study as it helps to confirm the effectiveness of the FC in comparison with traditional CI lecture in higher education. This is the first comprehensive meta-analysis on the FC focusing on postsecondary education capturing studies from 2000-2017, and it showed the FC significantly outperformed CI with an overall effect size of +0.30.

With the focus on improving student success in STEM education to improve scientific innovation and economic growth (DeCoito, 2016), there is interest in whether there are instructional approaches that are more effective in STEM disciplines. While Vo et al. (2017) found the effect of BL to be significantly greater with STEM courses, this meta-analysis found there to be no significant difference between the impact of the FC on STEM and non-STEM average effect sizes. This meta-analysis found that how disciplines are categorized into STEM or non-STEM is rarely discussed and not universally agreed upon, especially when it comes to professional health-related disciplines. This study recommends creating a new category called *health-related* to absorb nursing, pharmacy, and medicine specific courses. The moderator variable of subject was expanded from STEM versus non-STEM during the preliminary coding process because some studies did not fit into the dichotomous options. Although nursing, medicine, and pharmacy seemed to require STEM subjects overall, some courses were not clearly STEM related or non-STEM. The idea of a new *health-related* category emerged to absorb the studies that were specific to nursing, pharmacy, and medicine. When analyzing the moderator variable of quizzes versus no quizzes a different pattern of results emerged when they were further categorized by STEM, non-STEM, and health-related. The effect sizes for studies with quizzes in STEM and non-STEM courses were higher than those without, however there was no difference between quizzes and no quizzes in the health-related category. While not statistically significant at $p = 0.058$ it was approaching significance enough to mention that health-related studies may act differently than STEM or non-STEM studies. To some degree these seemingly different moderator variables appear to be confounded. Perhaps in the health-related professional disciplines of nursing, medicine, and pharmacy there is already ample opportunity to put the new knowledge in practice so that quizzes are not needed as the driving force to ensure the material from the video lectures is understood.

As noted above, the results of this extensive FC meta-analysis showed the impact of the FC on achievement in higher education was comparable to that of BL in general. However, given that FC is a manifestation of BL, this supports the theoretical alignment of these approaches as presented in Figure 5.

In summary, this meta-analysis added not only to the FC literature but to a better understanding of the different impact of FC quizzes on achievement in STEM, non-STEM, and health-related disciplines.

Practical Significance. Practitioners such as university or college faculty and possibly secondary teachers may look to this study to help make more informed decisions when implementing BL and in particular the FC model. This study would reassure instructors and administrators that the FC had a small but practically significant (Lipsey et al., 2012) difference in improving achievement compared to traditional teaching ($\overline{g}$ = +0.30, k = 125, $p < 0.01$). With this kind of evidence supporting the FC, university administrators could use the results of this meta-analysis to help justify extra financial support, time required by instructors for development, technical support, and the pedagogical support of an instructional designer to guide the design. The results of the FC meta-analysis were not different than those from BL in general though. The FC is just one way to structure BL.

Limitations

Even though meta-analysis surpasses qualitative reviews and vote counts as a form of analysis because it is based on a systematic review process and uses statistical approaches to calculate magnitude of effect sizes and direction, there is still potential for bias. Where there is bias there may be misrepresented results. The limitations of this meta-analysis are reflected in the limitations of the data collected and how it is treated.

First it is important to note that the research question itself provides a limitation on the results possible from this study. The comparison of FC to CI provides an average overall effect size that then can be further analyzed according to moderator factors to possibly find more meaning. However, the question that compares FC to CI will not be able to answer finer instructional design questions. See the *suggestions for future research* section for more about how the research question could be improved.

Secondly, language is a limitation because the literature search was conducted only on studies or abstracts available in English. Any research available in other languages that might have contributed to the findings was excluded. Future analyses should incorporate additional research in this regard.

A third limitation is that there was only one main coder throughout the meta-analysis. Coding requires judgement calls and for that reason ideally two independent coders code everything and compare results. In this meta-analysis twenty-five percent of all search results were reviewed independently by two coders, and discrepancies were discussed until there was consensus. This consensus was negotiated to ensure consistency of coding for the remaining seventy-five percent of the results that were coded by one person, the author. Interrater reliability statistics were calculated as well to ensure a solid basic understanding and agreement. This approach was used again when reviewing the full-text of the studies, when calculating the effect sizes, and when coding the study features. The funnel plot along with other tests for publication bias, indicated there was no publication bias.

A fourth limitation is that all the variations of active learning used were not coded. Sometimes the lists were extensive and due to limited resources was put aside. Instead, levels of active learning within the control condition (some or none) became a reasonable variable to code, along with whether there were different instructors, and if the study took place over different

semesters. Sometimes these variations are specified and sometimes these details are missing. Meta-analyses in the social sciences are always prone to the realities of variability in treatment implementation, and/or incomplete reporting. That said, the very large sample obtained in the present study mitigates drawing misleading conclusions.

A fifth limitation is that the study cannot comment on how students received the FC. If students found the FC too difficult or too much work and decided to switch sections or drop the course, this information would not be detected by comparing the final exam marks of the FC with those of CI as the marks would only reflect the achievement of students who did finish the course. Students drop classes or change sections for any number of reasons not necessarily related to whether the students felt they would fail. Reporting on student perceptions is the scope of a different study than this one.

A sixth limitation comes from missing data in the primary studies. The researcher of the primary study is often the instructor in the treatment condition and may be invested in having the treatment work. Some studies described the extra steps needed to blind the student names from assessments, and/or have another instructor independently mark the exams to ensure fair marking. However, this information was not discussed in other studies. Whether the researcher was the instructor of the treatment group and whether the assessments were blinded before grading was not coded in this study. The effect size does not always show the full picture of possible bias when implementing the FC.

Suggestions for Future Research

Now that we know that on average the FC significantly outperforms CI to the same degree that BL outperforms CI, the field will benefit less from studies that compare FC with CI. What is needed now, in the opinion of this researcher, are comparisons between FC treatments or

between BL treatments, so that we can determine the impact of how instructional approaches differ between them. Within the FC model there are a range of active-learning approaches that can be used to varying degrees, with and without collaboration, as well as a range of video lecture styles with and without interaction, for varying durations, and purposes. Studies that compare FC to FC could help shed light on the sorts of interaction treatments and the conditions that would result in improved achievement. The heterogeneity of FC studies' outcomes suggests that there may be design and/or contextual issues that yield positive, neutral or even negative effects. More nuanced studies that examine the design and processes underlying FC environments will help inform better ways of utilizing this promising strategy.

Finally, for years DE, OL, and BL versus CI comparison studies have been conducted and many meta-analyses followed as seen in Table 1 in the literature review. In the move away from "all versus none" meta-analyses, when there were enough DE versus DE studies, the Bernard et al. (2009) meta-analysis built on Moore (1989), and Anderson's (2003) work by comparing different types of interaction treatments (i.e., instructional and/or media conditions designed to facilitate student-student, student-content, and student-teacher interactions) under various conditions (e.g., synchronous or asynchronous) to determine correlations with improved achievement. Schmid, Bernard, Borokhovski, Tamim, Abrami, Surkes, Wade, and Woods (2014) also used this approach to examine technology integration in both the control and treatment conditions to address questions about the impact different levels of technology use, and the purposes of their use (e.g., cognitive support versus presentation support). While Bernard, Borokhovski, Schmid, and Tamim, (2018) also encouraged moving from the "all versus none" to "all versus some" comparisons to answering questions about effective instructional design practices, they recognized that more time and effort are required for this type of coding. Researchers would need to move from what Cooper (2017) called low-inference coding (e.g., FC

versus CI) to high-inference coding (e.g., more active learning versus less active learning). Nonetheless, the time has come to take up this challenge in order to improve the instructional design of the future.

Recognizing that a meta-analyst is a "prisoner of the existing literature," meaning that there are limits on how much a reviewer can do to disentangle complex questions if they are not addressed by primary researchers, there is a need for these researchers to ask more subtle questions. In the absence of a new approach to experimentation, it is possible that we will not get to the core questions that teachers and instructional designers need to improve the conditions under which FC, or for that matter any other new instructional approach, work best.

Overall Summary

In conclusion, according to this study, the FC is significantly more effective (statistically and practically) in improving student achievement than the traditional CI in higher education. No study features were found to significantly moderate this effect. The FC version of BL outperformed the traditional CI to the same degree that BL did. Even though the FC included pedagogical guidance of active learning in the classroom and video lectures to watch at home, it was not more effective than the results of four meta-analyses on BL in general without specific pedagogical guidance. Meta-analyses whose questions involve differentiating between the impacts on STEM versus non-STEM courses should include a Health-related category for courses specific to nursing, pharmacy, and medicine or clearly articulate how these are being coded. Given the meta-analysis is clear that the FC is more effective than CI, it is time to stop this line of inquiry and pursue comparisons that involve variations of the FC in the treatment and the control in order to determine underlying attributes that can improve the instructional design of future courses.

REFERENCES

Starred references (*) are studies in the meta-analysis

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

*Adams, C., & Dove, A. (2016). Flipping calculus: The potential influence, and the lessons learned. *Electronic Journal of Mathematics & Technology, 10*(3), 154-164.

*Albert, M., & Beatty, B. J. (2014). Flipping the classroom applications to curriculum redesign for an introduction to management course: Impact on grades. *Journal of Education for Business, 89*, 419-424. http://dx.doi.org/10.1080/08832323.2014.929559

*AlJaser, A. M. (2017). Effectiveness of using flipped classroom strategy in academic achievement and self-efficacy among education students of Princess Nourah Bint Abdulrahman University. *English Language Teaching, 10*(4), 67-77.

Allen, M., Mabry, E., Mattrey, M., Bourhis, J., Titsworth, S., & Burrell, N. (2004). Evaluating the effectiveness of distance learning: A comparison using meta-analysis. *Journal of communication*, *54*(3), 402-420.

Allen, I. E., & Seaman, J. (2013). *Changing course: Ten years of tracking online education in the United States*. Sloan Consortium. PO Box 1238, Newburyport, MA 01950.

Anderson, L. W., & Krathwohl, D. (Eds.) (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman.

Anderson, T. (2003). Getting the mix right again: An updated and theoretical rationale for interaction. *The International Review of Research in Open and Distributed Learning*, *4*(2).

*Anderson Jr, H. G., Frazier, L., Anderson, S. L., Stanton, R., Gillette, C., Broedel-Zaugg, K., & Yingling, K. (2017). Comparison of pharmaceutical calculations learning outcomes achieved within a traditional lecture or flipped classroom andragogy. *American Journal of Pharmaceutical Education, 81*(4), 1-9. http://dx.doi.org/10.5688/ajpe81470

*Asarta, C. J., & Schmidt, J. R. (2017). Comparing student performance in blended and traditional courses: Does prior academic achievement matter? *Internet & Higher Education, 32*, 29-38. http://dx.doi.org/10.1016/j.iheduc.2016.08.002

*Asiksoy, G., & Özdamlı, F. (2016). Flipped classroom adapted to the ARCS model of motivation and applied to a physics course. *EURASIA Journal of Mathematics, Science & Technology Education, 12*(6), 1589-1603. http://dx.doi.org/10.12973/eurasia.2016.1251a

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

*Bagley, S. F. (2014). *Improving student success in calculus: A comparison of four college calculus classes* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3632076)

Bates, A. T. (2015). Teaching in a digital age: Guidelines for designing teaching and learning. Creative Commons Attribution-NonCommercial 4.0 International License, except where otherwise noted. Printed by SFU Document Solutions, Simon Fraser University, Burnaby, BC.

*Baytiyeh, H., & Naja, M. K. (2017). Students' perceptions of the flipped classroom model in an engineering course: A case study. European Journal of Engineering Education, 42(6), 1048-1061. http://dx.doi.org/10.1080/03043797.2016.1252905

Bergmann, J., & Sams, A. (2008). Remixing chemistry class. *Learning and Leading with Technology*, *36*(4), 24-27.

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

Bernard, R. M. (2017, April). *What meta-analyses say about the effectiveness of distance education, online learning and blended learning.* Invited presentation and panel discussion at the Online Learning Consortium (OLC) conference. New Orleans. LA.

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research*, *79*(3), 1243-1289.

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Wallet, P. A., Fiset, M. & Huang, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. *Review of educational research*, *74*(3), 379-439.

Bernard, R. M., Borokhovski, E., Schmid, R. F. Tamim, R. M. (2018). Gauging the effectiveness of educational technology integration in education: What the best-quality meta-analyses tell us. Manuscript submitted for publication.

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.

Betihavas, V., Bridgman, H., Kornhaber, R., & Cross, M., (2016). The evidence for 'flipping out': A systematic review of the flipped classroom in nursing education. Nurse Educ. Today 38, 15–21.

*Bishop, J. L. (2013). *A controlled study of the flipped classroom with numerical methods for engineers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3606852)

Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. In *ASEE National Conference Proceedings, Atlanta, GA*.

Bloom, B. S. (1968). Learning for mastery. Evaluation Comment (UCLA-CSIEP), (2), 1–12.

Bloom, B. S. (Ed.) (1956). Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York: Longman.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2014). Comprehensive Meta-Analysis (version 3.3.070). Biostat, Englewood, NJ.

*Bossaer, J. B., Panus, P., Stewart, D. W., Hagemeier, N. E., & George, J. (2016). Student performance in a pharmacotherapy oncology module before and after flipping the classroom. *American Journal of Pharmaceutical Education, 80*(2), 1-6. http://dx.doi.org/10.5688/ajpe80231

*Braun, I., Ritter, S., & Vasko, M. (2014). Inverted classroom by topic - A study in mathematics for electrical engineering students. *International Journal of Engineering Pedagogy, 4*(3), 11-17. http://dx.doi.org/10.3991/ijep.v4i3.3299

*Brooks, A. W. (2014). Information literacy and the flipped classroom: Examining the impact of a one-shot flipped class on student learning and perceptions. *Communications in Information Literacy, 8*(2), 225-235.

*Butzler, K. B. (2015). *The effects of motivation on achievement and satisfaction in a flipped classroom learning environment* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3637765)

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cavanaugh, C. S. (2001). The effectiveness of interactive distance education technologies in K–12 learning: A meta-analysis. International Journal of Educational Telecommunications, 7, 73–88.

Cavanaugh, C., Gillan, K. J., Kromrey, J., Hess, M., & Blomeyer, R. (2004). The effects of distance education on K-12 student outcomes: A meta-analysis. *Learning Point Associates/North Central Regional Educational Laboratory (NCREL)*.

*Caviglia-Harris, J. (2016). Flipping the undergraduate economics classroom: Using online videos to enhance teaching and learning. *Southern Economic Journal, 83*(1), 321-331. http://dx.doi.org/10.1002/soej.12128

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2018). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational*

Technology Research and Development, 1-32. https://doi.org/10.1007/s11423-018-9633-7

*Cheng, X., Ka Ho Lee, K., Chang, E. Y., & Yang, X. (2017). The "flipped classroom" approach: Stimulating positive learning attitudes and improving mastery of histology among medical students. *Anatomical Sciences Education, 10*(4), 317-327. http://dx.doi.org/10.1002/ase.1664

*Chiang, Y., & Wang, H. (2015). Effects of the in-flipped classroom on the learning environment of database engineering. *International Journal of Engineering Education, 31*, 454-460.

*Choi, J., & Lee, Y. (2018). To what extent does 'flipping' make lessons effective in a multimedia production class. *Innovations in Education and Teaching International, 55*(1), 3-12. http://dx.doi.org/10.1080/14703297.2015.1123105

Çırak Kurt, S., Yıldırım, İ., & Cücük, E. (2018). The effects of blended learning on student achievement: A meta-analysis study. *Hacettepe University Journal of Education*, 33(3), 776-802. doi: 10.16986/HUJE.2017034685

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

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008). Internet-based learning in the health professions: a meta-analysis. *Jama*, *300*(10), 1181-1196.

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed., Vol. 2). Sage Publications.

*Cotta, K. I., Shah, S., Almgren, M. M., Macías-Moriarity, L. Z., & Mody, V. (2016). Effectiveness of flipped classroom instructional model in teaching pharmaceutical

calculations. *Currents in Pharmacy Teaching & Learning, 8*(5), 646-653. http://dx.doi.org/10.1016/j.cptl.2016.06.011

*Cruzado, I., & Roman, E. M. (2015). Inverted classroom and its influence on students' attitudes across learning styles. *Transportation Research Record, 2480*(1), 38-44. http://dx.doi.org/10.3141/2480-05

*Davies, R., Dean, D., & Ball, N. (2013). Flipping the classroom and instructional technology integration in a college-level information systems spreadsheet course. *Educational Technology Research & Development, 61*, 563-580. http://dx.doi.org/10.1007/s11423-013-9305-6

*Day, L. J. (2018). A gross anatomy flipped classroom effects performance, retention, and higher-level thinking in lower performing students. *Anatomical Sciences Education*, Advance online publication. http://dx.doi.org/10.1002/ase.1772

DeCoito, I. (2016). STEM education in Canada: A knowledge synthesis, Canadian Journal of Science, Mathematics and Technology Education, 16(2) 114-128, DOI: 10.1080/14926156.2016.1166297

Dewey, J., & Jackson, P. W. (1990). *The school and society and the child and the curriculum*. A Centennial Publication.

Driscoll, M. P. (2005). *Psychology of learning for instruction: Pearson new international edition*. Pearson Higher Ed.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463.

*Eichler, J. F., & Peeples, J. (2016). Flipped classroom modules for large enrollment general chemistry courses: A low barrier approach to increase active learning and improve student grades. *Chemistry Education Research and Practice, 17*(1), 197-208.

*Faretta, R. S. (2016). *A causal-comparative inquiry into the significance of implementing a flipped classroom strategy in nursing education* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10107409)

*Files, D. D. (2017). *Instructional approach and mathematics achievement: An investigation of traditional, online, and flipped classrooms in college algebra* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10100739)

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

*Fraga, L. M., & Harmon, J. (2014). The flipped classroom model of learning in higher education: An investigation of preservice teachers' perspectives and achievement. *Journal of Digital Learning in Teacher Education, 31*(1), 18-27. http://dx.doi.org/10.1080/21532974.2014.967420

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415.

Garnham, C., & Kaleta, R. (2002). Introduction to hybrid courses. *Teaching with Technology Today, 8*(6). Retrieved from http://www.uwsa.edu/ttt/

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, *7*, 95–105. doi: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, CA: Jossey-Bass.

*Geist, M. J., Larimore, D., Rawiszer, H., & Al Sager, A. W. (2015). Flipped versus traditional instruction and achievement in a baccalaureate nursing pharmacology course. *Nursing Education Perspectives, 36*, 114-115. http://dx.doi.org/10.5480/13-1292

Gillette, C., Rudolph, M., Kimble, C., Rockich-Winston, N., Smith, L., & Broedel-Zaugg, K. (2018). A meta-analysis of outcomes comparing flipped classroom and lecture. *American Journal of Pharmaceutical Education*. 82 (5), Article 6898. https://doi.org/10.5688/ajpe6898

*Gillispie, V. (2016). Using the flipped classroom to bridge the gap to Generation Y. *The Ochsner Journal, 16*(1), 32-36.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research 1. *Educational researcher*, *5*(10), 3-8.

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, CA: Pfeiffer.

*Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., & Graham, M. J. (2015). Increased preclass preparation underlies student outcome improvement in the flipped classroom. *CBE Life Sciences Education, 14*(4). Retrieved from http://dx.doi.org/10.1187/cbe.15-02-0040

*Guerrero, S., Beal, M., Lamb, C., Sonderegger, D., & Baumgartel, D. (2015). Flipping undergraduate finite mathematics: Findings and implications. *PRIMUS, 25*, 814-832. http://dx.doi.org/10.1080/10511970.2015.1046003

*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). Retrieved from http://www.amstat.org/publications/jse/v23n1/gundlach.pdf

Guskey, T. R. (1987). Rethinking mastery learning reconsidered. *Review of Educational Research*, *57*(2), 225-229.

Guskey, T. R. (2007). Closing the achievement gap: Revisiting Benjamin S. Bloom's "learning for mastery." *Journal of Advanced Academics, 19*, 8–31.

Halili, S. H., & Zainuddin, Z. (2015). Flipping the classroom: What we know and what we don't. *The Online Journal of Distance Education and e-Learning*, *3*(1), 28-35.

*Harrington, S. A., Vanden Bosch, M., Schoofs, N., Beel-Bates, C., & Anderson, K. (2015). Quantitative outcomes for nursing students in a flipped classroom. *Nursing Education Perspectives, 36*, 179-181. http://dx.doi.org/10.5480/13-1255

*Harvey, S. (2014). The "flipped" Latin classroom: A case study. *Classical World, 108*(1). Retrieved from http://muse.jhu.edu/article/562560

*Haughton, J., & Kelly, A. (2015). Student performance in an introductory business statistics course: Does delivery mode matter? *Journal of Education for Business, 90*, 31-43. http://dx.doi.org/10.1080/08832323.2014.968518

*He, W., Holton, A., Farkas, G., & Warschauer, M. (2016). The effects of flipped instruction on out-of-class study time, exam performance, and student perceptions. *Learning and Instruction, 45*, 61-71. http://dx.doi.org/10.1016/j.learninstruc.2016.07.001

* Kim, H., & Jang, Y. (2017). Flipped learning with simulation in undergraduate nursing education. Journal of Nursing Education, 56(6), 329-336. http://dx.doi.org/10.3928/01484834-20170518-03

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

*Heyborne, W. H., & Perrett, J. J. (2016). To flip or not to flip? Analysis of a flipped classroom pedagogy in a general biology course. *Journal of College Science Teaching, 45*(4), http://dx.doi.org/10.2505/4/jcst16_045_04_31

*Horton, D., Craig, M., Campbell, J., Gries, P., & Zingaro, D. (2014). Comparing outcomes in inverted and traditional CS1. In *Proceedings of the 2014 conference on Innovation & technology in computer science education* (pp. 261-266). New York, NY: ACM. Retrieved from http://dx.doi.org/10.1145/2591708.2591752

*Hotle, S. L., & Garrow, L. A. (2016). Effects of the traditional and flipped classrooms on undergraduate student opinions and success. *Journal of Professional Issues in Engineering Education & Practice, 142*(1), 1-11. http://dx.doi.org/10.1061/(ASCE)EI.1943-5541.0000259

Hu, R., Gao, H., Ye, Y., Ni, Z., Jiang, N., & Jiang, X. (2017). Effectiveness of flipped classrooms in Chinese baccalaureate nursing education: A meta-analysis of randomized controlled trials. *International Journal Of Nursing Studies, 79*, 94-103. http://dx.doi.org/10.1016/j.ijnurstu.2017.11.012

*Hu, Y., Montefort, J. M., & Tsang, E. (2017, June). An analysis of factors affecting student performance in a statics course. In ASEE Annual Conference and Exposition, Conference Proceedings (Vol. 2017).

*Hudson, D. L., Whisenhunt, B. L., Shoptaugh, C. F., Visio, M. E., Cathey, C., & Rost, A. D. (2015). Change takes time: Understanding and responding to culture change in course redesign. *Scholarship of Teaching and Learning in Psychology, 1*, 255-268. http://dx.doi.org/10.1037/stl0000043

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

*Hussey, H. D. (2014). Promoting active learning through a flipped course design. In J. Keengwe, G. Onchwari & J. Oigara (Eds.), *Promoting Active Learning through the Flipped Classroom Model* (pp. 23-46). Hershey, PA: IGI Global. Retrieved from http://dx.doi.org/10.4018/978-1-4666-4987-3.ch002

Jahng, N., Krug, D., & Zhang, Z. (2007). Student achievement in online distance education compared to face-to-face education. *European Journal of Open, Distance and E-Learning*, *10*(1).

*Jeavons, T., Flecknoe, S., Davies, A.N. & White, G. (2013). 'Lecture-flip' pedagogy in bioscience education. In J. Herrington, A. Couros & V. Irvine (Eds.), *Proceedings of EdMedia: World Conference on Educational Media and Technology 2013* (pp. 2082-2087). Waynesville, NC: Association for the Advancement of Computing in Education. Retrieved from http://www.editlib.org/p/112261/

*Jensen, J. L., Kummer, T. A., & Godoy, P. D. d. M. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE Life Sciences Education, 14*(1). Retrieved from http://dx.doi.org/10.1187/cbe.14-08-0129

Johnson, D. W., Johnson, R. T., & Smith, K. A. (2014). Cooperative learning: Improving university instruction by basing practice on validated theory. *Journal on Excellence in College Teaching, 25*(3&4), 85-118.

*Kang, N. (2015). The comparison between regular and flipped classrooms for EFL Korean adult learners. *Multimedia-Assisted Language Learning, 18*(3), 41-72. http://dx.doi.org/10.15702/mall.2015.18.3.41

Karabulut-Ilgu, A., Jaramillo Cherrez, N., & Jahren, C. T. (2018). A systematic review of research on the flipped learning method in engineering education. British Journal of Educational Technology, 49(3), 398-411.

*Karaca, C., & Ocak, M. A. (2017). Effects of flipped learning on university students' academic achievement in algorithms and programming education. *International Online Journal of Educational Sciences, 9*(2), 527-543. http://dx.doi.org/10.15345/iojes.2017.02.017

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A-M. K., Hammerstrøm, K. T., & Sathe, N. (2016). *Searching for studies: A guide to information retrieval for Campbell Systematic Reviews*. Retrieved from Campbell Collaboration website: https://www.campbellcollaboration.org/images/Campbell_Methods_Guides_Information_Retrieval.pdf

Kuhn, D. (2007). Is direct instruction an answer to the right question?. *Educational psychologist*, *42*(2), 109-113.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education* 31: 30-43.

*Lape, N. K., Levy, R., Yong, D. H., Haushalter, K. A., Eddy, R., & Hankel, N. (2014). Probing the inverted classroom: A controlled study of teaching and learning outcomes in undergraduate engineering and mathematics. In *Proceedings of the 121st ASEE Annual Conference & Exposition*. Retrieved from https://www.asee.org/public/conferences/32/papers/9475/view

*Lee, A. M. (2016). *An examination of student outcomes and student satisfaction in a flipped learning environment: A quasi-experimental design* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10126159)

*Lee, N., Lee, L. W., & Kovel, J. (2016). An experimental study of instructional pedagogies to teach math-related content knowledge in construction management education. *International Journal of Construction Education & Research, 12*(4), 255-269. http://dx.doi.org/10.1080/15578771.2016.1141440

*Lewis, J. S., & Harrison, M. A. (2012). Online delivery as a course adjunct promotes active learning and student success. *Teaching of Psychology, 39*(1), 72-76. http://dx.doi.org/10.1177/0098628311430641

*Li, S., & Dan, F. (2015, July). *Influences from university students on the flipped classroom*. Paper presented at the 10th International Conference on Computer Science & Education, Cambridge University, UK. Retrieved from https://ieeexplore.ieee.org/document/7250373

*Liebert, C. A., Lin, D. T., Mazer, L. M., Bereknyei, S., & Lau, J. N. (2016). Effectiveness of the surgery core clerkship flipped classroom: A prospective cohort trial. *American Journal of Surgery, 211*(2), 451-457.e451. http://dx.doi.org/10.1016/j.amjsurg.2015.10.004

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*. Retrieved from https://eric.ed.gov/?id=ED537446.

*Lockman, K., Haines, S. T., & McPherson, M. L. (2017). Improved learning outcomes after flipping a therapeutics module: Results of a controlled trial. *Academic Medicine: Journal Of The Association Of American Medical Colleges, 92*(12), 1786-1793. http://dx.doi.org/10.1097/ACM.0000000000001742

Lopes, A. P., & Soares, F. (2018). Flipping a mathematics course, a blended learning approach. In *INTED2018 Proceedings, 12th International Technology, Education and Development Conference* (pp. 3844-3853). IATED Academy.

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

*Luna, Y. M., & Winters, S. A. (2017). "Why did you blend my learning?" A comparison of student success in lecture and blended learning introduction to sociology courses. *Teaching Sociology, 45*(2), 116-130. http://dx.doi.org/10.1177/0092055X16685373

Machtmes, K., & Asher, J. W. (2000). A meta-analysis of the effectiveness of telecourses in distance education. *American Journal of Distance Education*, *14*(1), 27-46.

*Maciejewski, W. (2016). Flipping the calculus classroom: An evaluative study. *Teaching Mathematics and Its Applications, 35*(4), 187-201. http://dx.doi.org/10.1093/teamat/hrv019

*Marcey, D. J., & Brint, M. E. (2012, November). *Transforming an undergraduate introductory biology course through cinematic lectures and inverted classes: A preliminary assessment of the CLIC model of the flipped classroom*. Paper presented at the 2012 NABT Biology Education Research Symposium, Dallas, TX. Retrieved from http://www.nabt.org/websites/institution/File/docs/Four%20Year%20Section/2012%20Proceedings/Marcey%20&%20Brint.pdf

*Marchalot, A., Dureuil, B., Veber, B., Fellahi, J.-L., Hanouz, J.-L., Dupont, H., . . . Compère, V. (2017). Effectiveness of a blended learning course including e-learning and flipped classroom in first year anaesthesia training. *Anaesthesia, Critical Care & Pain Medicine, 37*(5), 411-415. http://dx.doi.org/10.1016/j.accpm.2017.10.008

*Margoniner, V. (2014). Learning gains in introductory astronomy: Online can be as good as face-to-face. *Physics Teacher, 52*, 298-301. http://dx.doi.org/10.1119/1.4872414

Margulieux, L. E., Bujak, K. R., McCracken, W. M., & Majerich, D. M. (2014, January). Hybrid, blended, flipped, and inverted: Defining terms in a two dimensional taxonomy. In *Paper accepted to the 12th Annual Hawaii International Conference on Education. Honolulu, HI January* (Vol. 2014, pp. 5-9).

Margulieux, L. E., McCracken, W. M., & Catrambone, R. (2015). Mixing in-class and online learning: Content meta-analysis of outcomes for hybrid, blended, and flipped courses. In O. Lindwall, P. Hakkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.) Exploring the Material Conditions of Learning: The Computer Supported Collaborative

Learning (CSCL) Conference (pp. 220-227), 2. Gothenburg, Sweden: The International Society of the Learning Sciences.

*Mason, G. S., Shuman, T. R., & Cook, K. E. (2013). Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course. *IEEE Transactions on Education, 56*, 430-435. http://dx.doi.org/10.1109/TE.2013.2249066

*Mattis, K. V. (2015). Flipped classroom versus traditional textbook instruction: Assessing accuracy and mental effort at different levels of mathematical complexity. *Technology, Knowledge & Learning, 20*(2), 231-248. http://dx.doi.org/10.1007/s10758-014-9238-0

*McCray, G. E. (2000). The hybrid course: Merging online instruction and traditional classroom. *Information Technology and Management, 1*(4), 307-327. http://dx.doi.org/10.1023/A:1019189412115

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

Means, B, Toyama, Y., Murphy, R. F., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature, *Teachers College Record*, 115(3), 1–47. Retrieved from http://www.tcrecord.org/library/content.asp?contentid=16882.

*Melton, B., Graf, H., & Chopak-Foss, J. (2009). Achievement and satisfaction in blended learning versus traditional general health course designs. *International Journal for the*

Scholarship of Teaching and Learning, 3(1). Retrieved from http://digitalcommons.georgiasouthern.edu/ij-sotl/vol3/iss1/26

*Mennella, T. A. (2016). Comparing the efficacy of flipped vs. alternative active learning in a college genetics course. *American Biology Teacher, 78*(6), 471-479. http://dx.doi.org/10.1525/abt.2016.78.6.471

*Moffett, J., & Mill, A. C. (2014). Evaluation of the flipped classroom approach in a veterinary professional skills course. *Advances in Medical Education and Practice, 5*, 415-425. http://dx.doi.org/10.2147/amep.s70160

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269.

Moore, M. G. (1989). Editorial: Three types of interaction. *American Journal of Distance Education*, 3(2).1-6. https://doi.org/10.1080/08923648909526659

*Mooring, S. R., Mitchell, C. E., & Burrows, N. L. (2016). Evaluation of a flipped, large-enrollment organic chemistry course on student attitude and achievement. *Journal of Chemical Education, 93*(12), 1972-1983. http://dx.doi.org/10.1021/acs.jchemed.6b00367

*Mortensen, C. J., & Nicholson, A. M. (2015). The flipped classroom stimulates greater learning and is a modern 21st century approach to teaching today's undergraduates. *Journal of Animal Science, 93*, 3722-3731. http://dx.doi.org/10.2527/jas.2015-9087

*Munson, A., & Pierce, R. (2015). Flipping content to improve student examination performance in a pharmacogenomics course. *American Journal of Pharmaceutical Education, 79*(7), 1-7. http://dx.doi.org/10.5688/ajpe797103

*Murray, L., McCallum, C., & Petrosino, C. (2014). Flipping the classroom experience: A comparison of online learning to traditional lecture. *Journal of Physical Therapy Education, 28*(3), 35-41. http://dx.doi.org/10.1097/00001416-201407000-00006

Njie-Carr, V.P., Ludeman, E., Lee, M.C., Dordunoo, D., Trocky, N.M., Jenkins, L.S., (2017). An integrative review of flipped classroom teaching models in nursing education. *Journal of Professional Nursing*. 33 (2), 133–144.

*O'Connor, E. E., Fried, J., McNulty, N., Shah, P., Hogg, J. P., Lewis, P., . . . Reddy, S. (2016). Flipping radiology education right side up. *Academic Radiology, 23*(7), 810-822. http://dx.doi.org/10.1016/j.acra.2016.02.011

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education*, *25*, 85-95.

*Ojennus, D. D. (2016). Assessment of learning gains in a flipped biochemistry classroom. *Biochemistry and Molecular Biology Education, 44*(1), 20-27. http://dx.doi.org/10.1002/bmb.20926

*Olitsky, N. H., & Cosgrove, S. B. (2016). The better blend? Flipping the principles of microeconomics classroom. *International Review of Economics Educaiton, 21*, 1-11. http://dx.doi.org/10.1016/j.iree.2015.10.004

Osguthorpe, R. T., & Graham, C. R. (2003). Blended learning environments: Definitions and directions. *Quarterly Review of Distance Education, 4,* 227–234. Retrieved from http://www.infoagepub.com/index.php?id=89&i=58

*Overmyer, G. R. (2015). *The flipped classroom model for college algebra: Effects on student achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3635661)

Owston, R., York, D., & Murtha, S. (2013). Student perceptions and achievement in a university blended learning strategic initiative. *The Internet and Higher Education*, *18*, 38-46.

*Papadopoulos, C., Santiago-Roman, A., & Portela, G. (2010). Work in progress – Developing and implementing an inverted classroom for engineering statics. In *Proceedings of 40th ASEE/IEEE Frontiers in Education Conference.* Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5673198

*Pereira, J. A., Pleguezuelos, E., Meri, A., Molina-Ros, A., Molina-Tomas, M., & Masdeu, C. (2007). Effectiveness of using blended learning strategies for teaching and learning human anatomy. *Medical Education, 41*(2), 189-195. http://dx.doi.org/10.1111/j.1365-2929.2006.02672.x

*Peterson, D. J. (2016). The flipped classroom improves student achievement and course satisfaction in a statistics course. *Teaching of Psychology, 43*(1), 10-15. http://dx.doi.org/10.1177/0098628315620063

*Prefume, Y. E. (2015). *Exploring a flipped classroom approach in a Japanese language classroom: A mixed methods study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10024278)

*Prepose, L. S. (2015). *Online, flipped, and traditional instruction: A comparison of student performance in higher education* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10820497)

*Prescott, W. A., Jr., Woodruff, A., Prescott, G. M., Albanese, N., Bernhardi, C., & Doloresco, F. (2016). Introduction and assessment of a blended-learning model to teach patient assessment in a doctor of pharmacy program. *American Journal Of Pharmaceutical Education, 80*(10), 176-176. http://dx.doi.org/10.5688/ajpe8010176

Presti, C. R. (2016). The flipped learning approach in nursing education: A literature review. *Journal of Nursing Education*, *55*(5), 252-257.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of engineering education*, *93*(3), 223-231.

Publications, A. P. A., on Journal, C. B. W. G., & Standards, A. R. (2008). Reporting standards for research in psychology: Why do we need them? What might they be?. *The American Psychologist*, *63*(9), 839.

*Quint, C. L. (2015). *A study of the efficacy of the flipped classroom model in a university mathematics class* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3707108)

*Rau, M. A., Kennedy, K., Oxtoby, L., Bollom, M., & Moore, J. W. (2017). Unpacking "active learning": A combination of flipped classroom and collaboration support is more effective but collaboration support alone is not. *Journal of Chemical Education, 94*(10), 1406-1414. http://dx.doi.org/10.1021/acs.jchemed.7b00240

*Reza, S. B., M. (2015). A study of inverted classroom pedagogy in computer science teaching. *International Journal of Research Studies in Educational Technology, 4*(2). Retrieved from http://www.editlib.org/p/151048/

Rosenshine, B. (2009). 11 The empirical support for direct instruction. In Tobias, S. & Duffy, T. M. (Eds.) *Constructivist Instruction*, (pp. 201-220). New York, NY: Routledge.

*Ruddick, K. (2012). *Improving chemical education from high school to college using a more hands-on approach* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3529991)

*Rutz, E., Eckart, R., Wade, J. E., Maltbie, C., Rafter, C., & Elkins, V. (2003). Student performance and acceptance of instructional technology: Comparing technology-enhanced and traditional instruction for a course in statics. *Journal of Engineering Education, 92*, 133-140. http://dx.doi.org/10.1002/j.2168-9830.2003.tb00751.x

*Ryan, M. D., & Reid, S. A. (2015). Impact of the flipped classroom on student performance and retention: A parallel controlled study in general chemistry. *Journal of Chemical Education, 93*(1), 13-23. http://dx.doi.org/10.1021/acs.jchemed.5b00717

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., Wade, A. & Woods, J. (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. *Computers & Education*, 72, 271-291.

*Şengel, E. (2014). Using the 'flipped classroom' to enhance physics achievement of the prospective teacher impact of flipped classroom model on physics course. *Journal of the Balkan Tribological Association, 20*(3), 488-497.

*Seyedmonir, B., Barry, K., & Seyedmonir, M. (2014). Developing a community of practice (CoP) through interdisciplinary research on flipped classrooms. *Internet Learning Journal, 3*(1), 85-94. Retrieved from http://digitalcommons.apus.edu/internetlearning/vol3/iss1/9/

Shachar, M., & Neumann, Y. (2003). Differences between traditional and distance education academic performances: A meta-analytic approach. *The International Review of Research*

in Open and Distributed Learning, *4*(2). Retrieved from http://www.irrodl.org/index.php/irrodl/%20article/viewArticle/153/234.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.Staker, H., & Horn, M. B. (2012). Classifying K-12 Blended Learning. *Innosight Institute*.

*Shattuck, J. C. (2016). A parallel controlled study of the effectiveness of a partially flipped organic chemistry course on student performance, perceptions, and course completion. *Journal of Chemical Education, 93*(12), 1984-1992. http://dx.doi.org/10.1021/acs.jchemed.6b00393

*Shyu, H. Y. (2014). Implementing the flipped classroom strategy into in-service education. In T. Bastiaens (Ed.), *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2014* (pp. 1819-1823). Chesapeake, VA: Association for the Advancement of Computing in Education. Retrieved from http://www.editlib.org/p/149001/

Sitzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. *Personnel psychology*, *59*(3), 623-664.

Spanjers, I. A., Könings, K. D., Leppink, J., Verstegen, D. M., de Jong, N., Czabanowska, K., & Van Merriënboer, J. J. (2015). The promised land of blended learning: Quizzes as a moderator. *Educational Research Review*, *15*, 59-74.

Spring, K., & Graham, C. (2017). Thematic patterns in international blended learning literature, research, practices, and terminology. *Online Learning Journal*, *21*(4).

Staker, H., & Horn, M. B. (2012). Classifying K-12 blended learning. *Innosight Institute*.

*Stickel, M. (2014). Teaching electromagnetism with the inverted classroom approach: Student perceptions and lessons learned. In *Proceedings of the 121st ASEE Annual Conference & Exposition*. Retrieved from https://www.asee.org/public/conferences/32/papers/10572/view

*Street, S. E., Gilliland, K. O., McNeil, C., & Royal, K. (2015). The flipped classroom improved medical student performance and satisfaction in a pre-clinical physiology course. *Medical Science Educator, 25*(1), 35-43. http://dx.doi.org/10.1007/s40670-014-0092-4

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, 0034654317751919.

Tallmadge, G. K. (1977). Ideabook: The joint dissemination review panel. *Washington, DC: US Office of Education*.

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

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology*, *53*(2), 207-216.

Trogden, B. G. (2015). The view from a flipped classroom: Improved student success and subject mastery in organic chemistry. In *Implementation and critical assessment of the flipped classroom experience* (pp. 119-137). IGI Global.

*Tune, J. D., Sturek, M., & Basile, D. P. (2013). Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. *Advances in Physiology Education, 37*, 316-320. http://dx.doi.org/10.1152/advan.00091.2013

*Turan, Z., & Goktas, Y. (2016). The flipped classroom: Instructional efficiency and impact on achievement and cognitive load levels. *Journal of E-Learning & Knowledge Society, 12*(4), 51-62. Retrieved from https://www.learntechlib.org/p/173672/

*Van Sickle, J. (2016). Discrepancies between student perception and achievement of learning outcomes in a flipped classroom. *Journal of the Scholarship of Teaching and Learning, 16*(2), 29-38.

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.

*Wasserman, N. H., Quint, C., Norris, S. A., & Carr, T. (2017). Exploring flipped classroom instruction in calculus iii. *International Journal of Science and Mathematics Education, 15*(3), 545-568. http://dx.doi.org/10.1007/s10763-015-9704-8

*Webb, M., & Doman, E. (2016). Does the flipped classroom lead to increased gains on learning outcomes in esl/efl contexts? *CATESOL Journal, 28*(1), 39-67.

*Webster, D. R., Majerich, D. M., & Madden, A. G. (2016). Flippin' fluid mechanics - Comparison using two groups. *Advances in Engineering Education, 5*(3), 1-20.

*Weng, P. (2015). Developmental math, flipped and self-paced. *PRIMUS, 25*, 768-781. http://dx.doi.org/10.1080/10511970.2015.1031297

*Whillier, S., & Lystad, R. P. (2015). No differences in grades or level of satisfaction in a flipped classroom for neuroanatomy. *The Journal Of Chiropractic Education, 29*, 127-133. http://dx.doi.org/10.7899/JCE-14-28

*Whitman Cobb, W. N. (2016). Turning the classroom upside down: Experimenting with the flipped classroom in American government. *Journal of Political Science Education, 12*(1), 1-14. http://dx.doi.org/10.1080/15512169.2015.1063437

*Willis, J. A. (2014). *The effects of flipping an undergraduate precalculus class* (Unpublished Doctoral dissertation). Appalachian State University, Boone, NC.

*Wilson, S. G. (2013). The flipped class: A method to address the challenges of an undergraduate statistics course. *Teaching of Psychology, 40*, 193-199. http://dx.doi.org/10.1177/0098628313487461

Wilson, S. M., & Peterson, P. L. (2006). Theories of learning and teaching: What do they mean for educators? Working paper. *National Education Association Research Department*.

Williams, S. L. (2006). The effectiveness of distance education in allied health science programs: A meta-analysis of outcomes. *The American Journal of Distance Education*, *20*(3), 127-141.

*Winter, J. B. (2013). *The effect of the flipped classroom model on achievement in an introductory college physics course* (Doctoral dissertation, Mississippi State University).

*Wong, T. H., Ip, E. J., Lopes, I., & Rajagopalan, V. (2014). Pharmacy students' performance and perceptions in a flipped teaching pilot on cardiac arrhythmias. *American Journal of Pharmaceutical Education, 78*(10), 1-6. http://dx.doi.org/10.5688/ajpe7810185

*Yelamarthi, K., Drake, E., & Prewett, M. (2016). An instructional design framework to improve student learning in a first-year engineering class. *Journal of Information Technology Education: Innovations in Practice, 15*, 195-222.

*Zhao, Y., & Ho, A. (2014). *Evaluating the flipped classroom in an undergraduate history course* (HarvardX Research Memo). Retrieved from Harvard Graduate School of Education website: http://harvardx.harvard.edu/files/harvardx/files/evaluating_the_flipped_classroom_-_zhao_and_ho.pdf

Zhao, Y., Lei, J., Yan, B., Lai, C., & Tan, H. S. (2005). What makes the difference? A practical analysis of research on the effectiveness of distance education. *Teachers College Record*, *107*(8), 1836

*Zhonggen, Y., & Guifang, W. (2016). Academic achievements and satisfaction of the clicker-aided flipped business english writing class. *Educational Technology & Society, 19*(2), 298-312.

*Ziegelmeier, L. B., & Topaz, C. M. (2015). Flipped calculus: A study of student performance and perceptions. *PRIMUS, 25*, 847-860. http://dx.doi.org/10.1080/10511970.2015.1031305

APPENDIX A. CATEGORIES, NUMBERS, AND % OF EXCLUDED FULL-TEXT STUDIES

Category	Excluded full-text studies	
	Number	%
Study quality insufficient (e.g., one group pre-test post-test)	26	22
Insufficient statistics to calculate an effect size	23	19
Student feedback; survey data	17	14
Wrong population; wrong topic; opinion; literature review	22	18
Definition of flipped classroom violated in the study	11	9
Duplicates	9	8
Multiple reasons	9	8
Language other than English or non-retrievable	3	3
Total	120	100

APPENDIX B. CODEBOOK

1) Publication Features

- Study number
- Author(s) Name(s)
- Year of publication
- Publication type (Peer Reviewed Journal, Dissertation/Thesis, other)

2) Methodological Features

- Design (experiment, quasi-experimental, pre-experimental)
- Instructor (Same or different)
- Semester (Same or different)
- Outcome Measurement (One, average, composite, e.g., final course grade)
- Sample size (experimental)
- Sample size (control)
- Sample size (total)
- Effect size magnitude (Cohen's d converted to hedge's g)
- Effect size direction (positive or negative)

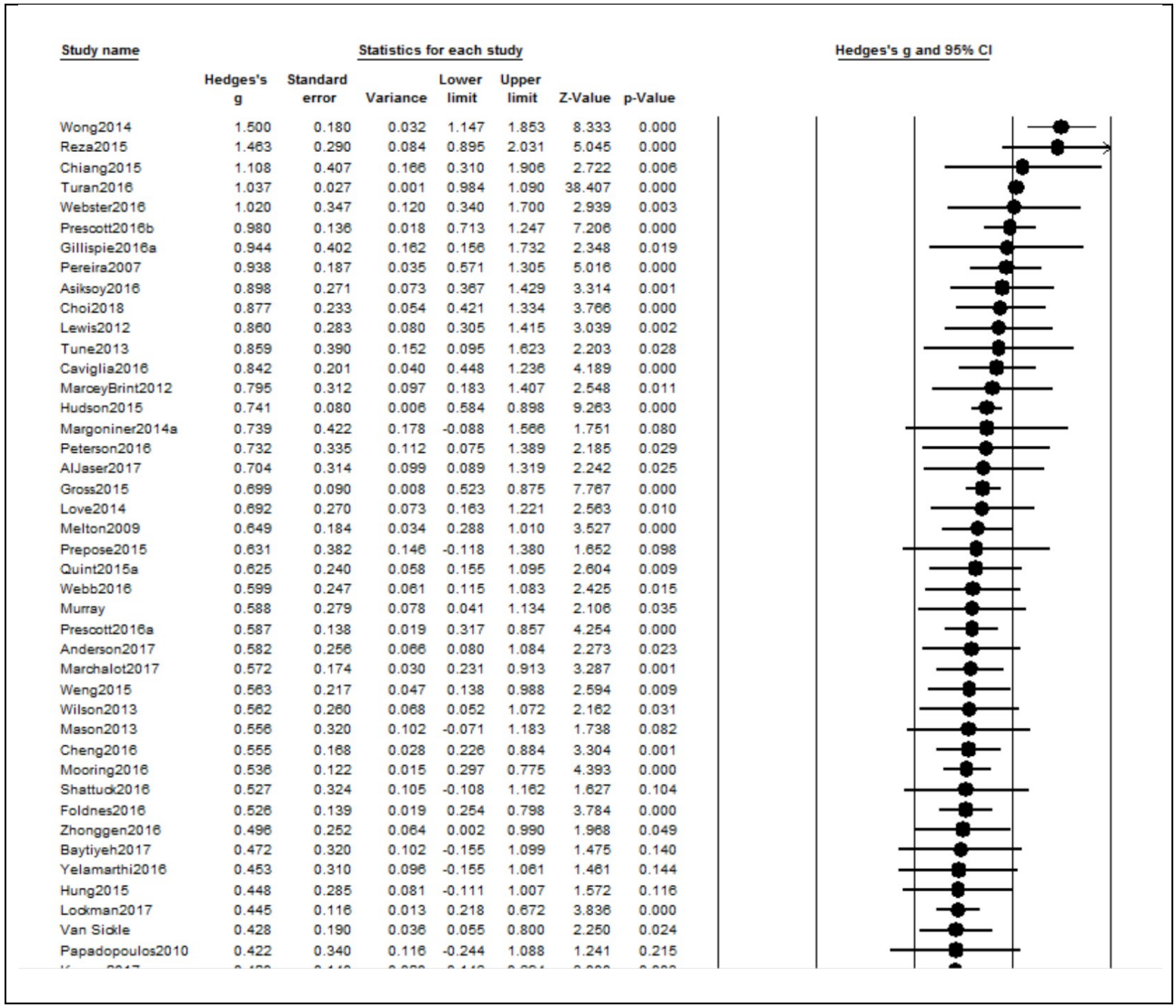
3) Course demographics

- Control Condition (F2F or Online)
- Control Description (describe)
- Course Year Level (Graduate or undergraduate)
- Course Subject matter (STEM or Non-STEM)
- STEM subject (Science, Technology, Engineering, or Math or Health)
- Instructor experience facilitating active learning

4) Pedagogical factors

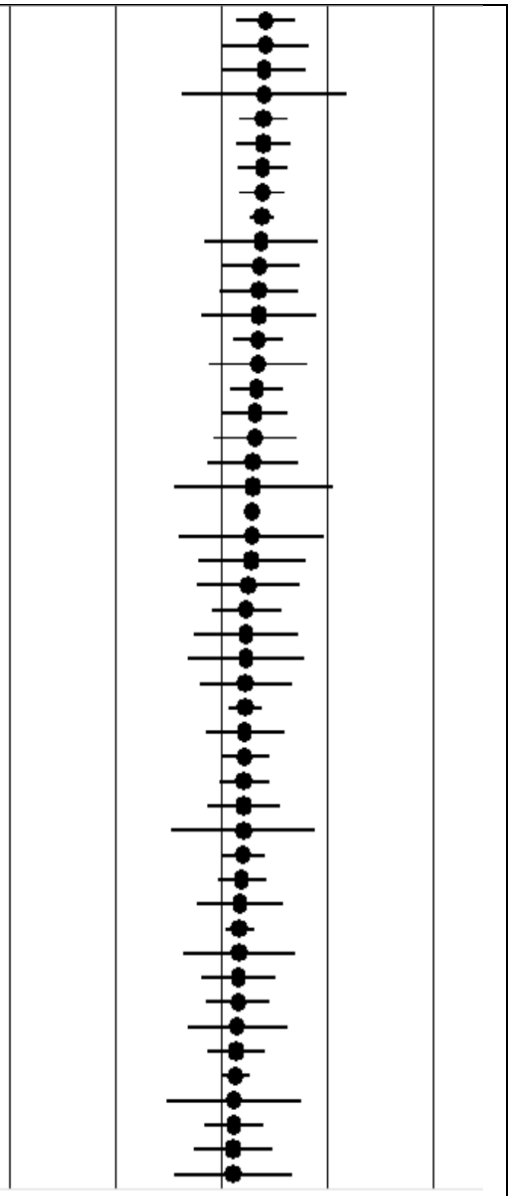
- Quizzes/pre-class assessment used regularly in treatment (yes or no)
- Time on Task (Equivalent/ greater than/ less than*)*

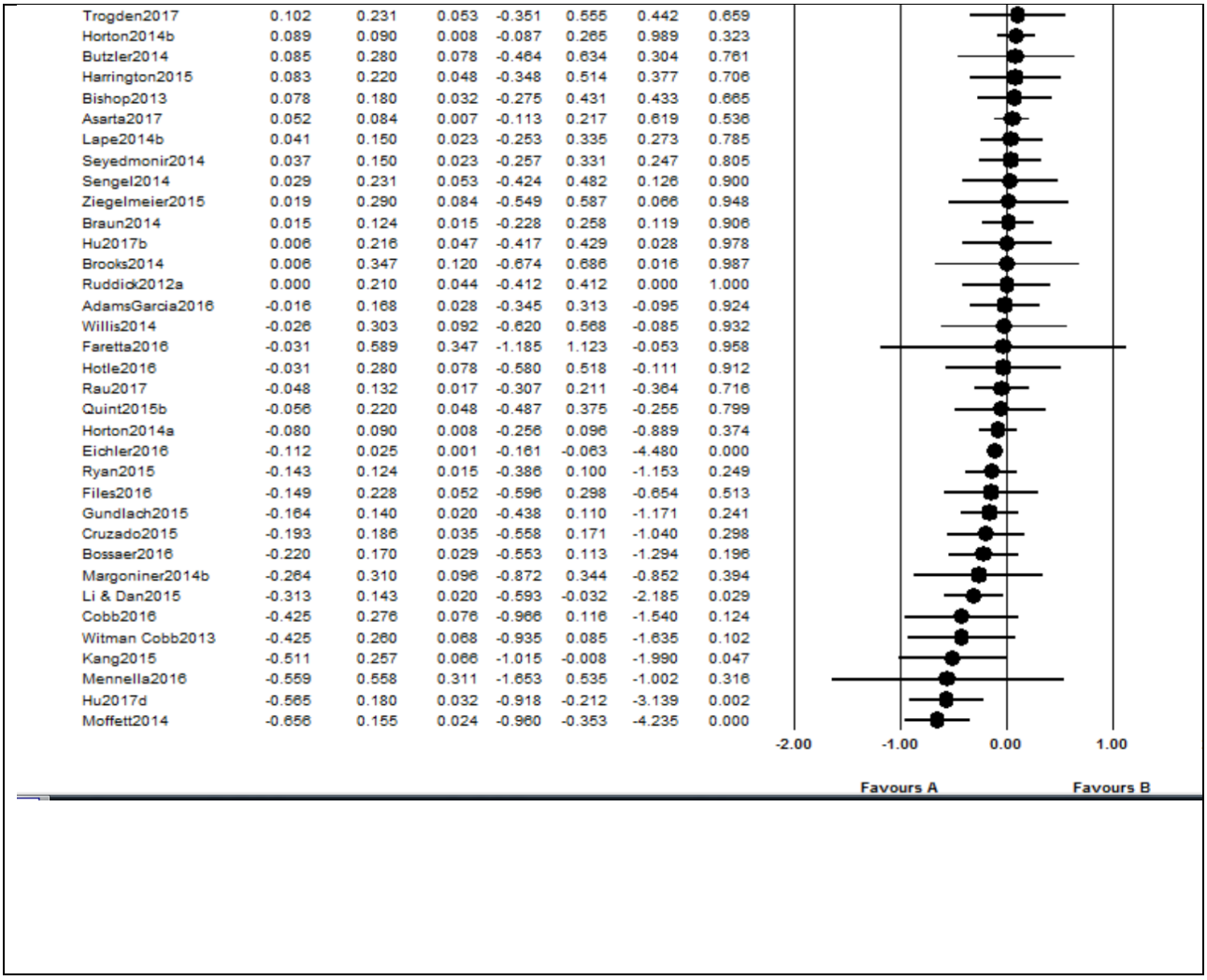
APPENDIX C: FOREST PLOT FOR THE FULL SET OF 125 EFFECT SIZES



Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Wong2014	1.500	0.180	0.032	1.147	1.853	8.333	0.000
Reza2015	1.463	0.290	0.084	0.895	2.031	5.045	0.000
Chiang2015	1.108	0.407	0.166	0.310	1.906	2.722	0.006
Turan2016	1.037	0.027	0.001	0.984	1.090	38.407	0.000
Webster2016	1.020	0.347	0.120	0.340	1.700	2.939	0.003
Prescott2016b	0.980	0.136	0.018	0.713	1.247	7.206	0.000
Gillispie2016a	0.944	0.402	0.162	0.156	1.732	2.348	0.019
Pereira2007	0.938	0.187	0.035	0.571	1.305	5.016	0.000
Asiksoy2016	0.898	0.271	0.073	0.367	1.429	3.314	0.001
Choi2018	0.877	0.233	0.054	0.421	1.334	3.766	0.000
Lewis2012	0.860	0.283	0.080	0.305	1.415	3.039	0.002
Tune2013	0.859	0.390	0.152	0.095	1.623	2.203	0.028
Caviglia2016	0.842	0.201	0.040	0.448	1.236	4.189	0.000
MarceyBrint2012	0.795	0.312	0.097	0.183	1.407	2.548	0.011
Hudson2015	0.741	0.080	0.006	0.584	0.898	9.263	0.000
Margoniner2014a	0.739	0.422	0.178	-0.088	1.566	1.751	0.080
Peterson2016	0.732	0.335	0.112	0.075	1.389	2.185	0.029
AlJaser2017	0.704	0.314	0.099	0.089	1.319	2.242	0.025
Gross2015	0.699	0.090	0.008	0.523	0.875	7.767	0.000
Love2014	0.692	0.270	0.073	0.163	1.221	2.563	0.010
Melton2009	0.649	0.184	0.034	0.288	1.010	3.527	0.000
Prepose2015	0.631	0.382	0.146	-0.118	1.380	1.652	0.098
Quint2015a	0.625	0.240	0.058	0.155	1.095	2.604	0.009
Webb2016	0.599	0.247	0.061	0.115	1.083	2.425	0.015
Murray	0.588	0.279	0.078	0.041	1.134	2.106	0.035
Prescott2016a	0.587	0.138	0.019	0.317	0.857	4.254	0.000
Anderson2017	0.582	0.256	0.066	0.080	1.084	2.273	0.023
Marchalot2017	0.572	0.174	0.030	0.231	0.913	3.287	0.001
Weng2015	0.563	0.217	0.047	0.138	0.988	2.594	0.009
Wilson2013	0.562	0.260	0.068	0.052	1.072	2.162	0.031
Mason2013	0.556	0.320	0.102	-0.071	1.183	1.738	0.082
Cheng2016	0.555	0.168	0.028	0.226	0.884	3.304	0.001
Mooring2016	0.536	0.122	0.015	0.297	0.775	4.393	0.000
Shattuck2016	0.527	0.324	0.105	-0.108	1.162	1.627	0.104
Foldnes2016	0.526	0.139	0.019	0.254	0.798	3.784	0.000
Zhonggen2016	0.496	0.252	0.064	0.002	0.990	1.968	0.049
Baytiyeh2017	0.472	0.320	0.102	-0.155	1.099	1.475	0.140
Yelamarthi2016	0.453	0.310	0.096	-0.155	1.061	1.461	0.144
Hung2015	0.448	0.285	0.081	-0.111	1.007	1.572	0.116
Lockman2017	0.445	0.116	0.013	0.218	0.672	3.836	0.000
Van Sickle	0.428	0.190	0.036	0.055	0.800	2.250	0.024
Papadopoulos2010	0.422	0.340	0.116	-0.244	1.088	1.241	0.215

Karaca2017	0.420	0.140	0.020	0.146	0.694	3.000	0.003
ZhaoHo2014	0.419	0.210	0.044	0.007	0.831	1.995	0.046
Hussey2015	0.406	0.200	0.040	0.014	0.798	2.030	0.042
Gillispie2016b	0.401	0.396	0.157	-0.375	1.177	1.013	0.311
Cotta2016	0.400	0.116	0.013	0.173	0.627	3.448	0.001
Olitsky2016	0.400	0.130	0.017	0.145	0.654	3.075	0.002
Maciejewski2016	0.387	0.119	0.014	0.154	0.620	3.252	0.001
McLaughlin2014	0.387	0.110	0.012	0.171	0.603	3.518	0.000
OConnor2016	0.384	0.060	0.004	0.266	0.502	6.400	0.000
Ojennus2016	0.373	0.271	0.073	-0.158	0.904	1.377	0.169
Luna2017	0.365	0.189	0.036	-0.005	0.735	1.931	0.053
Winter2013	0.355	0.190	0.036	-0.017	0.727	1.868	0.062
Fraga2014	0.355	0.278	0.077	-0.190	0.900	1.276	0.202
Mortensen2015	0.350	0.120	0.014	0.115	0.585	2.917	0.004
Guerrero2015	0.347	0.240	0.058	-0.123	0.817	1.446	0.148
Bagley2014	0.332	0.130	0.017	0.077	0.587	2.554	0.011
Haughten2015b	0.317	0.157	0.025	0.010	0.625	2.021	0.043
Rutz2003	0.316	0.200	0.040	-0.076	0.708	1.580	0.114
Geist2015	0.299	0.220	0.048	-0.132	0.730	1.359	0.174
LeeLeeKovel2016	0.299	0.382	0.146	-0.450	1.048	0.783	0.434
Kim2017	0.289	0.024	0.001	0.242	0.336	12.042	0.000
Shyu2014	0.288	0.350	0.123	-0.398	0.974	0.823	0.411
Whillier2015	0.286	0.260	0.068	-0.224	0.796	1.100	0.271
Adams2013	0.255	0.250	0.063	-0.235	0.745	1.020	0.308
Wasserman2017	0.236	0.168	0.028	-0.093	0.565	1.405	0.160
Ruddick2012b	0.232	0.250	0.063	-0.258	0.722	0.928	0.353
McCray2000	0.231	0.280	0.078	-0.318	0.780	0.825	0.409
LeeLiu2016	0.226	0.222	0.049	-0.209	0.661	1.018	0.309
Stickel2014	0.224	0.080	0.006	0.067	0.381	2.800	0.005
Davies2013	0.221	0.190	0.036	-0.151	0.593	1.163	0.245
Overmyer2014	0.220	0.116	0.013	-0.007	0.448	1.900	0.057
Munson2015	0.214	0.120	0.014	-0.021	0.449	1.783	0.075
Heyborne2016	0.211	0.176	0.031	-0.134	0.556	1.199	0.231
Prefume2015	0.209	0.346	0.120	-0.469	0.887	0.604	0.546
Street2015	0.206	0.106	0.011	-0.002	0.414	1.942	0.052
Haughten2015a	0.191	0.116	0.013	-0.036	0.419	1.649	0.099
Jeavons2013	0.175	0.210	0.044	-0.237	0.587	0.833	0.405
Albert2014	0.169	0.069	0.005	0.034	0.304	2.448	0.014
Lape2014a	0.167	0.270	0.073	-0.362	0.696	0.619	0.536
Hu2017c	0.165	0.180	0.032	-0.188	0.518	0.917	0.359
Liebert2016	0.160	0.153	0.023	-0.140	0.460	1.046	0.296
Hu2017a	0.149	0.241	0.058	-0.323	0.621	0.618	0.536
Day2018	0.142	0.139	0.019	-0.130	0.414	1.022	0.307
Baepler2014	0.131	0.070	0.005	-0.006	0.268	1.871	0.061
Harvey2014	0.119	0.322	0.104	-0.512	0.750	0.369	0.712
He2016	0.116	0.144	0.021	-0.166	0.398	0.806	0.420
Jensen2015	0.112	0.190	0.036	-0.260	0.484	0.589	0.556
Mattis2015	0.110	0.285	0.081	-0.448	0.669	0.387	0.698




Trogden2017 0.102 0.231 0.053 -0.351 0.555 0.442 0.659
Horton2014b 0.089 0.090 0.008 -0.087 0.265 0.989 0.323
Butzler2014 0.085 0.280 0.078 -0.464 0.634 0.304 0.761
Harrington2015 0.083 0.220 0.048 -0.348 0.514 0.377 0.706
Bishop2013 0.078 0.180 0.032 -0.275 0.431 0.433 0.665
Asarta2017 0.052 0.084 0.007 -0.113 0.217 0.619 0.536
Lape2014b 0.041 0.150 0.023 -0.253 0.335 0.273 0.785
Seyedmonir2014 0.037 0.150 0.023 -0.257 0.331 0.247 0.805
Sengel2014 0.029 0.231 0.053 -0.424 0.482 0.126 0.900
Ziegelmeier2015 0.019 0.290 0.084 -0.549 0.587 0.066 0.948
Braun2014 0.015 0.124 0.015 -0.228 0.258 0.119 0.906
Hu2017b 0.006 0.216 0.047 -0.417 0.429 0.028 0.978
Brooks2014 0.006 0.347 0.120 -0.674 0.686 0.016 0.987
Ruddick2012a 0.000 0.210 0.044 -0.412 0.412 0.000 1.000
AdamsGarcia2016 -0.016 0.168 0.028 -0.345 0.313 -0.095 0.924
Willis2014 -0.026 0.303 0.092 -0.620 0.568 -0.085 0.932
Faretta2016 -0.031 0.589 0.347 -1.185 1.123 -0.053 0.958
Hotle2016 -0.031 0.280 0.078 -0.580 0.518 -0.111 0.912
Rau2017 -0.048 0.132 0.017 -0.307 0.211 -0.364 0.716
Quint2015b -0.056 0.220 0.048 -0.487 0.375 -0.255 0.799
Horton2014a -0.080 0.090 0.008 -0.256 0.096 -0.889 0.374
Eichler2016 -0.112 0.025 0.001 -0.161 -0.063 -4.480 0.000
Ryan2015 -0.143 0.124 0.015 -0.386 0.100 -1.153 0.249
Files2016 -0.149 0.228 0.052 -0.596 0.298 -0.654 0.513
Gundlach2015 -0.164 0.140 0.020 -0.438 0.110 -1.171 0.241
Cruzado2015 -0.193 0.186 0.035 -0.558 0.171 -1.040 0.298
Bossaer2016 -0.220 0.170 0.029 -0.553 0.113 -1.294 0.196
Margoniner2014b -0.264 0.310 0.096 -0.872 0.344 -0.852 0.394
Li & Dan2015 -0.313 0.143 0.020 -0.593 -0.032 -2.185 0.029
Cobb2016 -0.425 0.276 0.076 -0.966 0.116 -1.540 0.124
Witman Cobb2013 -0.425 0.260 0.068 -0.935 0.085 -1.635 0.102
Kang2015 -0.511 0.257 0.066 -1.015 -0.008 -1.990 0.047
Mennella2016 -0.559 0.558 0.311 -1.653 0.535 -1.002 0.316
Hu2017d -0.565 0.180 0.032 -0.918 -0.212 -3.139 0.002
Moffett2014 -0.656 0.155 0.024 -0.960 -0.353 -4.235 0.000
-2.00
-1.00
0.00
1.00
Favours A
Favours B