UNIVERSITY OF CALIFORNIA,
IRVINE

Investigation of the Effects of Flipped Instruction on Student Exam Performance, Motivation and Perceptions

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Wenliang He

Dissertation Committee:
Professor George Farkas, Chair
Associate Professor Penelope Collins
Professor Diane O'Dowd

2016

Dedication

To

my wife, Zhuying

Sharing our life and love along this journey together is a blessing beyond words.

To

my grandfather, Jinpei

whose commitment to work and attitudes towards life inspire and motivate me.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express the deepest appreciation to my graduate advisor and committee chair, Professor George Farkas, whose invaluable guidance has got me through some difficult times and whose constant encouragement has given me the confidence to reach for new heights. His rigorous research approach and scholarship inspire me to conduct my research to high standards.

I would also like to thank my committee members, Associate Professor Penelope Collins and Professor Diane O'Dowd. Dr. Collins encourages me to explore new research interests and has been a strong influence on my choice of research direction. Dr. O'Dowd is a phenomenal instructor whose instructional practices and research have influenced my view on learning and instruction.

I am also grateful to Dr. Amanda Holton and Dr. Renee Link. I have collaborated with them over a variety of projects. Without her help and support, this dissertation would not be possible.

Curriculum Vitae

Wenliang He

EDUCATION

- **Ph.D. in Education, University of California Irvine (UCI), CA.** 09/2011 – 09/2016

Dissertation: Investigation of the effects of flipped instruction on student exam performance, motivation and perceptions.
Journal & Conference Reviewer: Computers & Education, AERA Conference, ASEE Conference.
Related coursework: Applied Regression Analysis, Structural Equation Modeling, Experimental Design, Educational Tests and Measurement, Introduction to Meta-analyses.

- **M.S. in Statistics, University of California Irvine, CA.** 09/2013 - 12/2015

Related coursework: Statistical Computation, Bayesian Data Analysis, Probabilistic Learning, Data Mining, Generalized Linear Models, Longitudinal and Survival Analysis, Survey Sampling.

- **M.A. in Education, University of California Irvine, CA.** 09/2011 - 06/2014
- **B.E. in Materials Science & Engineering, Beijing Univ. of Tech.** 09/2000 - 07/2004

TECHNICAL SKILLS

R Programming, Python (numpy, pandas, scikit-learn), Matlab, SQL, GitHub, Unix Shell, Stata

RESEARCH PROJECTS

IVLE Project, SoE and Department of Computer Science, UCI 09/2015 – 09/2016

- Conduct campus-wide educational ***data mining*** with online, hybrid, and flipped courses taught at UCI. Crawl and clean canvas data. Analyze student clickstream data using Markov Chain model, perform data visualization to understand click events and student online learning behavior, and apply unsupervised learning to cluster students. Extract interesting variables (e.g. procrastination) for inferential statistics.

General Chemistry, Department of Chemistry, UCI 09/2013 – 09/2016

- Conduct ***quasi-experiments*** to assess flipped instruction in a first-year general chemistry course using a ***mixed method*** approach. Partly ***automatize*** and perform extensive ***exploratory analysis*** to inform subsequent modeling. Apply ***linear regression*** and ***structural equation modeling*** to examine treatment, interaction and mediating effects. Use student comments to aid in interpreting quantitative results.
- Design ***survey*** and construct ***measurement scale*** on motivation and satisfaction. Perform ***exploratory*** and ***confirmatory factor analysis*** to validate the scale. Conduct systematic ***literature review***.
- Use results to iteratively improve the course over three years, which leads to three research papers.

Text Mining, School of Education, UCI 09/2013 – 07/2014

- Automatically generate frequently asked questions (FAQs) based on students inquires from surveys and online forums. Perform ***classification*** (e.g., ***naïve Bayes***, logistic regression, and ***support vector machine***) to separate real questions from non-questions,

and implement several ***clustering*** algorithms (e.g., ***K-means***, ***EM***, and ***latent Dirichlet allocation***) to produce question clusters.
- Derive and code ***topic modeling*** from scratch and simulate data to examine its performance with documents of limited number of tokens.

Biology Laboratories, School of Biological Sciences, UCI 09/2012 – 07/2015
- Provide ***statistical consulting and analyses*** to four faculty members from two departments on multiple projects using ***experiments*** and ***quasi-experiment design*** (e.g. assessing the impact of writing, online videos, open-notes, visualization and group size on knowledge transfer), which led to two research papers.
- Use ***linear regression***, ***logistic regression***, ***mixed effects model*** to examine treatment, interaction and mediating effects. Draw a variety of ***graphs*** to visualize and summarize results.

Organic Chemistry, Department of Chemistry, UCI 09/2013 – 07/2014
- Collect ***longitudinal data*** and perform ***interrupted time series*** to study the effectiveness of flipped instruction in an organic chemistry course. The results led to an AERA conference paper.

Digital Design, Department of Electrical & Computer Engineering, UCI 06/2012 – 08/2013
- Use ***fixed effects*** and ***generalized least squares models*** to analyze ***longitudinal data*** to evaluate student performance in a flexible hybrid electrical engineering course. Use ***qualitative methods***, e.g. class ***observation***, coding videos and conducting ***interviews***. Record and edit Khan Academy style instructional videos. The result led to a research paper.

ACADEMIC WORKSHOPS

Workshop Presenter, Intro to R, UCI Data Science Initiative 06/2016
- Co-led workshop in ***Introduction to R*** designed for graduate students interesting in R programming and analysis. It covers basic R syntax, functions, programming fundamentals, and statistical analysis and visualization techniques.

Workshop Presenter, SoCAL PKAL Annual Meeting, UC Irvine 02/2016
- Led workshop in ***statistical data analysis*** using R designed specifically for STEM instructors. It began with a short crash course in R, followed by data management and preliminary data analysis using a combination of descriptive statistics and graphs. I wrote R functions to semi-automate data management and analysis routines and taught how to use preliminary analyses to inform subsequent modeling.

Quasi-Experiment Workshop, Northwestern University, IL 08/2015
- Led by Thomas Cook and William Shadish, the two-week intensive training program covered regression discontinuity design, interrupted time series, propensity score matching, and instrumental variable method.

Software Carpentry Workshop, University of California Irvine, CA 02/2015
- Learned data management with Unix Shell and Pandas, and version control using GitHub.

Learning Analytics Summer Institute (LASI) 2014, Harvard University, MA 06/2014

- Participated in three workshops, i.e. data visualization and learning analytics, multimodal learning analytics, and text mining and educational discourse.

LearnLab Summer School 2014, Carnegie Mellon University, PA 06/2014

- Participated in the one-week Educational Data Mining (EDM) track and completed the final project on evaluating different algorithms for estimating parameters in Bayesian Knowledge Tracing.

PUBLICATIONS & CONFERENCE PRESENTATIONS

- **He, W.**, & Farkas, G. (under preparation). A critical review of the flipped pedagogy: research to date and directions for the future.
- **He, W.**, Holton, A., & Farkas, G. (to be submitted). Impact of flipped instruction on immediate and subsequent course performance in a large undergraduate chemistry course.
- **He, W.**, Holton, A., Warschauer, M., & Farkas, G. (under review). Differentiated impact of flipped instruction in a large undergraduate chemistry course.
- **He, W.**, Song, Y., Sato, B., & Kadandale, P. (under review). Confidence in confidence: Deconstructing the components of self-reported confidence, *CBE Life Sciences Education*.
- Lie, R., Abdullah, C., **He, W.**, & Tour, E. (accepted). Perceived barriers to engagement with the primary literature: effects of experience and instruction, *CBE Life Sciences Education*.
- **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.
- **He, W.**, Gajski, D., Farkas, G., & Warschauer, M. (2015). Implementing flexible hybrid instruction in an electrical engineering course: The best of three worlds? *Computers & Education*, 81, 59-68.
- **He, W.**, & Link, R. (2015). Bridging the achievement gap: A longitudinal study of the differential lingering effects of flipped instruction. AERA 2015 Annual Conference, Chicago, IL.
- Sato, B.K., Kadandale, P. **He, W.**, & Warschauer, M. (2015). The grass isn't always greener: Perceptions of and performance on open note exams, *CBE Life Sciences Education*, *14*(2) ar11.
- Sato, B.K., Kadandale, P. **He, W.**, Murata, P.M., Latif, Y., & Warschauer, M. (2014). Practice makes pretty good: Assessment of primary literature reading abilities across multiple large-enrollment biology laboratory courses. *CBE Life Sciences Education*, *13*(4), 677-686.
- Jiang, S., Williams, A.E., Warschauer, M., **He, W.**, & O'Dowd, D.K. (2014). Influence of incentives on performance in a pre-college biology MOOC. *The International Review of Research in Open and Distributed Learning*, *15*(5).
- Gajski, D.D., Dang, Q.V., & **He, W.** (2013). An online methodology for individualized education, *2013* International Conference on e-Learning, e-Business, EIS, & e-Government (EEE'13), Las Vegas, NV.
- **He, W.**, Gallway, E.P., Hsu, J., White, C., Lawrence, J.F., & Snow, C.E. (2012). Patterns of students' vocabulary improvement from one-time instruction and review instruction, LRA 62nd Annual Conference, San Diego, CA.

- **He, W.,** Fischer, T., & Hess, H. (2008). Surface Patterning and Functionalization for Biomolecular Motor Nanotechnology. In *ACS symposium series*, 986, 354-374. Oxford University Press.
- **He, W.**, Wang, H., &Yan, H. (2005). On the mechanism of electro-deposition of hydrogen-free diamond-like carbon films. *Carbon*, 43, 2000-2006.
- Xu, H., **He**, **W.,** Wang, H., & Yan, H. (2004). Solvothermal synthesis of $K_2V_3O_8$ nanorods. *Journal of Crystal Growth*, 260, 447-450.
- Xu, H., Wang, H., Zhang, Y., **He, W.**, & Zhu, M. (2004). Hydrothermal synthesis of zinc oxide powders with controllable morphology. *Ceramics International*, 30, 93-97.
- **He, W.**, Zhang, Y., Zhang, X., Wang, H., & Yan, H. (2003). Low temperature preparation of nanocrystalline Mn_2O_3 via ethanol-thermal reduction of MnO_2. *Journal of Crystal Growth*, 252, 285-288.
- **He**, **W.**, Wang, H., & Yan, H. (2003). Electro-deposition of diamond-like carbon and carbon nitride thin films in liquid phase. The 8th IUMRS International Conference on Advanced Materials, Yokohama, Japan.
- **He**, **W.,** Zhang, Y., & Yan, H. (2002). The synthesis of nanocrystalline $(LaxNd_{1-x})_{0.7}Sr_{0.3}MnO_3$ through non-alkoxide sol-gel method. MRS 2002 Chinese Conference, Beijing, China.

Abstract of the Dissertation

Investigation of the Effects of Flipped Instruction on
Student Exam Performance, Motivation and Perceptions.

By

Wenliang He

Doctor of Philosophy in Education

University of California, Irvine, 2016

Professor George Farkas, Chair

The goal of this dissertation is to investigate the effects of flipped instruction on student exam performance, motivation, and perceptions. Flipped instruction was implemented in three consecutive years in the same introductory chemistry course taught by the same instructor. Surveys were delivered to measure out-of-class study time, student motivation, perceived instruction clarity and quality. Our studies have consistently shown that flipped instruction did not appreciably increase students' overall study time outside the classroom. It only causes a shift in student workload. Our first study shows that flipped instruction had a small and statistically significant effect on student final exam performance with no marked interaction effect. Student responses to the flipped pedagogy was distinctly lukewarm with about one fifth of the students showing polarized feelings. Non-compliance with pre-class study was found to be a serious implementation issue, which might lead to the small treatment effect and absence of interaction. Giving assignments and quizzes associated with each video effectively reduced non-compliance, as shown by the second study. However, technological failures in class seemed to result in flipped students consistently rating the class to be of lower quality. Accordingly, flipped students were shown to underperform their control counterparts. Moreover, second-year students and

females benefit more from flipped instruction. The variety of issues exposed during the first two years prompted us to reflect upon the resilience of traditional lectures, where its simplicity might be its greatest virtue. We therefore caution against overreliance on complex technologies or teaching techniques. Finally, with non-compliance and technology failures solved, the third study adopted a softer approach to introducing flipped instruction by periodically adjusting the balance between lecturing and active learning components. Although the results showed no treatment effect on student final exam performance, students from the flipped section who enrolled into a subsequent course outperformed their control counterparts in post-course grade. Moreover, students with lower high school GPA to start with benefited more from the flipped pedagogy. Collectively speaking, it is advisable that flipped instructors in first-year introductory courses should start simple and be cautious of deviating from traditional lectures too much too fast.

Chapter 1 Introduction

1.1 Introduction to the Flipped Instruction Model

Flipped instruction is a recent phenomenon that has attracted growing attention from both teaching and research communities (Bishop & Verleger, 2013). The excitement over flipped instruction is primarily due to its capacity to fuse two existing directions of research, i.e. online instruction and active learning techniques, into a new pedagogy that is more than the sum of its parts. The essence of flipped instruction is to stage learning of new material before class in order to free up class time for more practice and productive use of knowledge via a variety of active learning techniques (Tucker, 2012). This reorganization of the sequence of teaching and learning, hence the name "flipped" instruction, is in stark contrast to the traditional lecture format, where new material is typically introduced by teachers during the class, followed by students reviewing the content at a later time with subsequent homework for practice and tests for evaluation of learning outcomes.

For advocates of flipped instruction, the rationale for using this pedagogy is threefold. First, over the years, a number of published meta-analyses comparing student academic outcomes using online versus face-to-face instruction have generally shown that on average online instruction is about as effective as classroom instruction (Machtmes & Asher, 2000; Bernard et al., 2004; Sitzmann, Kraiger, Stewart, & Wisher, 2006; Means, Toyama, Murphy, Bakia, & Jones, 2009), even though great heterogeneity exists from case to case and students tend to emotionally favor traditional classrooms (Mackey & Freyberg, 2010). This finding is cited by advocates to justify the use of online videos as a valid instrument for delivering instruction outside the classroom. Second, some have criticized traditional lectures for being overly passive (Gewertz, 2008; King, 2012) and have argued for more widespread adoption of

active learning techniques (Prince, 2004; Michael, 2006). These authors state that in traditional classrooms, learning is a largely unidirectional process of knowledge transfusion from teachers to students. Without adequate opportunities to engage with the material, students tend to concentrate on surface indicators rather than underlying principles (Jaques, 1992), thus neglecting deep learning (Marton & Säljö, 1976). In contrast, active learning techniques encourage productive use of knowledge rather than passive transfer, which in theory leads to better comprehension and retention. Over the years, a number of purported benefits of active learning have been cited, including increased student performance (Michael, 2006; Chaplin, 2009), stronger retention (Dougherty et al., 1995), improved mastery of conceptual reasoning (Crouch & Mazur, 2001), enhanced problem-solving skills (Gijbels, Dochy, Van den Bossche, & Segers, 2005), and greater motivation and general satisfaction with the courses (Colliver, 2000; Newman, 2003). Third, regardless of the numerous claimed benefits of active learning, many researchers have pointed out that the integration of these methods into the classroom is hindered by the pressure to cover a wide variety of topics in an already packed curriculum, leaving little room for innovative practices (Moravec, Williams, Aguilar-Roca, & O'Dowd, 2010; Dove, 2013; Bishop & Verleger, 2013). Flipped instruction is particularly suitable to resolving this dilemma by moving instruction of factual information outside the class to free up class time for deeper processing of course material with more practice and problem solving. It is for this reason that flipped instruction has attracted growing attention from both practitioners and researchers alike.

1.2 Flipped Instruction Defined

Researchers have not yet reached a consensus on a formal definition of flipped instruction. For early pioneers (Bergmann & Sams, 2008), the intent of flipped instruction was to

eliminate lectures in class (i.e., students could watch video podcasts at home before class) so that material that had traditionally been assigned as homework could be completed in class with more student-centered and inquiry-based activities. This idea has led to the general conception of flipped instruction as "events that have traditionally taken place inside the classroom now take place outside the classroom and vice versa" (Lage, Platt, & Treglia, 2000). Bishop and Verleger (2013) suggested including only the studies with computer-based pre-class instruction. We believe qualifying instructional medium is unnecessary and unjustified. It had been shown in a quasi-experimental study with over 800 students that pre-class reading assignments supplemented with worksheets could be as effective as pre-class videos in increasing exam performance using the flipped pedagogy (Moravec, Williams, Aguilar-Roca, & O'Dowd, 2010).

Based on the discussion above, in this study, we define flipped instruction as having three attributes. Flipped classrooms should feature (a) learning of new material before class followed by (b) in-depth explanation, practice, and productive use of knowledge in class through active learning techniques, where (c) both pre-class learning and in-class attendance are mandatory. All three features are necessary. First, pre-class learning is an integral part of instruction. Long before flipped instruction was studied as a distinct pedagogy, instructors were known to assign textbook material for students to read before class. In traditional classrooms, however, pre-class learning was often not enforced and instructors would cover the pre-assigned material in class anyway. In a flipped classroom, pre-class instruction is designated for teaching factual knowledge that will not be repeated in class except for brief reviews. Secondly, productive use of knowledge should dominate class time in order to promote conceptual understanding. Traditional classrooms also adopt active learning techniques. Due to a packed schedule, however, active learning often accounts for a small proportion of the class time. In a flipped classroom, since

lectures are, to a large extent (if not entirely) replaced by active learning activities, class time is largely reserved for productive use of knowledge. Finally, class attendance must be mandatory. In-class instruction is geared towards promoting conceptual understanding, which is a crucially important aspect of learning. Therefore, a "flipped" classroom that adopts an optional attendance policy is not genuinely flipped. It resembles an online class more than a flipped class, since the instruction is already offered online and hence a student can afford not to attend class. In other words, pre-class study and in-class activities should be integrated. They complement each other and are integral parts of learning as a whole.

1.3 Purpose of the Dissertation

The primary purpose of the study is to assess the treatment effect of flipped instruction on student exam performance, as well as on student motivation and perceptions. Any differentiated treatment effect of flipped instruction as moderated by student prior performance (e.g., high school GPA) and demographics (e.g., gender, ethnicity) is also of interest. In addition, I am also interested in how specific measures regarding the implementations of flipped instruction would influence student perceptions and satisfaction. By iteratively implementing, measuring, and improving the flipped pedagogy in the same course, I would like to examine how teaching practices and instructional effects evolve over time.

1.4 Structure of the Dissertation

The dissertation begins with a literature review. It first surveys the theories with regard to the pre-class and in-class components of flipped instruction. Practical implementation issues are discussed highlighting the complexity of the pedagogy and the variety of decisions needed to be made for pre-class and in-class instruction. Most importantly, empirical evidence for the

treatment effect of flipped instruction on student exam performance is thoroughly examined with a strong emphasis on factors that influence the effectiveness of flipped instruction.

The main part of the dissertation is composed of three chapters, each corresponding to one iteration of implementing the flipped pedagogy. The first implementation focuses on (a) comparing the overall out-of-class study time of flipped instruction versus traditional instruction, (b) examining the overall treatment effect of flipped instruction on student exams, (c) studying student's responses to the pedagogy, and (d) exposing potential implementation issues impacting treatment effect.

The second iteration addresses the implementation issues exposed and hence is intended as a replication study to examine (a) overall out-of-class study time, (b) overall treatment effect and differentiated treatment effect moderated by prior performance and demographics, and (c) student responses. It is of interest to know whether previous results can be replicated and if flipped instruction would become more effective as the instructor gains more experience implementing the pedagogy.

The third study builds on what have been found with the previous two studies particularly regarding the implementations issues. By addressing such issues, it is expected that our flipped instruction would become increasingly effective. While the overall main effect and the differentiated impact of flipped instruction on student performance are still of interest, it is important to examine if flipped instruction can exert any influence on student performance in a subsequent course closely related in content to the current course under study.

A concluding chapter summarizes the findings and highlights the implications of the three studies. Practical suggestions are provided based on our experience of implementing the pedagogy for three consecutive years.

References

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

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., & 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.

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

Chaplin, S. (2009). Assessment of the impact of case studies on student learning gains in an introductory biology course. *Journal of College Science Teaching, 39*(1), 72–79.

Colliver, J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine, 75*(3), 259-266.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Dougherty, R.C., Bower, C.W., Berger, T., Rees, W., Mellon, E.K., and Pulliam, E. (1995) Cooperative learning and enhanced communication: effects on student performance, retention, and attitudes in general chemistry. *Journal of Chemical Education 72*(9): 793-797.

Dove, A. (2013). Students' Perceptions of Learning in a Flipped Statistics Class. *In Society for Information Technology & Teacher Education International Conference 2013*(1), 393-398.

Gewertz, C. (2008). States press ahead on "21st century skills.". *Education Week, 28*(8), 21-23.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: a meta-analysis from the angle of assessment. *Review of Educational Research, 75*(1) 27-61.

Jaques, D. (1992). *Learning in groups.* Houston: Gulf.

King, C. J. (2012). Restructuring engineering education: Why, how and when?. *Journal of Engineering Education, 101*(1), 1-5.

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*(1), 30-43.

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.

Mackey, K. R., & Freyberg, D. L. (2010). The effect of social presence on affective and cognitive learning in an international engineering course taught via distance learning. *Journal of Engineering Education*, *99*(1), 23-34.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I - Outcome and process. *British Journal of Educational Psychology, 46*(1), 4-11.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. *US Department of Education*.

Michael, J. (2006). Where's the evidence that active learning works?. *Advances in Physiology Education, 30*(4), 159-167.

Moravec, M., Williams, A., Aguilar-Roca, N., & O'Dowd, D. K. (2010). Learn before lecture: a strategy that improves learning outcomes in a large introductory biology class. *CBE-Life Sciences Education, 9*(4), 473-481.

Newman, M. J. (2005). Problem-based learning: An introduction and overview of the key features of the approach. *Journal of Veterinary Medical Education*, *32*(1), 12-20.

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

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.

Tucker, B. (2012). The flipped classroom. *Education Next, 12*(1), 82-83.

Chapter 2 Literature Review

2.1 Theoretical Framework

In principle, direct benefits of flipped instruction can result from two sources, the flipped structure itself and the various active learning techniques involved. A number of theories have been proposed to validate these two aspects of flipped instruction. Some have found support from research on flipped instruction, while others remain untested. In the following sections, theories underlying flipped instruction are examined and available evidence is cited whenever applicable.

2.1.1 Theories Supporting Pre-class Instruction

Schema theory and cognitive load theory have been proposed to explain the benefit of offloading learning material to the pre-class time period. Schema theory suggests that knowledge acquisition is most robust when new information can be incorporated into existing knowledge networks composed of numerous cognitive constructs called schemas (Bartlett, 1932). People use schemas to organize current knowledge, which provides a framework for future reference and understanding (Anderson, Reynolds, Schallert, & Goetz 1977). Acquisition of schemas, however, is a gradual process. When a concept is first learned, the ability to use it is limited. Applying new concepts requires deliberate thought and controlled processing. With time and effort, controlled processing gradually gives way to automatic processing that occurs without conscious control (Shiffrin & Schneider, 1977). Given our limited working memory (Miller, 1956), schema acquisition and automation have positive implications for learning. Schemas can greatly alleviate the limitations of our working memory by increasing the amount of information that can be held in it through chunking separate elements into meaningful blocks. Meanwhile, automatic processing greatly enhances its efficiency by allowing information to be processed

without conscious intervention to free up working memory for more demanding tasks. Working together, schema acquisition and automatic processing ensure that we can use our limited working-memory efficiently to process new information and over time relay it into our seemingly unlimited long-term memory.

Cognitive load theory (CLT) builds upon the ideas of schema and automation to further explain what determines the difficulty of any learning material. CLT predicts that putting too much information (i.e. high cognitive load) into working memory beyond its loading capacity leads to failures in processing and comprehending the material (Sweller, 1994). CLT further distinguishes between intrinsic and extrinsic cognitive loads (Maybery & Bain, 1986). Intrinsic cognitive load stems from the interaction of elements of useful information contained in any learning material. Extraneous cognitive load is determined by all the extra information provided in context. If many elements of information interact and need to be processed simultaneously, rather than sequentially, intrinsic cognitive load will be high and the material difficult to learn. Moreover, extraneous cognitive load only interferes with learning under conditions of high cognitive load caused by high element interactivity (Sweller, 1994). In flipped classrooms, relocating material outside the classroom gives students the opportunity to build up prior knowledge before coming to class. CLT implies that students who are better prepared before class should perceive the same material to be of lower intrinsic cognitive load and thus easier to understand, since fewer numbers of new schemas are encountered and students are less likely to be impeded by extraneous cognitive load.

Results from pre-lecture and flipped instruction have supported these conjectures. In a psychology course enrolling 162 students, Narloch, Garbin, and Turnage (2006) examined the effect of pre-class quizzes on exam performance in five consecutive semesters. Students from the

pre-class quiz groups were found to be better prepared in general compared to the control group with no quizzes. Moreover, the treatment students significantly outperformed their control counterparts in all exams. Most interestingly, post-course surveys indicated that the treatment students perceived lectures to be clearer, more organized, and more effective in preparing students for exams, despite the fact that lecture contents were similar between the two conditions. Similarly, Stelzer, Brookes, Gladding, and Mestre (2010) used longitudinal data from a flipped classroom to show that pre-class preparation drastically reduced the perceived difficulty of the course, while raising perceived lecture vales even when class time was reduced by 50 minutes per week. These results imply that pre-class learning reduces intrinsic cognitive load by allowing students to build up a stronger knowledge base before class, and use this knowledge to make better sense of in-class activities, leading to less vulnerability to the interference of extraneous information, increased clarity of knowledge structure, and stronger comprehension and exam performance.

In addition, by offloading part of the instruction before class, the flipped structure incidentally induces spaced learning practice. For over a 100 years since Ebbinghaus (1913), numerous studies have shown that given the same amount of time in total, spaced learning leads to better retention compared to massed learning, a phenomenon known as the spacing effect (Donovan & Radosevich, 1999). In flipped classrooms, pre-lecture assignments force students to allocate time for studying before class. Pre-class study increases preparation and thus perceived clarity and usefulness of the class, which in turn leads to increased class attendance (Stelzer et al., 2010; Deslauriers, Schelew, & Wieman, 2011). In addition, when asked to compare the overall study time in flipped classrooms to that in regular classes (Foertsch et al., 2002; Narloch et al., 2006; Papadopoulos & Roman, 2010; Mason, Shuman, & Cook, 2013), students in all but

one study (Papadopoulos & Roman, 2010) reported having spent no more, or even less time, studying in flipped classes. Although the accuracy of students' self-reports was not verified, the findings that the overall study time did not appreciably increase and that students spent more time studying before and inside of class strongly imply that students in flipped classrooms tended to distribute their study time more evenly.

2.1.2 Theories Supporting In-class Active Learning

Active learning activities are an integral part of flipped classrooms. Some even believe that the claimed benefits of flipped instruction largely result from its capacity to accommodate various active learning techniques in the classroom, rather than its inverted format per se (Tucker, 2012). Compared with a small number of pre-class activities, the choice of in-class activities is large, and most flipped classrooms have employed multiple active learning techniques interspersed throughout class time. Since active learning is an umbrella term that encompasses a wide spectrum of instructional techniques, it is impractical to review all available theories proposed to support each technique. In the following, we will only review the theories that deal with active learning in general and interested readers should consult the research literature on each specific technique for further information.

The levels of processing theory explores the link between the intensity of information processing and long-term retention (Craik & Lockhart, 1972). It postulates that the rate of forgetting is a function of the type and depth of elaboration of new information through various levels of processing. The more types and greater depth of information processed, the better the retention. According to the theory, knowledge processing can be broken down into three processes: First, noticing various properties of new information and the conditions under which it applies; second, integrating aspects of the information gained into existing knowledge networks;

and third, consolidating the knowledge gain through repeated further exposure to the same information in varying meaningful contexts. Adequate implementation of the processes will lead to robust knowledge networks embedded in the long-term memory for efficient retrieval. The theory is germane to supporting the use of active learning practices, as it directly suggests that productive use of new information enhances learning and that learning in context is preferable to learning in isolation. In flipped classrooms, students often work with complex problems that necessitate synthesis of scattered knowledge. This process allows students to practice with key concepts repeatedly over time, increases student motivation (Antepohl & Herzig, 1999; Martin, West, & Bill, 2008), and facilitates subsequent knowledge retrieval in similar contexts (Johnson, Ahlgren, Blount, & Petit, 1980; Clement, 1982). The levels of processing theory has long been supported by research on vocabulary instruction (Nelson, 1977; Friederici, 1985).

By employing an array of instructional practices, flipped classrooms are also believed to offer a more inclusive learning environment that caters to diverse student needs. A number of researchers have suggested that students learn with distinctly different learning styles and preferences (Reichmann & Grasha, 1974; Kolb, 1981) and the mismatch between a student and an instructor's styles and personality types could result in lowered student performance (Borg & Shapiro, 1996; Ziegert, 2000) and more negative attitude towards the course (Charkins, O'Toole, & Wetzel, 1985). Given the diversity of student preferences, some have theorized that traditional instruction is ill-suited for delivering optimal instruction, for it adopts a one-size-fits-all approach that heavily relies on lecturing. Regardless of what theory might imply, evidence collected from flipped classrooms showing the benefits of an inclusive environment is flimsy. Admittedly, students do have different preferences when it comes to learning. Some may prefer books; others videos. Some may favor direct instruction. Others do better with self-directed

learning using examples. However, just because a student has a learning preference does not guarantee that the preference is the most effective way for the student to learn. This is the critical link that current studies from flipped instruction fail to address. In one study, for example, Lage, Platt, & Treglia (2000) constructed a table mapping all of the fourteen different instructional methods used in a flipped classroom to the corresponding learning styles featured in four theoretical frameworks on the subject. However, the mapping was done entirely on a conceptual basis and no evidence was shown that a preferred learning style for a student would also be the most effective way for the student to learn the material. Moreover, convincing evidence on the very existence of meaningfully different and stable "learning styles" is lacking, and the concept of learning styles has come under harsh criticism from cognitive psychologists (Reynolds, 1997; Garner, 2000). For example, Garner (2000) claimed that substantial problems existed with Kolb's Learning Style Inventory (LSI) (Kolb, 1984) that is frequently used within many areas of study assigning students to a given learning style. Garner attributed the poor reliability of the LSI measure to its lack of theoretical rigor and examined the contradictions and confusion around whether Kolb was arguing for learning styles as stable traits or flexible states. Garner finally concluded that Kolb's learning style theory "lacks any coherent foundations and clear links to psychology." Given these issues, the claim that students from flipped classrooms would benefit from the diversity of in-class activities is largely speculative.

2.2 Practical Implementation Issues

2.2.1 Practical Issues with Pre-class Instruction

Practical issues exist over the specifics of conducting pre-class instruction. So far, the most commonly adopted media for delivering pre-class activities have been online videos, reading assignments (Deslauriers et al., 2011), and animated PowerPoint slides. Little research

has been done to assess the comparative effectiveness and impact of the three options in flipped settings. Prior research in multimedia and multi-modal instruction suggests that students prefer videos to readings (Day & Foley, 2006; Stelzer et al., 2010), and that multi-modal instruction with graphs and animations leads to better performance than using texts alone (Stelzer, Gladding, Mestre, & Brookes, 2009; Gellevij et al., 2002). However, the only study on flipped instruction that employed a quasi-experimental design to examine this issue found that pre-class reading assignments combined with worksheets was as effective as pre-class videos in increasing student exam performance relative to traditional lectures (Moravec et al., 2010). Since the time and monetary cost of assigning readings is much lower than that of producing videos, carefully designed reading assignments with supplements may be a good method of choice.

Moreover, the use of videos requires considerable input of effort and careful planning. To date, the bulk of published studies on flipped instruction adopted online videos as the standard format for staging pre-class activity. Concerns have been voiced over the time needed for developing the videos. The typical amount of time for producing one hour's worth of videos and accompanying material was estimated to range from 3.7 (Enfield, 2013) to 15 hours (Mason et al., 2013). For covering the same material, the length of the videos, however, can be reduced to about half of the length of lectures by removing administrative announcements, instructor tangents, student questions, and pauses in presentation while writing (Day & Foley, 2006; Mason et al., 2013). If class time is about three hours each week, then the total time of video production will range between 55.5 to 225 hours for a ten-week quarter-long course and 83.25 to 337.5 hours for a fifteen-week semester-long course, which would clearly pose some challenge on the part of the instructors. Using camera-recorded lectures from prior years or videos from online resources is an easy solution. However, camera-recorded videos are not as viewable as

screencasts made from commercial software such as Camtasia or Microsoft Producer. Further, quality videos suitable for specific courses might be difficulty to identity and some have raised concerns that using others' videos might also undermine the authority of junior instructors (Pearson, 2012; Enfield, 2013). In addition, videos are comparatively more difficult to browse through than texts. In some courses where many videos were assigned, students had reported difficulty of finding the right parts of the videos to watch when working on group projects that required multiple elements of knowledge from different videos (Mason et al., 2013). Moreover, the maximum length for each video has been suggested to be no more than twenty minutes (Zappe, Leicht, Messner, Litzinger, & Lee, 2009), which is roughly the duration of an average listener's attention span (Bonwell & Eison, 1991). However, little discussion was held in deciding what the optimal total amount of hours required for pre-lecture study should be.

Non-compliance with pre-class study causes concern especially in less radically flipped classrooms that still reserve some proportion of class time for lectures. Long before flipped instruction becomes a buzzword, instructors were known to assign textbook readings to prepare students, hoping to cover the material in more depth through class discussion rather than introducing new content (Ryan, 2006; Dobson, 2008). In traditional classrooms, since instructors would cover the pre-assigned materials anyway, students' non-compliance with the reading assignments was high (Connor-Greene, 2000) and was getting worse over the years (Burchfield & Sappington, 2000). In flipped classrooms, student resistance to pre-lecture assignments was still reported, especially in the beginning of the course (Herreid & Schiller, 2013), but the issue of non-compliance is alleviated for two reasons. First, instructors often assign for-credit quizzes or exercises to ensure compliance. However, what percentage of total credit should be allocated to provide enough incentive has not been adequately discussed in research literature. Secondly,

some flipped classrooms barely gave lectures and students were explicitly told that class time would not be used to repeat the basics from assigned material, which effectively forced students to learn in advance. However, some instructors would deliberately choose to adopt a less radically inverted classroom by re-teaching some portion of lectures in class, so students would felt less disoriented (Strayer, 2012). Under this circumstance, since lectures are still part of the class, students are most likely not to do pre-lecture study especially when the proportion of allocated credits is low.

2.2.2 Practical Issues with In-class Active Learning

The excitement over flipped instruction has fueled the sentiment that flipped instruction would work simply because voluminous research has shown that active learning works. This optimism is unjustified for two reasons. First, numerous active learning techniques exist and whether all have shown generally positive impact is open to discussion. Second, even if this is true, the heterogeneity of effect sizes with different class settings must not be ignored. For example, problem-based learning (PBL) is one of the most heavily studied active learning techniques, whose efficacy has been examined by a number of meta-analyses (Albanese & Mitchell, 1993; Vernon & Blake, 1993; Colliver, 2000; Gijbels et al., 2005; Strobel & van Barneveld, 2009). Gijbels et al. (2003) found that the overall combined effect size of PBL over traditional lectures approaches zero, while it differed wildly by subject matter. Others also showed the effect of PBL differed greatly depending on the types of questions used (Strobel & van Barneveld, 2009). In fact, Koedinger, Corbett, and Perfetti (2012) has proposed a knowledge-learning-instruction (KLI) framework suggesting that the choice of instructional methods must match the complexity of the learning material and the underlying learning processes involved. The KLI framework implies that the effectiveness of any given teaching

method is not a constant, and that simply applying active learning would not magically improve learning. Instead, it might cause problems by giving more time for bad pedagogies. It also cautions that one must carefully select and implement active learning techniques and be mindful of interpreting results in context.

Since the primary goal of flipped instruction is to engage all students, a range of active learning techniques, from simple to complex, have been invented to tackle this issue. In general, however, simple techniques are easier to implement and are less likely to go wrong. To promote engagement, the simplest way is to ask students to come up with individual solutions first and then turn to their neighbors to share, a technique known as *think-pair-share* (Angelo & Cross, 1993). The virtue of this technique is that it can be used in classrooms of almost any size. Alternatively, several students can form a group either in an ad hoc manner or in a pre-determined way, where each member comes up with a solution and then works together to reach a consensus, which is referred to as *collaborative learning* (Bruffee, 1984). In the event that students within the same group or from different groups are divided between several equally popular solutions, when only one is correct, students are asked to find someone holding opposites views and are given time to convince each other of their choices, a practice known as *peer-instruction* (Crouch & Mazur, 2001). As the complexity of the problems grows, certain problems will require each member making unique contribution to the group so that each student achieves his or her goal if and only if other members achieve theirs, a practice known as *cooperative learning* (Johnson, Johnson, & Smith, 1998). Cooperative learning requires orchestrated group effort; simply assigning students to groups and telling them to work together does not result in cooperative efforts. Cooperative learning also has different degrees of

complexity with simple ones such as *group demonstration* and *role-play* and complex ones such as *problem-based learning* (PBL) and *project-based learning*.

All techniques discussed above need students to work in groups and staging group activities entails making multiple decisions about group size, levels of sharing, and time allotment under the constraints of question difficulty, class size, and overall class schedule. Any mismatch between the decision and the class could cause undesirable outcomes (Miller, Trimbur, & Wilkes, 1994). For example, if the questions are comparatively straightforward, then pairs of students might suffice. Harder problems require more students and probably some planning to match student skills. Time for group work should be allocated to the extent that it is enough for groups to work through the problems, but not too much to induce boredom (Csíkszentmihályi, 1990) and elicit off-topic conversation (Hess, 2004). In addition, having students work in groups might not be as useful as one would expect. An important reason for assigning groups is to pool partial knowledge from individuals to form a more complete knowledge and skill base to produce better solutions. However, psychologists have pointed out that sometimes discussion could be dominated by information that members hold in common before discussion, and by information that supports members' existent preferences (Stasser & Titus, 1985). This implies that after completing group work students could end up knowing exactly what they used to know and having their prior misconceptions strengthened, if they are spending time working only using the knowledge and skills we already possess and unwittingly reinforcing each other's biases. Most importantly, it was shown in flipped classrooms that students actually ranked group discussion in the bottom of the list of preferred teaching practices (Enfield, 2013). Moreover, group activities also makes it hard for instructors to get a sense of students' collective response patterns, which is crucial for adjusting and planning for future instruction.

Depending on the active learning techniques involved, class time can be either highly structured or loosely organized, which has different implications in practice. Frequently discussed in research on classroom response systems (Bruff, 2009), *formative assessment* is particularly helpful in crafting a highly structured class. Formative assessment helps students to understand teachers' expectations, exposes common mistakes and misconceptions, and provides teachers with instantaneous feedback for planning further instruction. In a highly structured class, class time is usually divided into several segments. Each segment focuses on a specific topic. It usually starts with a question and uses formative assessment via clickers to gauge student initial understanding on the subject. Paired or group discussion follows, where students debate and negotiate their solutions. Upon conclusion of discussion, poll again to see the collective patterns of students' updated understanding. If major problems still exist, teachers can intervene to iron things out. This carefully curated class structure is ideal for ensuring that students remain on task and are not easily frustrated for having to decide what to do next.

Contrary to the closely-knit class structure, classes can also be organized in an open form. Open structure is particularly common for inquiry-based activities such as problem and project based learning. In PBL, for example, learning is organized around solving ill-defined real-world problems that involves multiple domains of knowledge. Students learn course contents on a need-to-know basis and seek only to acquire the right amount of knowledge needed to solve current problems. PBL constantly requires cooperative work, for the complexity of the problems necessarily exceeds an individual's ability to solve. When flipped classrooms adopt a PBL-like format, class structure can be regarded as loosely organized, where class time is not carefully planned, students are frequently assigned into groups, the instructor roams among the groups to answer questions only when asked, and students rely on each other for information and learn in

an ad hoc manner from examples rather than direct systematic instruction. Not surprisingly, loosely organized classes require considerable self-discipline and self-directed learning on students' part. This had caused serious concerns in some flipped classrooms (Warter-Perez & Dong, 2012), where students were less motivated to master course material or not adequately prepared with prior education.

2.3 Effects on Student Performance

Previous sections reviewed the theories and practical issues with flipped instruction. In this section, we will examine the overall impact of flipped instruction on student performance and perception using empirical studies that included at least some control. The scope of the search and the inclusion criteria will be discussed first, followed by a brief overview of the general patterns emerged from subsequent review. This section concludes with a detailed discussion of some specific factors influencing the treatment effectiveness of flipped instruction on student performance.

2.3.1 Search Scope and Inclusion Criteria

According to the definition of flipped instruction presented previously, I have expanded the search of literature to include the use of non-video medium for delivering pre-class instruction. The literature search was conducted at two stages. First, electronic databases, i.e. ERIC, PsycINFO, and Web of Science, were searched using a combination of terms including flipped, inverted, pre-lecture, instruction, classroom, and pedagogy. Second, subsequent pedigree search was conducted. Relevant studies identified from the first stage were carefully read to screen for related citations particularly from the introduction and literature review sections. Candidate papers were also entered into Google Scholar to generate lists of other related studies that cited them, from which new studies on the topic could be found.

To qualify for inclusion, studies must (1) use an empirical research design that includes at least a comparison group; (2) be conducted on the post-secondary level; (3) be published in a peer-reviewed journal, conference, master thesis, or doctoral dissertation; (4) be published after the year 2000, and (5) present adequate information that allows computation of effect sizes. Applying this screening scheme produced 34 studies, which are shown in Table 2.1 below.

2.3.2 Overview of Treatment Effects

The majority of the studies were conducted in STEM disciplines with only a few exceptions. The over representation of STEM subjects is not surprising considering that the emergence of flipped instruction is largely driven by the desire to improve teaching quality. STEM disciplines, more than others, are most frequently put under the microscope due to their implications for technological and social progress. In fact, the over representation of STEM subjects in flipped instruction literature parallels that in active learning literature, where a number of techniques, e.g. just-in-time teaching, collaborative learning, and PBL, were actually invented in STEM disciplines.

The identified studies involved both lower and upper level courses from physics, biology, statics, control systems, statistics, and user interaction design. Total sample sizes ranged from 40 up to 1500 students. Some studies only employed flipped instruction for one week, while others implemented it for the entire quarter or semester. Online videos, readings, and PowerPoint slides were the three most commonly used media for staging pre-class instruction. In-class activities ranged from simple techniques such as paired discussion to complicated ones such as PBL and project-based learning. Both open and close class forms were employed and effect sizes differed widely from close to zero up to over 2.5 standard deviations.

Table 2.1 Empirical Studies Examining Effect of Flipped Instruction on Student Performance

First Author, Year	Course	Grade Level	Number of Cohorts	Treatment (Sample Size)	Control (Sample Size)	Effect Size (Cohen's *d*)
Day, 2006	UI Design	Upper Level	1	28	18	0.69
Moravec, 2010	Biology	Lower Level	2	752	430	1.42
Papadopoulos, 2010	Statics	Unknown	1	43	11	0.20
Stelzer, 2010	Physics	Lower Level	8	750	750	0.20
Deslauriers, 2011a	Physics	Freshman	1	211	171	2.50
Deslauriers, 2011b	Physics	Upper Level	2	62	48	1.14
Pierce, 2012	Therapeutics	Upper Level		71	missing	0.86
Bishop, 2013	Numerical Methods	Sophomore	1	55	63	0
Choi, 2013	Software Eng.	Upper Level	1	38	35	0.11
Guerrero, 2013	Mathematics	Unknown	1	15	29	0.20
Lemley, 2013	Thermodynamics	Upper Level	2	15	23	1.02
Mason, 2013	Control Systems	Senior	2	20	20	0.75
McLaughlin, 2013	Pharmaceutics	Professional	2	162	153	-0.13
Morin, 2013	Eng. Programming	Freshman	2	255	237	0
Wilson, 2013	Statistics	Lower Level	2	45	45	0.54
Albert, 2014	Management	Upper Level	2	321	596	0.19
Baepler, 2014	Chemistry	Lower Level	3	375 / 375	350	0.14 & -0.07
Findlay-Thompson, 2014	Intro Business	Unknown	1	30	42	0
Fraga, 2014	English	Unknown	1	25	26	0.36
Ghadiri, 2014	Electronics	Unknown	1	78	50 & 75	0.57 & 0.87
Overmyer, 2014	Algebra	Lower Level	1	136	165	0.22
Rais-rohani, 2014	Statics	Unknown	1	53	57	0.17
Street, 2014	Physiology	Professional	2	177	180	0.29
Willis, 2014	Pre-calculus	Lower Level	2	22	22	-0.03
Winquist, 2014	Statistics	Lower Level	11	53	58	0.36
Wong, 2014	Pharmacology	Professional	2	101	103	0.38
Yelamarthi, 2014	Digital Circuits	Lower Level	2	17	24	0.46
Flynn, 2015	Chemistry	Lower Level	2	398	724	0.11
Hung, 2015	English	Lower Level	1	25	24	1.54
Kennedy, 2015	Calculus	Lower Level	1	77	76	-0.11
Quint, 2015a	Calculus III	Upper Level	1	39	41	0.19
Quint, 2015b	Calculus III	Upper Level	1	35	36	0.51
Schroeder, 2015	Calculus	Lower Level	1	63	49	0.32
Eichler, 2016	Chemistry	Lower Level	1	452	294	-0.07

Although only a handful of empirical studies assessing treatment effect on student exam performance were published before 2012, recent years have seen a surge in the number of such studies. Among the 35 studies, eight studies showed negative or null impact; eleven showed small effect ($0 < d < 0.3$); eleven showed moderate to large effect ($0.3 \leq d < 1.0$); and five showed surprisingly large effect ($d \geq 1.0$). For the eight studies showing negative or null impact, all results were statistically non-significant with the largest negative effect size of -0.114. In other words, one in four flipped classrooms was about as effective as traditional classrooms, and three in four of them would outperform their traditional counterparts.

Closer examination of the studies revealed several patterns. First, studies using authentic settings, i.e. using flipped instruction for an entire quarter or semester and assessing student performance with high-stakes exams, tended to produce smaller effect sizes than those conducted under special circumstances, i.e. using flipped instruction for a short period of time and assessing performance with low-stakes end-of-period tests. Second, compared to small flipped classrooms, larger class tends to be associated with smaller effect sizes. Third, the extent flipped classrooms should be structured was largely determined by student motivation and skills levels. Highly structured classrooms were most suitable for students with lower motivation and weaker study skills, whereas loosely organized classes featuring difficult authentic problems could be successfully implemented with academically mature students driven to master the contents. Fourth, some studies have shown improved treatment effect over time as the instructor had implemented the flipped pedagogy more than once. Fifth, effect of flipped instruction differs by the type of questions used for evaluating performance. Sixth, fewer studies have examined differentiated impacts of flipped instruction as moderated by student demographics and the its treatment effect on subsequent courses were still unclear.

2.3.3 Special vs. Authentic Settings

Several studies had reported surprisingly large effect sizes. Deslauriers et al. (2011) compared two sections of a first-year physics course taken by 538 engineering students at University of British Columbia (UBC). For most of the semester, two experienced professors, one for each section, taught with traditional lectures in a similar manner. Flipped instruction was introduced only in the twelfth week, when a postdoctoral researcher with limited teaching experience replaced one professor. Prior to the flipped treatment, measured scores with student performance in two midterms, conceptual knowledge of physics, attitudes about physics, prior class attendance and engagement were practically identical in the two sections. A twelve-item posttest was administered in both sections during the first class of the following week. Test results showed that the average scores were 41% (*SD* = 13%) in the control section and 74% (*SD* = 13%) in the treatment, which gives a staggering effect size of about 2.5 standard deviations. Moreover, during the twelfth week, student engagement in class nearly doubled and class attendance increased by 20% in the experimental section, whereas both measures remained unchanged in the control. Post-survey also showed that 77% of respondents in the experimental condition agreed that they would have learned more if the whole physics course would have been taught in this new format and 90% agreed that they enjoyed the flipped classroom. These results are interesting in that they suggested that even "novice" instructors, once equipped with the flipped pedagogy, could outperform their senior counterparts in raising student performance, class attendance and engagement.

Although this study is impressive in demonstrating the potential of flipped instruction, its surprisingly large effect size invites suspicion. To start with, the authors themselves cautioned that the immediate posttest primarily reflected the result of learning achieved from pre-class

study and the class itself. Other studies with smaller effect sizes often measured student performance with end-of-term final exams that reflected all the learning done inside and outside of the classrooms. Since learning is a multi-faceted process, it is conceivable that the impact of any single learning channel is likely to be diluted when more opportunities are available for students to acquire knowledge. Therefore, the effect size might have become smaller if measured using an end-of-term authentic final exam. In addition, the large effect size could also result from a sense of novelty introduced by the presence of a young and energetic instructor with a distinctly different teaching style, which might incidentally mobilize the experimental students out of a slumber state. Prior to intervention, class attendance and engagement were as low as about 56% and 45% respectively in both sections. During intervention, attendance and engagement jumped to 75% and 85% in the treatment condition, while remained unchanged in the control. Moreover, nearly 80% of the treatment students took the posttest, compared to only 63% in the control. These differences suggested that the control students remained in a slumber state for having no reasons to change, hence caring less and preparing less for the low-stakes posttest. Were instructors not changed, duration of the treatment extended longer, and the posttest items interspersed in a high-stakes final exam, the effect size could have become much smaller and more realistic.

In fact, some evidence for the above conjecture was provided by research done by the same authors, where the flipped pedagogy was employed for the duration of an entire eleven-week course. Deslauriers and Wieman (2011) compared two cohorts of students taking a quantum mechanics course during summer at the end of their second year at UBC. In summer 2008, 57 control students received traditional instruction taught by a superb lecturer who was a recent recipient of an annual award for teaching excellence. In the following summer, 67

experimental students were taught with flipped instruction by an unspecified instructor. The Quantum Mechanics Concept Survey (QMCS) was delivered one week prior to the final exam as an un-graded mock exam. The QMCS was completed by 48 control students averaging 67% (*SD* = 18%) and 62 treatment students averaging 85% (*SD* = 14%), which gives an effect size of 1.14 standard deviations, still quite impressive but less than half of that presented previously.

Further evidence for the conjecture came from another study that also only partially flipped the class in the equivalent of one week during a ten-week quarter, but used the same instructors and an authentic final exam. Moravec et al. (2010) used flipped instruction in three out of thirty lectures during the fall 2009 in two sections of an introductory cellular biology course, totaling 752 students. Students from one section in a previous year (either fall 2008 or fall 2007) with about 430 students were used as the control. Cohort comparison using SAT math, student demographics and concept assessment in cell biology showed no statistical difference, which was quite convincing considering the large sample size. The same instructor team taught in all three years and the effect of flipped instruction was assessed using six pairs of matched questions from the final exams. Although the study did not directly give an effect size estimate, nor provided detailed statistics for computing one, the average increase in percentage correct was 21.3%, which gives an estimated effect size of 1.42 standard deviations if we assume the original pooled standard deviation to be around 15% (it was 13% from Deslauriers et al.'s 2006 study), which is only a little more than half of the staggering effect size of 2.5. Given the above results, for further research, it is advisable to use the same set of instructors over an extended period of time using high-stakes exams to assess treatment effect, so the results are more useful for guiding practical decisions.

In contrast to prior studies, two studies involving the same instructor(s) in authentic settings showed more moderate impact. Day and Foley (2006) assessed flipped instruction using two sections of the same course, *User Interaction Design*, with 18 control and 28 treatment students. The same instructor taught both sections with identical homework assignments, projects and exams. The two sections were also similar in student demographics and prior GPAs. Most importantly, the instructor adjusted the amount of time required for pre-class learning and canceled seven class meetings for the treatment section to ensure equal overall instruction time. Students' final grades were composed of thirteen lecture homework assignments, three homework assignments, one midterm, one final, and one project. All grading for the two sections was blind, except for the project. Based on the reported statistics, the effect size was estimated to be 1.50 for the overall grade ($p < 0.01$), 0.68 for the midterm ($p = 0.10$), and 0.69 for the final ($p = 0.055$); those for lecture homework and the project cannot be estimated for lack of information. In compliance with our prior recommendation to use high-stakes exams to evaluate effect size, we adopt 0.69 standard deviations (pooling midterm and the final) as the effect size for this study. It should be noted that statistical significance must be viewed along with the magnitude of effect sizes in this study because of the limited sample size.

In the other study, Mason et al. (2013) assessed flipped instruction with two cohorts of students in a senior-level course taken exclusively by mechanical engineering majors. Both cohorts had 20 students and were taught in winter quarters in the same time slot by the same professor using the same textbook and weekly homework assignments in two consecutive years. The class met four days a week and 50 minutes each time. Seventeen problem pairs under seven categories from quizzes, midterms and finals were carefully matched to assess student performance. Cohort equivalence was examined using grades from two prior engineering

courses, college GPA, and number of credits taken; no statistical differences were found. The overall effect size estimated from the presented statistics was 0.75 standard deviations, which is of comparable size to the previous study.

2.3.4 Small vs. Large Classes

Generally speaking, large effect sizes are more likely to be associated with small classrooms. Among the five studies having reported effect sizes greater than one standard deviation, three studies had flipped classrooms with fewer than 70 students, while the other two studies were conducted not in authentic settings. In addition, with flipped classrooms of fewer than 100 students, six studies have reported effect sizes around or larger than 0.5 standard deviations (Day & Foley, 2006; Mason, Shuman, & Cook, 2013; Wilson, 2013; Ghadiri, 2014; Yalamarthi & Darke, 2014; Quint, 2015b).

In contrast, five studies conducted in authentic settings with flipped classrooms of more than 250 students showed consistently smaller effect sizes of less than 0.20 standard deviations (Albert & Betty, 2014; Baepler, Walker, & Driessen, 2014; Morin, Kecskemety, Harper, & Clingan, 2013; Flynn, 2015; Eichler & Peeples, 2016). For example, in an upper level Introductory to Management course with 596 control and 321 treatment students, Albert and Betty (2014) found an effect size of 0.19, the largest effect size among the five studies. Morin et. al. (2013) implemented flipped instruction in the first semester of a first year Engineering Programming course with 255 flipped students while using the course taught face-to-face in the second quarter of the previous year (there was a quarter to semester change) as the control with 237 students. Although students showed positive attitudes towards the flipped pedagogy, the overall treatment effect measured by the final exam is practically zero. Eichler and Peeples (2016) implemented a relatively rigorous quasi-experiment with two sections of the same course taught

by the same instructor in the same quarter in an Introductory Chemistry course. With 294 students in control and 452 students in the flipped classroom, the overall effect size was -0.07 standard deviations.

Two factors might contribute to small effect sizes associated with large classrooms. First, the major advantage of flipped instruction is to stage more active learning in class. Small classes allow the instructors to employ a variety of active learning techniques that involve heavy instructor-student interactions. In large classrooms, however, instructor-student interactions are much difficult to sustain, which thus limits the types and effectiveness of active learning techniques used in class. Second, large classrooms are usually associated with introductory courses where the students enrolled are mostly likely to be freshmen from diverse majors. Compared with juniors and seniors, first-year college students might be less motivated and less academically skilled in self-directed learning both before and in class.

2.3.5 Loosely vs. Closely Structured Class

The way flipped classrooms are structured also has impact on student performance and perception. Strayer (2012) conducted a mixed-method research focusing on student perception of the learning environments between two sections of an introductory statistics course enrolling 26 traditional and 23 flipped students from as diverse as fifteen majors. Student perception was measured by the College and University Classroom Environment Inventory on seven constructs, i.e. personalization, innovation, student cohesion, task orientation, cooperation, individualization, and equity. Results indicated that the two sections differed in their ratings on all seven measures with non-negligible effect sizes (i.e. greater than 0.30), in which five measures favored flipped instruction and two favored traditional format (i.e. task orientation and equity). Most importantly, flipped students felt considerably more disoriented about in-class activities with *ES*

= -0.88 and $p < 0.01$. With the supplement of qualitative data, the author further recommended, "perhaps an inverted classroom is not the preferred design for an introductory course. Many students in an introductory course do not have a deep interest in the subject and could be frustrated when they encounter learning tasks that aren't clearly defined." This observation has important implications. It strongly suggests existence of interplay between student motivation, skills, and in-class activities. As discussed previously, flipped classrooms can adopt a spectrum of activities arranged either in a well-structured or loosely organized fashion. In Strayer's (2012) study, the 23 flipped students in the introductory statistics course came from fifteen different majors, who were mostly likely taking the course to fulfill curriculum requirements. During class, they met in a computer laboratory with no formal lectures provided and were required to use spreadsheet programs to solve data analysis problems. Students were encouraged to seek help from each other and from the instructor. This environment qualifies as a loosely organized classroom. If students do not have the necessary skills or are not motivated to master course material in the first place, flipped instruction adopting complex activities with loosely knit class structure is more likely to encounter problems.

In stark contrast, Mason et al.'s (2013) flipped classroom discussed in the foregoing section also adopted a relatively loose class organization featuring open problems, discussions, and group projects. The class, however, was an upper-level engineering course taken exclusively by senior engineering students. By the fourth week of the quarter, students already felt that flipped instruction was "a better use of class time and that the format better prepared them for engineering practice." Most surprisingly, however, the end-of-quarter survey indicated that none of the seniors believed that the flipped pedagogy would work in first-year courses, since "the freshmen lacked the academic maturity needed to succeed in this setting". Even the authors

conceded that flipped instruction "may be difficult for students who have not developed strong study skills." By contrasts, however, three studies, including two discussed previously, reported success with flipped instruction in lower-level introductory courses (Moravec et al., 2010; Stelzer et al., 2010; Deslauriers et al., 2011), all with predominately positive student attitude. Not surprisingly, a common feature of these courses was the use of highly organized class structure, where class time was divided into segments and each segment started with clickers, preceded to paired discussion or group work, and wrapped up with instructors' feedback and summary. Given the evidence, it is advisable that instructors should choose less demanding activities using highly organized structures when students are not particularly skilled or motivated. Difficult activities in an open classroom should only be attempted when students are motivated to succeed with strong skills sets.

2.3.6 Procedural vs. Conceptual Questions

Several studies have shown that the effect of flipped instruction on student performance differs by the types of questions used to measure performance. With an overall effect size of 0.75 standard deviations, Mason et al. (2013) clustered the questions into seven topics, where three topics were design-related while the others were not. The flipped students performed particularly well on design-based problems ($ES = 1.19$ from presented data and 0.98 from our estimate) versus non-design problems ($ES = 0.58$ from our estimate). This difference was remarkable, as the effect size for design-based problems was almost twice as large as that of non-design based ones. The differentiated impact by question type is understandable once we examine the way flipped classrooms were conducted. In this study, the control class had only five meetings at computer labs solving problems with Matlab's control system software, whereas the treatment

class met almost exclusively there. Given the amount of time flipped students spent on problem solving in labs, it is no surprise that they would outperform their peers in this regard.

In an advanced Statistics course for political science majors, with an overall sample size of 67 students, Touchton (2015) reported an overall effect size of 0.41 standard deviations. Flipped students performed particularly well in more challenging components of the final applied statistics research paper regarding methodology (*ES* = 0.84), evidence and diagnostics (*ES* = 1.17), and research implications and conclusions (*ES* = 1.34). Moreover, flipped students rated the course significantly higher in terms of total course quality, instructor quality, self-assessment of learning, and interest in taking additional methodology courses.

In an upper level Calculus course for engineering majors with 41 control and 39 treatment students, Quint (2015a) found that flipped instruction had stronger impact on conceptual questions (*ES* = 0.47) as compared to procedural ones (*ES* = -0.10). The study was repeated for a second time (Quint, 2015b) and the results showed the consistent pattern that flipped instruction had stronger impact with conceptual questions (*ES* = 0.54) versus procedural ones (*ES* = 0.32).

2.3.7 Prior Grades and Demographics

Although many empirical studies have examined the overall treatment effect, relatively fewer have examined potential moderation of treatment effect by student demographics such as gender, ethnicity, and enrollment year. In a large General Chemistry course, Baepler, Walker, and Drlessen (2014) enrolled students from three cohorts with 350 control and 375 treatment students. With 20 multiple choice questions developed by American Chemical Society's (ACS) Division of Chemical Education Examinations Institute, a small but statistically significant effect was found (*ES* = 0.23 with two-sample *t*-test and *ES* = 0.15 with OLS regression). Students were

then divided into four groups by prior GPA quartiles. No differentiated treatment effects were found across the quartiles. Similarly, Weaver and Sturtevant (2015) enrolled multiple cohorts of students over three semesters in a Chemistry course. While the overall effect size was consistently around 0.42 standard deviations, no differentiated impact was found with quartile analysis based on student grades from a preceding course. These results indicate that flipped instruction benefits student uniformly regardless of prior academic standing.

In contrast, Ran and Reid (2015) conducted a study with two sections of a Chemistry course enrolling 206 control and 117 treatment students. The overall treatment effects were larger in first four exams and much smaller in the final exam (*ES* = 0.18). Quartile analysis based on student exam scores from a preceding semester indicates that flipped instruction benefits only students from the bottom tier while not affecting students from the middle and top tiers, which implies that the actual performance increase received by students from the bottom tier should be greater than the overall effect sizes revealed. This result is interesting as it suggests that flipped instruction has the potential to bridge the achievement gaps.

2.3.8 First-time vs. Multi-time Implementation

Thus far, very few repetition studies were reported, making it difficult to understand whether the effect of flipped instruction might increase over time as instructors gain more experience with implementing the pedagogy. Only one study was found in this regard. Quint (2015a) studied a flipped classroom with 41 control and 39 treatment students in two sections of an advanced Calculus course taught by two instructors. To account for instructor effects, both sections were taught in the traditional face-to-face format for the first one-third of the semester. Identical test was given to both sections and no differences were identified. The treatment section was then switched to flipped instruction. With first exam scores controlled, OLS

regression analysis showed an overall effect size of 0.19 standard deviations for the first-time implementation of the flipped pedagogy. The study was reproduced in the following semester with two sections of the course taught by the same instructors. The sections were comparable in size to the previous ones with 36 control and 35 treatment students. Although flipped students in spring on average performed 17 points lower on the first exam compared to their fall counterparts, they performed higher on the second and third exams. In contrast, traditional students in spring performed lower on all exams compared to traditional students in fall. In the end, flipped students in spring outperformed their control counterparts by 0.51 standard deviations, which supports the claim that flipped instruction can be more effective as instructors gain more experience implementing the pedagogy.

2.3.9 Current vs. Subsequent Performance Outcome

While most empirical studies have focused on examining treatment effect of flipped instruction on student performance in the current course, some have investigated the flipped treatment effect in a subsequent course. Rais-rohani and Walters (2014) found a small positive effect ($ES = 0.17$) of flipped instruction on student final exam performance in a Statics course. However, the effects of flipped instruction on two subsequent courses (i.e., Dynamics and Mechanism of Materials) were practically zero. This result is not entirely convincing owing to the misalignment of subject matters between the three courses.

In contrast, He and Link (2015) has shown that academically disadvantaged flipped students did significantly better as compared to the advantaged flipped students in all three exams by about 0.60 standard deviations in a subsequent organic chemistry course, where the flipped course was the first one in the sequence.

References

Albanese, M.A., and Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine 68* (1) 52-81.

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*(8), 419-424.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal, 14(*4), 367-381.

Angelo, T.A., and Cross, K.P. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers*. San Francisco: Jossey-Bass.

Antepohl, W., & Hezrig, S. (1999). Problem-based learning versus lecture based-learning in a course of pharmacology: A controlled, randomized study. *Medical Education, 33*(2), 106-113.

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.

Bartlett, Frederic C. (1932). *Remembering: a study in experimental and social psychology.* Cambridge: Cambridge University Press.

Bishop, J. L. (2013). *A controlled study of the flipped classroom with numerical methods for engineers* (Doctoral dissertation, Utah State University).

Bonwell, C. C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education, The George Washington University, Washington, DC 20036-1183.

Borg, M. O., & Shapiro, S. L. (1996). Personality type and student performance in principles of

economics. *Journal of Economic Education, 27*(1), 3-25.

Bruff, D. (2009). *Teaching with Classroom Response Systems*. San Francisco: Jossey-Bass.

Bruffee, K. A. (1984). Collaborative learning and the "conversation of mankind". *College English, 46*(7), 635-652.

Burchfield, C. M., & Sappington, J. (2000). Compliance with required reading assignments. *Teaching of Psychology, 27*(1), 58–60.

Charkins, R. J., O'Toole, D. M., & Wetzel, J. N. (1985). Linking teacher and student learning styles with student achievement and attitudes. *Journal of Economic Education, 16*(2), 111-120.

Choi, E. M. (2013). Applying inverted classroom to software engineering education. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *3*(2), 121-125.

Clement, J. (1982). Analogical reasoning patterns in expert problem solving. *Proceedings of the Fourth Annual Conference of the Cognitive Science Society.* Ann Arbor, MI: University of Michigan.

Colliver, J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine, 75*(3), 259-266.

Connor-Greene, P. A. (2000). Assessing and promoting student learning: Blurring the line between teaching and testing. *Teaching of Psychology, 27*(2), 84–88.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior, 11*(6), 671-684.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Csíkszentmihályi, M. (1990), *Flow: The Psychology of Optimal Experience*, New York: Harper and Row.

Day, J. A., & Foley, J. D. (2006). Evaluating a web lecture intervention in a human–computer interaction course. *Education, IEEE Transactions on, 49*(4), 420-431.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 332*(6031), 862-864.

Deslauriers, L., & Wieman, C. (2011). Learning and retention of quantum concepts with different teaching methods. *Physical Review Special Topics - Physics Education Research, 7*(1), 010101-1-6.

Dobson, J. L. (2008). The use of formative online quizzes to enhance class preparation and scores on summative exams. *Advances in Physiology Education, 32*(4), 297-302.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795-805.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* New York: Teachers College, Columbia University.

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.

Enfield, J. (2013). Looking at the impact of the flipped classroom model of instruction on undergraduate multimedia students at CSUN. *TechTrends*, *57*(6), 14-27.

Flynn, A. B. (2015). Structure and evaluation of flipped chemistry courses: organic & spectroscopy, large and small, first to third year, English and French. *Chemistry Education Research and Practice*, *16*(2), 198-211.

Findlay-Thompson, S., & Mombourquette, P. (2014). Evaluation of a flipped classroom in an undergraduate business course. *Business Education & Accreditation*, *6*(1), 63-71.

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.

Friederici, A. D. (1985). Levels of processing and vocabulary types: Evidence from on-line comprehension in normals and agrammatics. *Cognition, 19*(2), 133-166.

Foertsch, J., Moses, G., Strikwerda, J., & Litzkow, M. (2002). Reversing the Lecture/Homework Paradigm Using eTEACH® Web-based Streaming Video Software. Journal of Engineering Education, 91(3), 267-274.

Garner, I. (2000). Problems and inconsistencies with Kolb's learning styles. *Educational Psychology*, *20*(3), 341-348.

Gellevij, M., Van Der Meij, H., De Jong, T., & Pieters, J. (2002). Multimodal versus unimodal instruction in a complex learning context. *The Journal of Experimental Education, 70*(3), 215-239. doi:10.1080/00220970209599507

Ghadiri, K. (2014). Developing and implementing effective instructional stratagems in STEM. In *121st ASEE Annual Conference and Exposition*.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: a meta-analysis from the angle of assessment. *Review of Educational Research, 75*(1) 27-61.

Guerrero, S., Baumgartel, D., & Zobott, M. (2013). The use of screencasting to transform traditional pedagogy in a preservice mathematics content course. *Journal of Computers in Mathematics and Science Teaching*, *32*(2), 173-193.

He, W., & Link, R. (2015). Bridging the achievement gap: A longitudinal study of the differential lingering effects of flipped instruction. AERA 2015 Annual Conference, Chicago, IL.

Herreid, C. F., & Schiller, N. A. (2013). Case studies and the flipped classroom. *Journal of College Science Teaching*, *42*(5), 62-66.

Hess, D. E. (2004). Discussion in social studies: Is it worth the trouble?. *Social Education, 68*(2), 151-157.

Hung, H. T. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning*, *28*(1), 81-96.

Johnson, P. E., Ahlgren, A., Blount, J. P., & Petit, N. J. (1980). Scientific reasoning: Garden paths and blind alleys. In J. Robinson (Ed.), Research in science education: New questions, new directions. Colorado Springs, CO: Biological Sciences Curriculum Study.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). Cooperative learning returns to college what evidence is there that it works?. *Change: the magazine of higher learning, 30*(4), 26-35.

Kennedy, E., Beaudrie, B., Ernst, D. C., & St. Laurent, R. (2015). Inverted pedagogy in second semester calculus. *PRIMUS*, *25*(9-10), 892-906.

Kolb, D. A. (1981). Learning styles and disciplinary differences. In Chickering and Associates, eds., *The modern American college*, 232-255. SanFrancisco: Jossey-Bass.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Englewood Cliffs, NJ: Prentice-Hall.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning.

Cognitive Science, 36(5), 757-798.

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*(1), 30-43.

Lemley, E. C., Jassemnejad, B., Judd, E., Ring, B. P., Henderson, A. W., & Armstrong, G. (2013). Implementing a Flipped Classroom in Thermodynamics. In *120th ASEE Annual Conference and Exposition*.

Martin, L., West, J., & Bill, K. (2008). Incorporating problem-based learning strategies to develop learner autonomy and employability skills in sports science undergraduates. *Journal of Hospitality, Leisure, Sport and Tourism Education, 7*(1), 18-30.

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. *Education, IEEE Transactions on, 56*(4), 430-435.

Maybery, M. T., Bain, J. D., & Halford, G. S. (1986). Information-processing demands of transitive inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(4), 600-613.

McLaughlin, J. E., Griffin, L. M., Esserman, D. A., Davidson, C. A., Glatt, D. M., Roth, M. T., Gharkholonarehe, N. & Mumper, R. J. (2013). Pharmacy student engagement, performance, and perception in a flipped satellite classroom. *American journal of pharmaceutical education*, *77*(9), 1-8.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Miller, J. E., Trimbur, J., & Wilkes, J. M. (1994). Group dynamics: Understanding group success and failure in collaborative learning. *New Directions for Teaching and Learning, 1994*(59), 33-44. doi:10.1002/tl.37219945906.

Moravec, M., Williams, A., Aguilar-Roca, N., & O'Dowd, D. K. (2010). Learn before lecture: a strategy that improves learning outcomes in a large introductory biology class. *CBE-Life Sciences Education, 9*(4), 473-481.

Morin, B., Kecskemety, K. M., Harper, K. A., & Clingan, P. A. (2013). The inverted classroom in a first-year engineering course. *In 120th ASEE Annual Conference & Exposition: Frankly We Do Give a D*mn.* Atlanta, Georgia, USA.

Narloch, R., Garbin, C. P., and Turnage, K. D. (2006). Benefits of prelecture quizzes. *Teaching Psychology, 33*(2), 109-112.

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior, 16*(2), 151-171.

Overmyer, G. R. (2014). *The flipped classroom model for college algebra: effects on student achievement* (Doctoral dissertation, Colorado State University).

Rais-Rohani, M., & Walters, A. (2014). Preliminary Assessment of the Emporium Model in a Redesigned Engineering Mechanics Course. *Advances in Engineering Education, 4(*1) 1-19.

Papadopoulos, C., & Roman, A. S. (2010). Implementing an inverted classroom model in engineering statics: Initial results. *In 117th American Society for Engineering Education. American Society for Engineering Education.* Louisville, Kentucky, USA.

Pearson, G. (2012) Biology teacher's Flipped Classroom: 'A simple thing, but it's so powerful'. *Education Canada, 52*(5), n5.

Reynolds, M. (1997). Learning styles: a critique. *Management learning*, *28*(2), 115-133.

Riechmann, S. W., & Grasha, A. F. (1974). A rational approach to developing and assessing the construct validity of a student learning style scales instrument. *The Journal of Psychology*, *87*(2), 213-223.

Pierce, R., & Fox, J. (2012). Vodcasts and active-learning exercises in a "flipped classroom" model of a renal pharmacotherapy module. *American Journal of Pharmaceutical Education*, *76*(10), 1-5.

Quint, C. L. (2015). *A study of the efficacy of the flipped classroom model in a university mathematics class* (Doctoral dissertation, Teachers College, Columbia University).

Ryan, T. E. (2006). Motivating novice students to read their textbooks. *Journal of Instructional psychology, 33*(2), 135-140.

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.

Schroeder, L. B., McGivney-Burelle, J., & Xue, F. (2015). To flip or not to flip? An exploratory study comparing student performance in calculus I. *PRIMUS*, *25*(9-10), 876-885.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review, 84*(2), 127-190.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4(*4), 295-312.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*(6), 1467-1478.

Stelzer, T., Brookes, D. T., Gladding, G., & Mestre, J. P. (2010). Impact of multimedia learning modules on an introductory course on electricity and magnetism. *American Journal of Physics, 78*(7), 755-759.

Stelzer, T., Gladding, G., Mestre, J. P., & Brookes, D. T. (2009). Comparing the efficacy of multimedia modules with traditional textbooks for learning introductory physics content. *American Journal of Physics, 77*(2), 184-190.

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 15*(2), 171-193.

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.

Strobel, J., & van Barneveld, A., (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-based Learning, 3* (1) 44-58.

Touchton, M. (2015). Flipping the classroom and student performance in advanced statistics: Evidence from a quasi-experiment. *Journal of Political Science Education*, *11*(1), 28-44.

Tucker, B. (2012). The flipped classroom. *Education Next, 12*(1), 82-83.

Vernon, D.T.A., & Blake, R.L. (1993). Does problem-based learning work? A meta-analysis of evaluation research. *Academic Medicine, 68* (7) 550-563.

Warter-Perez, N., & Dong, J. (2012). Flipping the classroom: How to embed inquiry and design projects into a digital engineering lecture. *In Proceedings of the 2012 ASEE PSW Section Conference.*

Weaver, G. C., & Sturtevant, H. G. (2015). Design, implementation, and evaluation of a flipped format general chemistry course. *Journal of Chemical Education*, *92*(9), 1437-1448.

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

Wilson, S. G. (2013). The Flipped Class a Method to Address the Challenges of an Undergraduate Statistics Course. *Teaching of Psychology*, *40*(3), 193-199.

Winquist, J. R., & Carlson, K. A. (2014). Flipped statistics class results: Better performance than lecture over one year later. *Journal of Statistics Education*, *22*(3), 1-10.

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.

Yelamarthi, K., & Drake, E. (2015). A flipped first-year digital circuits course for engineering and technology students. *Education, IEEE Transactions on*,*58*(3), 179-186.

Zappe, S., Leicht, R., Messner, J., Litzinger, T., & Lee, H. W. (2009). Flipping" the classroom to explore active learning in a large undergraduate course. *In Proceedings, American Society for Engineering Education Annual Conference & Exhibition,* Austin, TX.

Ziegert, A. L. (2000). The role of personality temperament and student learning in principles of economics: Further evidence. *The Journal of Economic Education, 31*(4), 307-322.

Chapter 3 First-year Implementation

3.1 Introduction

Although a number of empirical studies have assessed flipped instruction, with some having reported large effect sizes, compelling evidence for the benefit of flipped instruction is relatively scant. Quality empirical studies conducted in authentic settings (i.e., over extended periods of time with high-stakes final exams) using rigorous research designs (i.e., same instructors, large sample sizes, verified equivalent control groups, and overall study time measured) are still needed.

The current study was conducted in two sections of a first-year general chemistry course taught by the same instructor at a large public university in the western United States. With over 300 students enrolled into each condition, the current study is a quasi-experiment conducted to answer the following questions:

(1) Do flipped students spend more or less time studying outside the classroom?

(2) Do flipped students outperform their control counterparts in exams? If so, do flipped students of diverse background benefit equally?

(3) Do flipped students favor flipped pedagogy over traditional instruction?

3.2 Methodology

3.2.1 Course description

This course is the first course in a three-quarter series focusing on the fundamentals of chemistry, e.g., quantum mechanics, atomic structure and bonding, hybridization etc. Having taught the course six times in two years with the traditional lecture format, the instructor used the flipped pedagogy for the first time in the fall 2013. Both sections met three times per week on Mondays, Wednesdays and Fridays. The control class was scheduled from 1:00 to 1:50 p.m., and

the treatment class from 2:00 to 2:50 p.m. To avoid students attending alternate sections, class attendance was mandatory and was recorded via clicker questions, which accounted for 5% of the final grade.

The control section was taught in a traditional lecture format. Although all recommended reading assignments were posted from the beginning of the quarter, no specific reading was assigned before class. Class attendance was recorded using one iClicker question per meeting. The instructor gave lectures using presentation slides and demonstrated problem solving using a document camera. Other than the instructor occasionally pausing for questions, no significant active learning components were involved in the control class. Students completed for-credit homework assignments after class and attended discussion sessions on a voluntary basis.

Students in the flipped section were required to watch about two videos ($M = 1.96$, $SD = 0.932$) before each class meeting and complete associated assignments before class. The videos averaged 10 minutes in length ($M = 10.23$, $SD = 5.22$), most within the range of 5 to 20 minutes (range: 2–23 minutes). There were 49 videos totaling 501.27 minutes overall. The instructor created all of the videos. It took about 4 hours to produce a 10-minute video, including the time needed to prepare assignments associated with each video. To ensure compliance with pre-class preparation, four unannounced quizzes testing on pre-class material were randomly scheduled throughout the quarter. The quizzes accounted for 5% of the total grade.

In class, a typical flipped meeting was divided into three sessions. The instructor briefly reviewed pre-class material and went through each assignment for 10 to 15 minutes. The review did not repeat factual information from the videos, but instead fostered conceptual understanding by drawing connections among concepts. Students were encouraged to ask questions during this time. After the review, the instructor spent another 10 to 15 minutes with two relatively simple

problems. Students worked on the problems individually and submitted their answers via clickers. For the rest of the class time (about 25 minutes), students worked on two to three increasingly difficult worksheet problems. Solving worksheet problems required integration of multiple concepts through multiple steps. The instructor projected the problems onto a big screen and asked students to form ad hoc groups of two or three to solve problems collectively. Meanwhile, the instructor and teaching assistants walked around the classroom and offered help whenever needed. Students submitted answers via clickers and the results were dynamically displayed to the instructor. When the class faltered, the instructor provided more hints, showed polling results to foster discussion, or paused the activities to address common mistakes.

3.2.2 Participants

In total, 781 students were originally enrolled in the two sections, in which 54 were under the age of 18 and were excluded from the study based on our IRB stipulations. Among the eligible students, 343 (93.46%) control and 334 (92.78%) treatment students gave their written consent for participation. Student demographic information was directly collected from the University's Registrar. Table 3.1 shows a detailed breakdown by experimental conditions. In the combined sample, about 60% were female. Students came from 38 different majors and 13 ethnic groups. For convenience, majors were regrouped into 61.00% Biology/Chemistry, 10.49% other STEM (i.e., including all STEM majors except for biology and chemistry related ones), 4.14% Non-STEM, and 24.37% Undeclared. Similarly, ethnicity were regrouped into 11.52% White, 25.55% Black/Latino, 34.71% South Asia including Vietnamese, Thai, and Filipino, 24.67% East Asia including Korean, Chinese, and Japanese, and 3.55% Unstated. Freshmen constituted 86.12% of the students with 10.04% sophomores, 2.22% juniors, and 1.62% seniors. The average SAT math score was 591.54 (SD = 77.32).

3.2.3 Measures

3.2.3.1 Examinations. Two non-cumulative midterms and one cumulative final exam were delivered over the course of ten weeks. All exams were similar in form, comprised primarily of short- and long-answers. Raw scores were converted into percentages for the ease of comparison and interpretation. The treatment section took the midterms after the control section. To avoid cheating, different forms of the midterms were used with isomorphic questions. An identical final exam was given at the same time to the two sections in different rooms. Each midterm made up 23% of the total grade and final exam 36% (homework constituted the remaining 8%).

3.2.3.2 Main surveys. Two main surveys (i.e., a pre-survey and a post-survey) were delivered in the beginning and end of the quarter. The post-survey used in the treatment section is shown in the Appendix. Survey responses were kept away from the instructor and the teaching assistants and were not processed until after the quarter. To encourage participation, 0.5% extra credit was rewarded for completing each main survey. The survey response rates were 88.29% from the control and 86.33% from the treatment group for the pre-survey, and 82.80% and 82.92% for the post-survey.

All survey items were framed on 6-point scales from one (most negative) to six (most positive). An identical pre-survey was delivered to both sections, asking about students' perceived effectiveness of different instructional avenues (i.e., textbooks, lectures, discussion sessions, homework, and learning from peers), and about their general motivation towards this course. Our motivation items were adapted from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1993). Only the items on task value (including interest and utility) and self-efficacy from the MSLQ were used in our study, as

suggested by the expectancy-value theory (Wigfield & Eccles, 2000). Each sub-construct, i.e., interest, utility, self-efficacy, was measured by two items, with the reliability of each sub-construct as measured by Cronbach's alpha above .70 in the pre-survey and above .80 in the post-survey. Finally, the six items were averaged to produce a composite score as a higher order construct to measure general motivation with an overall alpha of .79 and .85 in the pre- and post-survey.

The pre-survey also had two questions asking about the estimated number of hours per week students spent studying before and after class for a typical mathematics or science course. In the post-survey, for the flipped condition, two items measured perceived clarity of instructional videos with a reliability of .86, and three items measured perceived instructional quality in class with a Cronbach's alpha of .88. Another two items measured students' preference of flipped instruction over traditional lectures with an alpha of .90. Two optional open-ended questions asked for students' opinions about the strengths and weaknesses of our implementation of the flipped pedagogy and their suggestions for improvement.

3.2.3.3 Mini-surveys. Ten identical mini-surveys, one for each week, measured students' study time outside the classroom. The mini-survey is included in the Appendix. All mini-surveys were delivered on Mondays. Each mini-survey had only two questions, asking students to give numeric estimates about the number of hours they had spent studying before and after class in the preceding week. To encourage participation, 0.1% extra credit was rewarded for completing each mini-survey. As a result, students could earn up to two extra percentage points for completing all surveys (i.e., two main surveys and ten mini-surveys). Despite the short length of the mini-surveys, the average response rates were considerably lower for the control (M = 68.63%, SD = 7.06%) and treatment (M = 66.77%, SD = 8.38%) sections. It should be cautioned

that replying on students to self-report study hours might result in over-estimation, since students are likely to exaggerate their study effort. Due to concerns with the reliability of study time measures, we did not use them in OLS regression analysis. Instead, study times were used only for between-group comparison purposes. Although the estimated study hours might be biased, the comparison could still be valid as long as students from the two sections have similar propensity for distorting their estimates.

3.3 Results

3.3.1 Group equivalence

To start with, Table 3.1 shows the descriptive statistics by section on survey responses, exam performance, and demographics. Reasonable group equivalence was found by all measures except for major and ethnicity. Chi-squared test showed statistically significant (p = .021) differences in major. It can be argued, however, that the size of the difference is not practically important. The flipped section had 6.51% less STEM majors, 3.09% more Chemistry/Biology majors, 2.47% more Non-STEM majors, and 0.94% more Undeclared majors. Such differences are reasonably small and should not raise serious concerns. Similarly, differences in ethnicity between groups are even smaller and the overall chi-squared test is only marginally significant (p = .069). These small p values are most likely due to our relatively large sample size (i.e., 677 in total), which tends to produce small standard errors that accentuate differences.
extreme values generated rather similar results leading to the same conclusions.

Table 3.1

Descriptive Statistics of Survey Responses, Exam Performance, and Demographics by Section

Measure	Control ($N = 343$)	Treatment ($N = 334$)	t (p) or $\chi^2(p)$	Cohen's d
	M (SD) or Percentage	M (SD) or Percent		
Motivation (pre-survey)	4.63 (0.75)	4.68 (0.83)	0.82 (0.414)	0.064
Before-class Effort (pre-survey)	4.93 (6.01)	4.91 (5.42)	-0.03 (0.977)	-0.003
After-class Effort (pre-survey)	7.27 (7.85)	7.75 (9.55)	0.66 (0.507)	0.055
Textbook (pre-survey)	4.98 (2.82)	5.08 (2.82)	0.43 (0.666)	0.035
Lecture (pre-survey)	4.49 (2.60)	4.71 (2.52)	1.08 (0.281)	0.086
Discussion Session (pre-survey)	4.46 (2.74)	4.59 (2.59)	0.62 (0.532)	0.049
Homework (pre-survey)	4.70 (2.87)	4.79 (2.72)	0.40 (0.692)	0.032
Peers (pre-survey)	4.90 (3.04)	5.24 (2.81)	1.37 (0.170)	0.113
SAT Math	595.23 (75.76)	587.79 (78.81)	-1.21 (0.226)	-0.096
Freshman	86.30%	85.92%	3.91 (0.272)	
Sophomore	10.79%	9.28%		
Junior	1.17%	3.29%		
Senior	1.75%	1.50%		
Chemistry/Biology	59.48%	62.57%	9.69 (0.021)	
STEM	13.70%	7.19%		
Non-STEM	2.92%	5.39%		
Undeclared	23.91%	24.85%		
Female	60.06%	61.26%	0.10 (0.749)	
Black/Latino	23.91%	27.25%	8.71 (0.069)	
East Asia	25.36%	23.95%		
South Asia	32.65%	36.83%		
Unstated	3.21%	3.89%		
White	14.87%	8.08%		
Midterm1	49.90 (21.48)	54.38 (22.80)	2.63 (0.009)	0.202
Midterm2 (non-cumulative)	64.57 (18.79)	63.66 (17.49)	-0.65 (0.515)	-0.050
Final Exam (cumulative)	42.5 (12.51)	43.9 (11.59)	1.51 (0.132)	0.116
Motivation (post)	4.13 (1.14)	4.24 (1.10)	1.22 (0.222)	0.098
Class Quality (post)	4.03 (1.10)	4.05 (1.12)	0.27 (0.785)	0.018
Prefer Flipped Instruction (post)		3.63 (1.54)		

3.3.2 Out-of-class study time

(1) Did flipped students spend more or less time studying outside the classroom?

Table 3.2

Self-reported Out-of-class Study Time in Hours by Section

	Week	Control Mean (SD)	Treatment Mean (SD)	*t*-statistic (*p*)	Cohen's *d*
Before-class	Weeks 1-10	3.71 (3.11)	4.14 (2.76)	1.92 (0.056)	0.146
	Weeks 1-9	3.55 (2.87)	3.98 (2.56)	2.00 (0.046)	0.158
	Week 10	5.12 (5.29)	5.24 (5.17)	0.29 (0.774)	0.023
After-class	Weeks 1-10	6.28 (4.16)	5.69 (3.58)	-1.99 (0.047)	-0.152
	Weeks 1-9	6.00 (4.01)	5.51 (3.47)	-1.65 (0.100)	-0.131
	Week 10	8.07 (6.49)	7.01 (6.09)	-2.00 (0.046)	-0.168
Out-of-class	Weeks 1-10	9.99 (6.64)	9.83 (5.85)	-0.33 (0.742)	-0.026
	Weeks 1-9	9.54 (6.29)	9.48 (5.43)	-0.12 (0.901)	-0.010
	Week 10	13.19 (10.88)	12.26 (10.18)	-1.05 (0.294)	-0.088

Note. Averaged study times were used for producing the estimates. Removing or truncating

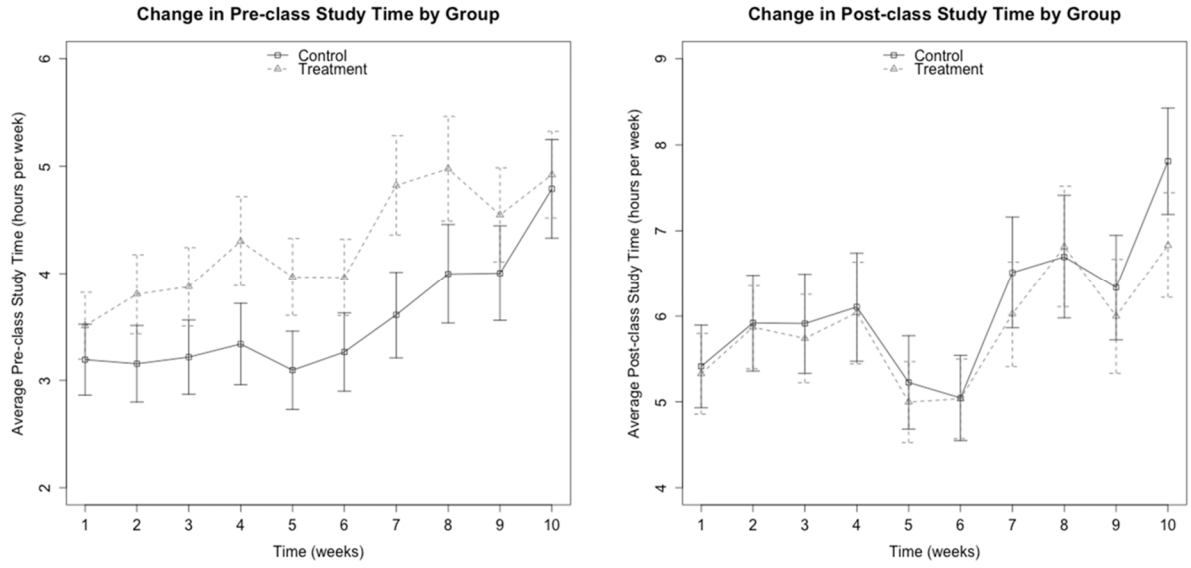

Figure 3.1 Changes in out-of-class study time over the ten-week quarter.

Trajectories of study time are shown in Figure 3.1 with descriptive and test statistics shown in Table 3.2. Effect sizes (*ES*) associated with two-sample *t*-tests are computed using Cohen's *d*. Two local maximums at weeks 4 and 8 correspond to the first and second midterms,

which explains the dips in average study time immediately afterwards. Flipped students on average spent more time before class (*ES* = 0.146, *p* = .056), especially in the first nine weeks (*ES* = 0.158, *p* = .046) and less so during the tenth week (*ES* = 0.025, *p* = .774). In contrast, flipped students tended to spend less time after class (*ES* = -0.152, *p* = .047), particularly during the tenth week (*ES* = -0.168, *p* = .046). Taken together, the overall out-of-class study time was roughly the same across weeks (*ES* = -0.026, *p* = .742). To note, the self-reported study times were highly skewed to the right. Therefore, we have conducted sensitivity analysis by removing or truncating extreme values, which resulted in smaller standard errors leading to the same conclusions with more distinct patterns.

3.3.3 Exam performance

(2) Did flipped students outperform their control counterparts in exams? If so, did flipped students of diverse background benefit equally?

As shown in Table 3.1, a simple two-sample *t*-test suggests that flipped instruction had a small and statistically non-significant effect on the final exam (*ES* = 0.116, *p* = .132). Although the combined sample size is large, due to the small mean difference in the final exam and the magnitude of the standard deviations, the statistical power computed *a posteriori* is only about 33%. In other words, the two-sample *t*-test does not have enough statistical power to detect the small treatment effect as observed. To investigate further, OLS regression models were used to include additional covariates, leading to smaller residual errors and thus more power to detect small differences. Another reason for using OLS regression is to account for minor imbalances as shown previously in group-equivalence check. The results are shown in Table 3.3. To note, all regression coefficients presented in this study are standardized beta coefficients (β) that can be readily interpreted as effect sizes (*ES*). The first two models used the final exam scores as the

Table 3.3

Effect of Flipped Instruction on Exam Performance with OLS Models

	Final Exam		First Midterm		Final Exam	
	Model 3.1	Model 3.2	Model 3.3	Model 3.4	Model 3.5	Model 3.6
(Intercept)	0.302*	0.333**	0.333**	0.388***	0.124	0.043
	(0.123)	(0.124)	(0.125)	(0.114)	(0.096)	(0.060)
Treatment	0.189**	0.192**	0.273***	0.248***	0.022	0.014
	(0.072)	(0.072)	(0.073)	(0.068)	(0.056)	(0.056)
First Midterm					0.626***	0.632***
					(0.033)	(0.031)
Motivation (pre)	0.086*	0.087*	0.046		0.058*	0.061*
	(0.036)	(0.036)	(0.036)		(0.028)	(0.028)
SATmath	0.477***	0.36***	0.340***	0.048	0.148**	0.128**
	(0.044)	(0.065)	(0.066)	(0.130)	(0.051)	(0.048)
Female	0.031	0.026	0.045		-0.001	-0.011
	(0.08)	(0.079)	(0.08)		(0.061)	(0.060)
SATmath×Female		0.197*	0.115		0.124*	0.122*
		(0.081)	(0.082)		(0.062)	(0.062)
STEM	-0.035	0.038	-0.123	-0.154	0.115	0.064
	(0.132)	(0.134)	(0.136)	(0.121)	(0.104)	(0.100)
Non-STEM	-0.625**	-0.552*	-0.487*	-0.508*	-0.248	-0.316*
	(0.231)	(0.232)	(0.235)	(0.223)	(0.180)	(0.157)
Undeclared	-0.312***	-0.302***	-0.336***	-0.297***	-0.094	-0.108
	(0.089)	(0.088)	(0.089)	(0.085)	(0.069)	(0.069)
Sophomore	0.101	0.098	0.292*	0.283*	-0.085	
	(0.118)	(0.118)	(0.119)	(0.117)	(0.092)	
Junior	-0.113	-0.176	-0.140	-0.067	-0.089	
	(0.312)	(0.312)	(0.316)	(0.312)	(0.241)	
Senior	0.332	0.290	0.926*	0.628*	-0.290	
	(0.387)	(0.386)	(0.391)	(0.312)	(0.300)	
Black/Latino	-0.488***	-0.502***	-0.657***	-0.680***	-0.093	
	(0.131)	(0.131)	(0.133)	(0.134)	(0.104)	
South Asia	-0.140	-0.161	-0.176	-0.243*	-0.051	
	(0.125)	(0.124)	(0.126)	(0.122)	(0.096)	
East Asia	-0.494***	-0.493***	-0.487***	-0.542***	-0.189+	
	(0.131)	(0.131)	(0.132)	(0.134)	(0.102)	
Unstated	-0.260	-0.245	-0.484*	-0.621**	0.057	
	(0.229)	(0.228)	(0.232)	(0.218)	(0.177)	
SATmath×Black/Latino				0.383*		
				(0.152)		
SATmath×South Asia				0.526***		
				(0.146)		
SATmath×East Asia				0.301*		
				(0.147)		
SATmath×Unstated				0.260		
				(0.201)		
Cases	555	555	556	635	555	555
Adj.R-squrared	0.305	0.311	0.3	0.3	0.588	0.588
AIC	1391.35	1387.22	1404.74	1608.96	1103.18	1096.07

Note. All estimates are standardized beta coefficients. Standard errors are in parentheses.

$^{+} < .10$, $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$

dependent variable. Model 3.1 is the main effect model, which includes all statistically significant covariates without interaction terms. It shows a small but statistically significant treatment effect ($\beta = 0.189$, $p = .010$). From Model 3.1, potential interaction terms were studiously explored and the only significant interaction found is between SAT math and gender ($\beta = 0.197$, $p = .015$), which is shown in Model 3.2 with larger adjusted R square and smaller AIC index. Post-regression diagnostics on Model 3.2 did not show noticeable violations of OLS assumptions; nor were any extremely influential cases identified. Therefore, Model 3.2 is accepted as our final model.

Closer inspection of Table 3.1 indicates that the treatment effect on exam performance was the strongest with the first midterm, disappeared by the second, and rebounded slightly with the final exam. It seems that the net benefit of flipped instruction diminished over time. As a result, Models 3.3–3.6 were fitted to investigate the mediating effect of the first midterm. Conceptually, the treatment effect on performance in Model 3.2 can be regarded as the total effect. We want to examine if the first midterm can be treated as a mediator. If so, we should see significant impact of treatment effect on the first midterm, but non-significant impact of treatment effect on the final exam with the first midterm included as a covariate.

Model 3.3 resembles Model 3.2 in all aspects except that the dependent variable was the first midterm scores. Model 3.4 improves upon Model 3.3 with significant interaction terms added and non-significant covariates removed. Models 3.5 and 3.6 used final exam scores as the dependent variable with the first midterm included as a covariate. Model 3.5 shows that (a) doing

well in the first midterm strongly correlated with the outcome in the final ($\beta = 0.626$, $p < .001$), (b) the direct effect of flipped instruction on final exam scores while controlling for the first midterm was negligibly small ($\beta = 0.022$, $p = .696$); and (c) the total effect of flipped instruction ($\beta = 0.192$) as shown in Model 3.2 was primarily due to the mediating effect of the first midterm (i.e., $0.157 = 0.248 \times 0.632$).

3.3.4 Perception and attitude

(3) Did flipped students favor the flipped pedagogy over traditional instruction?

All participating students in the treatment condition were asked to rate their attitudes towards flipped instruction in the post-survey. The overall flipped students' attitude was distinctly lukewarm. Students' ratings on their preference of flipped instruction over traditional lectures averaged 3.631 ($SD = 1.538$) on a 6-point scale, which is not significantly different from the neutral position of 3.50 ($p = .156$). A histogram of the preference measure reveals a broadly bell-shaped distribution with the global maximum around the mean and two local maximums peaking at the two extremes. This result indicates that opinions towards the flipped pedagogy were somewhat bipolar, as there were about as many students who strongly favored flipped instruction as those who fiercely opposed it (i.e., 30 students or 10.83% rated six and 34 students or 12.27% rated one on the 6-point scale).

To gain insight as to whether prior motivation and student demographics associate with perceived class quality and instructional clarity, we applied OLS regression and the results are shown in Table 3.4. With all students included, Model 4.1 shows that flipped instruction did not

raise perceived in-class quality. In contrast, students with stronger prior general motivation (β = 0.312, $p < .001$) and higher SAT math scores (β = 0.217, $p < .001$) tended to perceive in-class activities to be of higher quality. Moreover, non-Chemistry or non-Biology majors generally gave lower ratings. Likewise, flipped students' perception of video clarity demonstrated similar patterns of association, as students with higher SAT math scores and motivation rated instructional clarity more positively. Although major is not a statistically significant predicator, it is kept in the model for the sake of comparison. To note, the size of the coefficients for major in Model 4.2 is about as large as that in Model 4.1. The halved sample size increased standard errors, which might push the coefficients out of significance.

Table 3.4

Perceived In-class Quality for All Students and Video Clarity for Flipped Students

	Model 4.1 (Class Quality)	Model 4.2 (Video Clarity)
(Intercept)	4.117***	4.276***
	(0.08)	(0.099)
Treatment	-0.018	
	(0.101)	
SATmath	0.217***	0.231*
	(0.055)	(0.089)
Motivation (pre-survey)	0.312***	0.175*
	(0.052)	(0.084)
STEM	-0.348*	-0.496
	(0.177)	(0.366)
Non-STEM	-0.518+	-0.355
	(0.284)	(0.395)
Undeclared	0.039	0.070
	(0.125)	(0.208)
Cases	465	226

Note. All estimates are standardized beta coefficients. Standard errors are in parentheses.

+ < .10, * $p < .05$, ** $p < .01$, *** $p < .001$

In the flipped section, 82.92% of the students responded to the post-survey, and about half of them opted to express their likes and dislikes about the flipped pedagogy. The comments have offered valuable insights. Not surprisingly, student comments echoed their bipolar ratings. Some expressed unqualified approval, while others showed bitter resentment. Specifically, positive comments have confirmed some of the proposed benefits of flipped instruction, which include (a) the flexibility of *"watching videos at my own pace, whenever, wherever, and however many times I would like"*; (b) improved preparedness for class, as it is *"a means to make lecture portion much easier to understand"* and "*helping me feel more confident for knowing what the lecture is about, so you can just accumulate knowledge from that point"*; and (c) capacity to accommodate more elaboration, application, and teacher-student interaction, since "*by getting the dry stuff out of the way, the inverted method enables more teacher-student time during the classes,"* and *"we were able to do more practice problems in class and that really cleared things up for me."*

On the other hand, those who opposed flipped instruction voiced strong criticism. Some students were accustomed to the old ways and resented changes. *"When I attend a lecture, I wanted a professor to actually teach me the content. If I wanted to learn chemistry online then I could YouTube it myself rather than someone telling me, but since I'm paying for the course I feel that it would be more suitable for the professor to lecture during class."* Some students did not understand that the pre-class and in-class instructions are inseparable and could not see the value of instructor-guided problem solving in class. *"I did not like it at all, the inverted way made me feel like it was a waste of time to even go to class. I can do practice problems on my*

own time." The bulk of the criticism was leveled at video and class related issues; *"I disliked the voice quality of the videos. Some online videos had a very muffled voice from an older microphone"; "Sometimes it felt like some things were rushed or that we spent too much time on one thing";* and *"The practice problems in class seem to be easier compared to midterms' questions. I would have appreciated more exam style practice problems during lecture with the same level of difficulty of an exam."*

3.4 Discussion

3.4.1 Out-of-class study time

Several studies have reported without sufficient proof that students claimed to have spent no more, or even less time, studying in flipped classrooms (Foertsch, Moses, Strikwerda, & Litzkow, 2002; Mason et al., 2013; Narloch et al., 2006), except for one study reporting the opposite (Papadopoulos & Roman, 2010). Our data verified the claim that flipped instruction does not appreciably increase the overall workload of the students. This suggests that to assess the effectiveness of flipped instruction, it may not be necessary to adjust additional pre-class study time by reducing the number of class meetings (Street et al., 2015). The result also implies that any treatment effect on exam performance should be attributed to factors other than mere increase in study effort. One possibility is that flipped students might benefit from spaced learning, since they allocated their study time more evenly than regular students did. Given the same amount of time, spaced learning leads to better retention of information, a phenomenon known as the spacing effect (Donovan & Radosevich, 1999).

3.4.2 Exam performance

Our regression models suggest a small but statistically significant ($\beta = 0.192$, $p = .008$) effect of flipped instruction on student final exam performance. Compared to the effect sizes reported from other studies (Deslauriers, et al., 2011; Deslauriers & Wieman, 2011; Moravec et al., 2010), our result is appreciably smaller. We believe, however, an effect size of about 0.2–0.4 is much more likely in practice than some large effect sizes previously reported. As discussed earlier, studies conducted in relatively short periods of time using immediate end-of-term, low-stakes tests could arguably be more likely to produce more favorable results for several reasons. First, the novelty induced by a distinctly different instructional technique could temporarily intrigue and motivate students. Second, in shorter time periods, fewer things could possibly go amiss and hence make it more likely for a complex instructional technique such as flipped instruction to work, which entails making multiple decisions on pre-class and in-class components. In a flipped classroom, for example, an instructor need to consider the number and length of videos, accompanying practice questions, pre-class quizzes, percentage of lectures retained in class, the number and kinds of in-class active learning activities to adopt, and different ways to conduct them. The more decisions to make, the more it is likely that some steps can go wrong. Third, using immediate, low-stakes tests make it more difficult for other compensating mechanisms to work, hence overestimating the impact of flipped instruction. The final exam often accounts for the most of a student's grade. Therefore, students in both treatment and control conditions will do extra activities in their own time to prepare for it, which will somewhat drown out any positive impact brought about by flipped instruction. The residual

impact, after alternative overcompensating learning mechanisms are allowed to take effect, is therefore a more meaningful, practical measure of the overall effect of flipped instruction. As a result, we recommend using long-term, high-stakes, tests and hence regard results from cumulative final exams more highly than those from midterms.

In fact, our study has already demonstrated the possibility of diminishing treatment effect over longer time period: The effect size of flipped instruction from two-sample *t*-tests with the first midterm was about twice as large as that with the final exam. In addition, our regression results regarding the potential mediating effect of the first midterm on final exam performance indicate that the treatment effect on the final exam was largely attributed to the lingering effect of flipped instruction on the first midterm. In other words, flipped instruction contributed little to accruing benefits after the first midterm, which is partly supported by the practically null effect (ES = -0.050, p = .515 with two-sample *t*-test) by the second non-cumulative midterm.

Moreover, studies similar in design to ours (i.e., assessing flipped instruction using end-of-quarter or semester finals with large sample sizes) have reported comparable effect sizes (Street et al., 2015, Wong, Ip, Lopes, & Rajagopalan, 2014). On the other hand, it should also be noted that this was the first time our instructor had implemented flipped pedagogy. Since the instructor had spent most of the preceding summer (totaling about 200 hours) developing instructional videos, there was less time devoted to preparing materials such as pre-class assignments and quizzes that would help ensure compliance. It is possible that the lack of enough for-credit quizzes as incentives was accountable for non-compliance, which in turn lessened treatment effect in our case.

3.4.3 Perceptions and attitudes

Our results have shown that treatment students had mixed, polarized feelings about flipped instruction. Moreover, Table 3.4 implies that generally more motivated and academically well-prepared students might be more receptive to flipped instruction, as they tended to perceive the class to be of higher quality and the instruction of greater clarity. By closely examining students' comments, we began to understand some weaknesses of our implementation of the flipped pedagogy, which is helpful for understanding previous results.

Specifically, the essence of flipped instruction is to move certain instructional material outside the classroom to free up class time for problem-solving and teacher-student interaction. Its success, therefore, critically hinges on the success and effectiveness of pre-class study and in-class active learning activities. In our case, non-compliance with pre-class study was a serious issue. Three causes are identified. The first is habitual resistance. Many students commented that they were completely new to flipped instruction and were not fond of learning before class. Some customarily associated learning with instruction and firmly believed that "*being explicitly taught step-by-step is the most efficient way to learn*". Many could not see the critical importance of pre-class study and regarded it as "*extra work*" they were forced to do rather than a mere shift in workload. Second, time management skill and self-discipline are needed. One highly desired appeal of online videos is immortalized in the reprise, *anytime, anywhere, at any pace*. The irony is that having videos readily accessible any time online induces procrastination. In traditional classrooms, class time was designated for learning. In flipped classrooms, however, students had to decide when to study on their own and some simply lacked the self-discipline to do so in a

timely manner. When they studied online, some were easily distracted and lured away by other websites. Occasional lapses were a factor particularly during the second half of the quarter. For each class meeting, the minimum study time was about 20 minutes to simply play the videos. As time went on, the material became more challenging and pre-class study time multiplied. Meanwhile, other courses also became increasingly demanding towards the end, making unintended non-compliance more likely to happen. Once students missed watching the videos, attending class would not be as helpful. Some were clueless during group discussion and frustration led them to simply sit out the time in class rather than using the time to catch up. For whatever reasons, when students failed to adequately prepare before class and hence fared poorly in class, complaints would ensue and radiate in multiple directions (e.g., video examples could be more difficult; instructor should review video contents in class; instructor went through solutions too fast). By implication, non-compliance seemed to disproportionately affect students with low motivation, poor self-discipline, and weak time-management and academic skills.

The negative impact of non-compliance from some students rippled to affect others as well. While reading the comments, we identified an interesting association: If a student reported having enjoyed the videos and benefited from studying before class, the student would often lament the lack of enough challenging practice problems, complaining that the class time was not productively spent. In contrast, a major complaint voiced by the control students was that the instructor was too fast, too rushed, and covered too much content. Therefore, the question is why the instructor would not simply give more challenging practice problems to the treatment students. As one flipped student had observed, *"During office hours, Dr. B would take aside 10-*

20 students and teach them. Students were quiet and attentive while she answered their questions. Dr. B's explanations were concise and she was able to answer the follow-up questions in-depth. During lectures, Dr. B was forced to water down her explanations so that the least informed of the four hundred students could probably understand. The lectures usually went at the speed of the slowest student."

We believe that non-compliance is not only the root cause for various complaints, but also sheds light on the overall small treatment effect, absence of marked interaction, and diminishing treatment effect. Non-compliance affected both under-prepared and well-prepared students, as it made class activities less useful for the under-prepared while limiting the amount and difficulty of the practice problems that could have been solved in class. In other words, non-compliance hinders flipped instruction from reaching its full potential. The overall treatment effect was hence small.

With regard to interaction, we believe that an interaction effect occurs when the treatment conditions clearly agree with the characteristics of a specific subgroup. Others with characters departing from this niche group in varying degrees consequently enjoy the benefits to lesser extents. In our study, since neither well-prepared nor under-prepared students had enjoyed the full benefits of flipped instruction, no obvious niche group could be identified and hence no marked interaction was detected.

We also believe that non-compliance may have been amplified by the class composition, which included 86% freshmen. Due to the time spent on developing videos, the instructor did not create enough for-credit quizzes to ensure compliance. It is conceivable, when left to their own

devices, freshmen might be particularly vulnerable, since they were likely to have poor self-discipline, weak time-management and learning skills, compared to their more senior counterparts. Towards the end of the quarter, unintended non-compliance might become increasingly common, which in turn would translate into diminishing treatment effects in later weeks. Moreover, student attitudes did not remain static, but tended to grow towards the extremes over time, which eventually manifested as polarized feelings in the post-survey.

The conjecture that freshmen might be less receptive to flipped instruction has also been suggested elsewhere. Mason et al. (2013) assessed flipped instruction in an upper-level engineering course taken exclusively by seniors. By the fourth week of the quarter, students already agreed that flipped instruction was "a better use of class time". Most surprisingly, however, the end-of-quarter survey indicated that none of the 20 seniors believed that the flipped pedagogy would work in first-year courses, since the freshmen "lacked the academic maturity needed to succeed in this setting". As a result, the authors conceded that flipped instruction "may be difficult for students who have not developed strong study skills."

3.5 Limitations

While response rates to the main surveys were reasonably high, the weekly mini-surveys had much lower rates, averaging 67.7% (SD = 7.7%). The low response rates could result from the absence of reminder emails. For main surveys, response rates always jumped at the prompt of each remainder email. Since mini-surveys were delivered weekly, sending remainder emails was considered overly intrusive. If frequent responders to mini-surveys were more likely to possess

greater self-discipline and time management skills essential for adequate pre-class study, our measured study times would be biased.

In addition, despite our efforts to ensure students that survey responses would not be analyzed until after the quarter, students could still be motivated to exaggerate their study effort, a cause for upwardly biased study times. Although students with different characteristics might have varying propensities for over-estimation, between-group comparison of overly estimated study times might still be valid as long as students from the two sections have comparable characteristics and hence similar propensities for over-estimation. Future studies should consider alternative ways to measure out-of-class study time more accurately.

External validity is also a matter of concern. Student demographics in our study were over 60% Asian, about 12% Caucasian, and less than 2% Black or African-American in the combined sample. This composition is clearly different from many other institutions in the US and across the globe. It is unclear how much the ethnicity mix would impact the generalization of our results and findings.

3.6 Conclusions and Implications

Flipped instruction did not increase students' overall study time; it only caused a shift in student workload. By implication, any impacts of flipped instruction should be attributed to factors other than mere increase in study effort. Moreover, to assess flipped instruction, it might be unnecessary to adjust additional pre-class study time by reducing the number of class

meetings. By allocating study time more evenly, treatment students might in theory benefit from spaced learning.

We believe measuring student performance using long-term, high-stakes exams gives more practically meaningful results. With flipped instruction implemented for the first time, our OLS models showed a small, but statistically significant, treatment effect ($ES = 0.192$, $p = .008$) with the final exam. No marked interaction was identified, indicating flipped instruction benefited all students equally. The overall treatment effect was more pronounced in the beginning, but diminished over time.

Flipped instruction did not increase student motivation and perceived overall class quality. Treatment students' preference of flipped instruction over traditional lectures was lukewarm with about one fifth of the students displaying polarized feelings. Prior motivation and SAT math scores were positively associated with perceived class quality and instructional clarity, which suggests that highly motivated and academically well-prepared students might be more receptive to flipped instruction.

Positive student comments confirmed some proposed benefits of flipped instruction, including learning at one's own time and pace, better preparation for class, and more problem solving and teacher-student interaction. Pre-class study non-compliance was a serious issue in our study and was believed to be closely associated with negative student attitudes. On students' part, three possible causes of non-compliance were identified, i.e., habitual resistance, procrastination and distraction, and unintended lapses. By implication, non-compliance seemed to disproportionately affect students with poor self-discipline, low motivation, and weak time-

management and academic skills. Moreover, non-compliance seemed to affect all students, as under-prepared students tended to have difficulty following the class while well-prepared students reported boredom and demanded for more challenging practice problems during the class. As a result, the overall treatment effect may have diminished and no marked interaction was detected (since no niche group had enjoyed the full benefits of flipped instruction). The predominance of freshmen in the class may also lead to diminished treatment effect. Since not enough for-credit quizzes were created to ensure compliance, freshmen were particularly vulnerable due to weak self-disciple, time-management, and learning skills. Unintended lapses became increasingly frequent towards the end of quarter, leading to diminished treatment effect.

In summary, our implementation of the flipped pedagogy caused a shift in student workload from post-class to pre-class study without appreciably increasing the overall amount. Flipped instruction slightly, but uniformly, increased final exam performance for all subgroups of students. However, it did not increase student motivation and perceived overall class quality with about one fifth of the students showing polarized feelings. Non-compliance to pre-class study, lack of enough pre-class quizzes, and the predominance of freshmen in class are believed to result in the diminished treatment effect and absence of marked interaction. Future practitioners of flipped instruction should take measures to ensure pre-class study compliance, particularly in large introductory undergraduate courses, as the success and effectiveness of flipped instruction critically depends on both the pre-class and in-class components.

Appendix

Post-survey (for Treatment Section)

1. Based on your learning experience in this quarter, please rate the following items.

Please rate your agreement with the following statements	Strongly Disagree Strongly Agree
I am very interested in the content area of this course.	1 2 3 4 5 6
Beyond this quarter, this course will still be useful to me.	1 2 3 4 5 6
I am confident that I will do well in this course.	1 2 3 4 5 6
I find studying the course material enjoyable.	1 2 3 4 5 6
I will need the contents from this course in subsequent courses	1 2 3 4 5 6
Given my current situation, I am confident of getting a good grade.	1 2 3 4 5 6
Professor's online videos were crystal clear to me.	1 2 3 4 5 6
Professor's in-class instruction was crystal clear to me.	1 2 3 4 5 6
I prefer this inverted class format to a traditional "lecture" format.	1 2 3 4 5 6
I would prefer to take more science classes using this type of class format.	1 2 3 4 5 6

Please rate the overall quality of the following items	Poor Excellent
online component of the instruction	1 2 3 4 5 6
in-class component of the instruction	1 2 3 4 5 6
this course as a whole	1 2 3 4 5 6

2. Open-ended Questions (optional)

2.1. How do you like or dislike about the "inverted" method of teaching this course?

2.2. What is your major complaint about this course and how do you recommend us to improve?

References

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

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

Bourne, J., Harris, D., and Mayadas, F., (2005). Online engineering education: learning anywhere, anytime. *Journal of Engineering Education, 94*(1), 131-146.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Day, J. A., & Foley, J. D. (2006). Evaluating a web lecture intervention in a human–computer interaction course. *Education, IEEE Transactions on, 49*(4), 420-431.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 332*(6031), 862-864.

Deslauriers, L., & Wieman, C. (2011). Learning and retention of quantum concepts with different teaching methods. *Physical Review Special Topics - Physics Education Research, 7*(1), 010101-1-6.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795-805.

Foertsch, J., Moses, G., Strikwerda, J., & Litzkow, M. (2002). Reversing the lecture/homework paradigm using eTEACH® web□based streaming video software. *Journal of Engineering Education, 91*(3), 267-274.

Herreid, C. F., & Schiller, N. A. (2013). Case studies and the flipped classroom. *Journal of College Science Teaching*, *42*(5), 62-66.

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*(1), 30-43.

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. *Education, IEEE Transactions on, 56*(4), 430-435.

McDonald, K., & Smith, C. M. (2013). The flipped classroom for professional development: part I. Benefits and strategies. *The Journal of Continuing Education in Nursing*, *44*(10), 437.

Moravec, M., Williams, A., Aguilar-Roca, N., & O'Dowd, D. K. (2010). Learn before lecture: a strategy that improves learning outcomes in a large introductory biology class. *CBE-Life Sciences Education, 9*(4), 473-481.

Narloch, R., Garbin, C. P., and Turnage, K. D. (2006). Benefits of prelecture quizzes. *Teaching Psychology, 33*(2), 109-112.

Papadopoulos, C., & Roman, A. S. (2010). Implementing an inverted classroom model in engineering statics: Initial results. *In 117th American Society for Engineering Education. American Society for Engineering Education.* Louisville, Kentucky, USA.

Pintrich, P. R., Smith, D. A., García, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and psychological measurement*, *53*(3), 801-813.

Stelzer, T., Gladding, G., Mestre, J. P., & Brookes, D. T. (2009). Comparing the efficacy of multimedia modules with traditional textbooks for learning introductory physics content. *American Journal of Physics, 77*(2), 184-190.

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.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary educational psychology*, *25*(1), 68-81.

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.

Chapter 4 Second-year Implementation

4.1 Introduction

The current study is a follow-up to our work during the first year of this study. Our prior study showed a small and statistically significant treatment effect ($ES = 0.192$, $p = 0.008$). Student survey responses revealed non-compliance to pre-class study as a major implementation issue that we believe led to the small treatment effect and lack of interaction between treatment effect and student demographics or prior performance. In theory, pre-class learning is critical, as it arguably promotes spaced learning and better prepares students for class due to reduced cognitive load. With pre-class learning properly implemented, students could also benefit from deeper levels of processing due to the variety of active learning techniques employed in class. Therefore, non-compliance with pre-class study could potentially seriously undermine the effectiveness of flipped instruction. As a result, it is of critical importance to investigate effective means to ensure compliance and measure the resulting treatment effects.

The primary goal of this study is to continue our quest to measure overall treatment impact and explore moderation effects. It is of interest to see whether including pre-class for-credit quizzes would provide enough incentive to ensure compliance. Moreover, we are also attentive to students' perception of the flipped classroom and to any further implementation issues. Finally, our prior study indicated that flipped instruction caused a shift in workload from post-class to pre-class without appreciably changing the overall study time. This study will check

if the result is reproducible. Hence our current study intends to answer the following research questions:

(1) Did flipped students comply with pre-class study requirement and did they spend more or less time studying outside the classroom?

(2) Did flipped instruction increase student exam performance and motivation? If so, did students of diverse background benefit equally? Did flipped instruction have sustained impact on student overall performance in a subsequent course?

(3) Did flipped instruction impact perceived overall class quality? Were there further implementation issues?

4.2 Methodology

4.2.1 Course description

The present study was conducted in fall 2014 in two sections of a first-year general chemistry course taught by the same instructor at a large public university in the western United States. Previously, the instructor has taught the course seven times in three consecutive years using a traditional lecture format. Flipped instruction was implemented for the first time in fall 2013. In fall 2014, a new cohort of 607 students was enrolled into two sections. Both sections met three times a week on Mondays, Wednesdays and Fridays for ten weeks. The control class was scheduled from 1:00 to 1:50 pm, and the treatment class from 2:00 to 2:50 pm. To avoid students taking alternative sections, class attendance was mandatory and was recorded via

Learning Catalytics, a cloud-based learning analytics and assessment system, which accounted for 5% of the final grade.

The control courses were taught in a traditional lecture format. Book reading was recommended, though not "assigned" or tightly correlated with the lecture each day. No homework or accountability measures were taken to ensure the students read as recommended. In class the instructor lectured for the full class time. The bulk of the lecture was delivered with PowerPoint slides, with more complex problems being worked out on the document camera. A mixture of definitions, introductory concepts, and conceptual discussions and problem based discussions were used. While the lectures did occasionally pause for reflections, and simple questions with one or two word answers were given to the students, time was not set aside to allow them to properly solve or think through a problem on their own. No free work time was given for problem solving. Learning Catalytics were used once per class for a low level question. It was typically given half way through the class period on the material that had just been lectured about. This was used to control for required attendance in the control section and the questions were generally simple definition based.

For each 50-minute class meeting, the treatment students were required to watch about two online videos before class. The videos created for the previous flipped class were reused. From student feedback, five videos were recreated to increase audio quality, and three long videos were split into short ones. The combined length of the videos remained practically unchanged, totaling 53 videos and 514 minutes with most videos within the range of 5–15 minutes ($M = 9.70$, $SD = 5.01$). To ensure compliance, each video was accompanied by an

assignment and each class would begin with a quiz with straightforward questions to test on video material. Students were expected to spend 60 to 90 minutes per week studying before class. The quizzes accounted for 5% of the total grade.

In the flipped section, a typical meeting was divided into three segments. First, the instructor would briefly review pre-class material and go through each assignment for 10 to 15 minutes. This included a brief two-minute open-note "quiz" to check for understanding and to increase accountability for watching videos. The review itself did not repeat factual information but aimed to foster conceptual understanding. The instructor would spend another 10 to 15 minutes with two relatively simple problems. Students worked on the problems in small ad hoc groups (typically 2-4 students) and submitted their answers via Learning Catalytics. Finally, the rest of class time would feature two to three increasingly difficult worksheet problems. The instructor and teaching assistants would roam over the classroom and offer help whenever needed. Students could submit and change answers at any time and the results were dynamically displayed to the instructor. The collective responses from the class were shown to the students, and the students were given time to discuss within their groups and change their answers if needed. If the majority of the class faltered, the instructor would either provide more hints or adjourn current activities to address common mistakes.

For both control and treatment sections, identical homework was given after class, which constituted 10% of the total grade. Homework was delivered via Mastering Chemistry, which has multiple functionalities but used in this course primarily for homework. These were a

mixture of conceptual, definition, and problem solving questions, varied in difficulty from very simple definitions and one step questions, to complex multi-topic and multi-stepped problems.

4.2.2 Participants

In total, 657 students were initially enrolled into the control (N = 313) and treatment (N = 344) sections. During the first class meeting, students were informed of the study and were invited to participate. After excluding students who either dropped the class or did not participate in any exams, the effective sample size was 287 students in the control and 320 in the treatment section, among whom most agreed to participate in the study (i.e., 97.56% or N = 280 and 95.94% or N = 307 respectively). Participants' demographics information was collected from the University's Registrar.

Student demographics were similar between sections, and a detailed comparison is shown in Table 4.1. Students came from 36 different majors and 12 ethnic groups. For simplicity, majors were regrouped into Biology/Chemistry, STEM (i.e. all STEM majors except for Biology and Chemistry), Non-STEM, and Undeclared. Similarly, ethnicity was regrouped into White, Black/Latino, South Asia, East Asia, and Unstated. High school GPA was collected, since the majority were freshmen who took this course as one of their first college-level courses.

4.2.3 Measures

A number of measures, including exam performance, out-of-class study time, motivation, and perceived class quality, were collected from exams and surveys.

Examinations. Three non-cumulative exams in weeks 3, 6, and 9 and one cumulative final exam in week 11 were administered, accounting for 15%, 20%, 20%, and 25% of the total grade respectively. All exams were similar in form and were administered back to back. To avoid cheating, different forms of the exams were used with isomorphic questions. Raw scores were converted into percentages. Students' letter grades were collected from a subsequent chemistry course, where our course is the first one in the sequence. The letter grades were converted into numeric values in such a way that an A+ corresponds to 13 and an F to 1.

Surveys. Five surveys, a pre-survey and four post-surveys (see Appendix B), were delivered to measure students' study effort, motivation, and perceptions. The pre-survey was given after the first class meeting. Each post-survey was administered three days before the corresponding exam. To encourage participation, 0.4 extra credits were rewarded for completing each survey, leading up to two extra credits in total. All survey responses were kept separate from the instructor and not processed until after the quarter. Survey items were framed on a 6-point scale with one being the most negatively keyed and six the most positively keyed responses. The survey response rate was higher (over 85%) in the beginning and lower (slightly below 80%) towards the end, averaging 82.64% (SD = 4.44%) in the control and 80.91% (SD = 3.93%) in the treatment sections.

Our survey motivation items were adapted from the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, Smith, Garcia, & McKeachie, 1993). Compliant to the expectancy-value theory (Wigfield & Eccles, 2000), items on interest, utility, achievement values, and self-efficacy from MSLQ were used in our study. Three items measured each

construct, whose reliability was assessed by Cronbach's alpha. In all surveys, the averaged alpha was over 0.80 for all constructs. A general motivation measure was hence constructed by averaging the twelve items with an average alpha of 0.89 (range: 0.85–0.92) over the surveys.

To measure study effort, the pre-survey asked students to provide numeric estimates of the average number of hours per week they spent studying before and after class for a typical science or mathematics class. Post-surveys asked for estimated average pre- and post-class study time per week during the intervening weeks between the previous exam and the incoming one.

Four post-surveys asked about students' perceived effectiveness of different instructional avenues. Student ratings on lecture quality and class quality were averaged to construct a measure of the overall class quality with a Cronbach's alpha averaging 0.81 (SD = 0.03). Post-surveys also included two items asking about the extent flipped students completed all pre-class videos and assignments. Students' narrative comments were collected from the university-wide end-of-quarter optional instructor evaluation.

4.3 Results

4.3.1 Preliminary Comparisons

Group equivalence. Descriptive statistics by section are presented in Table 4.1. Student demographics and pre-survey results suggest reasonable group equivalence on all measures except for high school GPA and majors. Specifically, the flipped students on average had lower GPA by -0.09 points out of 4.00, which is a small effect in size (ES = -0.148, p = 0.076). The treatment section, however, had notably 11.43% more Chemistry/Biology majors, and less

STEM, undeclared, and non-STEM majors (i.e., 4.59%, 4.49%, and 2.35% respectively); and the chi-squared test showed statistically significant ($p = 0.021$) difference in majors. In subsequent OLS analyses, student demographics were included to address minor group imbalances.

Table 4.1

Descriptive Statistics of Demographics, Pre-Survey Results, and Exam Outcomes by Group

Measure	Control ($N = 280$)	Treatment ($N = 307$)	t (p) or $\chi^2(p)$	Cohen's d
	M (SD) or Percentage (N)	M (SD) or Percentage (N)		
SAT Math	604.37 (72.03)	600.19 (76.19)	-0.67 (0.506)	-0.056
High School GPA	2.87 (0.62)	2.78 (0.60)	-1.78 (0.076)	-0.148
Chemistry/Biology	51.97% (145)	63.40% (194)	11.15 (0.011)	
STEM	11.83% (33)	9.48% (29)		
Non-STEM	7.53% (21)	2.94% (9)		
Undeclared	28.67% (80)	24.18% (74)		
Freshman	88.53% (247)	92.81% (284)	3.38 (0.184)	
Sophomore	8.24% (23)	5.56% (17)		
Junior/Senior	3.23% (9)	1.63% (5)		
Male	43.84% (121)	42.81% (131)	0.06 (0.802)	
Female	56.16% (155)	57.19% (175)		
White	11.11% (31)	16.67% (51)	4.28 (0.370)	
Black/Latino	31.54% (88)	28.43% (87)		
South Asia	27.96% (78)	28.76% (88)		
East Asia	26.52% (74)	23.53% (72)		
Unstated	2.87% (8)	2.61% (8)		
Interest	4.21 (0.93)	4.18 (0.96)	-0.28 (0.779)	-0.032
Utility	5.25 (0.84)	5.22 (0.80)	-0.32 (0.750)	-0.037
Importance	4.79 (0.92)	4.77 (0.94)	-0.31 (0.760)	-0.022
Self-efficacy	4.23 (0.87)	4.24 (0.87)	0.13 (0.893)	0.011
Motivation	4.80 (0.61)	4.79 (0.58)	-0.32 (0.749)	-0.017
Pre-class Study Time	5.27 (4.72)	5.35 (4.40)	0.21 (0.834)	0.018
Post-class Study Time	7.44 (5.50)	6.61 (5.94)	-1.61 (0.108)	-0.145
Midterm1	52.69 (17.54)	51.65 (16.86)	-0.73 (0.468)	-0.060
Midterm2	68.85 (15.14)	70.15 (14.85)	1.05 (0.294)	0.087
Midterm3	61.75 (19.23)	61.61 (17.97)	-0.09 (0.926)	-0.008
Final	67.98 (16.28)	64.70 (15.96)	-2.45 (0.014)	-0.204
Post-course Grade	7.01 (2.84)	6.32 (2.92)	-2.49 (0.013)	-0.239

Outcome comparisons. From Table 4.1, two-sample *t*-tests showed no significant impact of flipped instruction on all three non-cumulative midterms, as the magnitude of the effect sizes was consistently smaller than 0.10 standard deviations. In the cumulative final exam, flipped students on average underperformed their control counterparts by 3.28% (ES = -0.204, p = 0.014), which is close to a half-letter grade difference. Furthermore, in the post-chemistry course, the flipped students also underperformed their control counterparts (ES = -0.239, p = 0.013).

4.3.2 Compliance and Study Time

(1) Did flipped students comply with pre-class study requirement and did they spend more or less time studying outside the class?

Compliance. To ensure compliance, each class meeting started with a quiz. Flipped students generally did quite well in the quizzes, indicating a high degree of pre-class study compliance. Survey results corroborated this claim. On average, 83.71% (SD = 5.13%) of the flipped students indicated that they often finished all the videos before class, among which 36.11% (SD = 2.06%) reported to have always finished them. On the contrary, 16.29% (SD = 5.13%) claimed that they were often unable to watch all the videos, among which 2.51% (SD = 1.79%) claimed that they never watched videos.

Study time. Table 4.2 shows the self-reported estimates of pre- and post-class study time for each section. Three midterms and one final exam naturally delimited the class into four periods. Flipped students consistently spent more time before class (ten-week average: ES = 0.165, p = 0.055) and less time thereafter (ES = -0.194, p = 0.024). As a result, the overall out-of-

class study time was roughly the same (*ES* = -0.024, *p* = 0.768). These results confirmed what we had shown in the previous study that flipped instruction did not put extra burden on students, as increase in pre-class study time was offset by decrease in post-class study effort.

Table 4.2

Self-reported Out-of-class Study Time in Hours by Section

	Week	Control Mean (SD)	Treatment Mean (SD)	*t*-statistic (*p*)	Cohen's *d*
Before-class	Weeks 1-3	4.641 (3.714)	5.298 (3.363)	2.087 (0.037)	0.186
	Weeks 4-6	5.347 (4.078)	5.822 (3.707)	1.326 (0.185)	0.122
	Weeks 7-8	5.241 (4.005)	6.191 (3.915)	2.563 (0.011)	0.240
	Weeks 9-10	6.039 (4.548)	6.86 (4.293)	1.762 (0.079)	0.186
	Weeks 1-10	5.444 (3.834)	6.043 (3.427)	1.927 (0.055)	0.165
After-class	Weeks 1-3	9.67 (5.595)	8.463 (5.378)	-2.482 (0.013)	-0.220
	Weeks 4-6	9.772 (5.63)	8.694 (5.777)	-2.056 (0.040)	-0.189
	Weeks 7-8	9.381 (5.635)	9.032 (5.637)	-0.662 (0.508)	-0.062
	Weeks 9-10	10.29 (6.709)	9.279 (6.263)	-1.477 (0.141)	-0.156
	Weeks 1-10	9.834 (5.472)	8.805 (5.168)	-2.26 (0.024)	-0.194
Out-of-class	Weeks 1-3	12.566 (9.331)	12.124 (8.839)	-0.588 (0.557)	-0.049
	Weeks 4-6	12.688 (9.763)	11.671 (10.212)	-1.233 (0.218)	-0.102
	Weeks 7-8	11.658 (9.656)	11.979 (10.589)	0.385 (0.701)	0.032
	Weeks 9-10	9.896 (11.646)	10.546 (11.373)	0.682 (0.495)	0.057
	Weeks 1-10	11.943 (8.613)	11.73 (8.795)	-0.295 (0.768)	-0.024

4.3.3 Exam Performance and Motivation

(2) Did flipped instruction increase student exam performance, motivation, and subsequent course grade? If so, did students of diverse background benefit equally?

Exam performance. To account for minor imbalances over GPA and majors, OLS regression was employed and the results are shown in Table 4.3. The first three models used final exam scores as the dependent variable. In our study, the cumulative final exam was valued more than non-cumulative midterms, because it revealed the overall long-term impact of flipped

instruction. Moreover, 70.36% ($N = 197$) control and 75.89% ($N = 233$) treatment students were enrolled into a subsequent chemistry course in the following quarter. Their letter grades were

Table 4.3

Effect of Flipped Instruction on Exam Performance with OLS Models

	Final Exam Score			**Post-course Grade**		
	Model3.1	**Model3.2**	**Model3.3**	**Model3.4**	**Model3.5**	**Model3.6**
(Intercept)	0.086	0.189*	0.156+	0.040	0.115	0.134+
	(0.055)	(0.081)	(0.086)	(0.064)	(0.075)	(0.077)
Treatment	-0.107+	-0.276**	-0.207+	-0.129*	-0.269**	-0.301**
	(0.063)	(0.104)	(0.118)	(0.061)	(0.095)	(0.098)
Motivation (pre-survey)	0.066*	0.061+	0.060+			
	(0.033)	(0.033)	(0.033)			
High School GPA	0.688***	0.685***	0.683***	0.835***	0.834***	0.838***
	(0.035)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
SATmath	0.140***	0.146***	0.148***	0.093**	0.094**	0.095**
	(0.035)	(0.036)	(0.036)	(0.035)	(0.035)	(0.035)
Female		-0.162+	-0.168+	-0.175**	-0.302**	-0.315***
		(0.094)	(0.094)	(0.063)	(0.091)	(0.093)
Treatment:Female		0.249+	0.246+		0.233+	0.252*
		(0.129)	(0.131)		(0.122)	(0.123)
Sophomore		-0.161	-0.236			-0.180
		(0.196)	(0.205)			(0.209)
Junior/Senior		0.412	0.288			-0.332
		(0.275)	(0.29)			(0.312)
Treatment:Sophomore		0.545*	0.725*			0.323
		(0.274)	(0.300)			(0.288)
Treatment:Junior/Senior		-0.381	-0.049			NA
		(0.394)	(0.466)			NA
STEM	0.130	0.085	0.185	0.192+	0.189+	0.197+
	(0.122)	(0.126)	(0.168)	(0.109)	(0.109)	(0.114)
Non-STEM	-0.348*	-0.460*	-0.242	-0.572**	-0.609**	-0.469+
	(0.159)	(0.194)	(0.242)	(0.217)	(0.217)	(0.255)
Undeclared	-0.092	-0.094	-0.014	0.015	0.02	0.024
	(0.074)	(0.075)	(0.103)	(0.078)	(0.078)	(0.078)
Treatment:STEM			-0.204			
			(0.243)			
Treatment:Non-STEM			-0.586			
			(0.407)			
Treatment:Undeclared			-0.165			
			(0.146)			
Cases	470	469	469	406	406	406
Adj. R-squared	0.541	0.543	0.543	0.649	0.651	0.650
AIC	980.70	980.12	982.84	744.73	743.02	746.68

Note. All estimates are standardized beta coefficients. Standard errors are in parentheses.

$^{+}\ p < .10$, $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$

used as the dependent variable for models 3.4–3.6 in Table 4.3. In all six models, continuous variables were standardized and the estimates are hence standardized beta coefficients that can be interpreted as effect sizes.

Model 3.1 is the main effect model that included student demographics and prior motivation as covariates without adding any interaction terms; non-significant terms were not included in the model. High school GPA and majors were statistically significantly associated with the final exam scores, and the treatment effect was somewhat negative (ES = -0.107, p = 0.091). Potential interaction effects were studiously explored, and Model 3.2 suggests that females and sophomores benefited from flipped instruction more than males and freshmen. Specifically, while first-year males in the flipped section did significantly worse than their control counterparts (ES = -0.276, p = 0.008), first-year females did better than first-year males (ES = 0.249, p = 0.055) and sophomores did remarkably better than freshmen (ES = 0.545, p = 0.047) in the treatment condition. By implication, it is second-year females who benefited most from flipped instruction. In fact, by changing the reference groups, the OLS model revealed that second-year females in treatment condition outperformed their control counterparts (ES = 0.517, p = 0.060). It is worth mentioning that due to the small presence of sophomores (i.e., 6.84% or N = 40), statistical significance as indicated by p values should be considered together with the size of the effect that signifies practical importance. Model 3.3 included the interaction between treatment and majors. Although none of the terms were statistically significant, the size of the coefficients suggests the possibility that non-Biology/Chemistry majors did worse in the flipped condition than their Biology/Chemistry counterparts.

Model 3.4 is the main effect model with post-course chemistry grade as the dependent variable, where flipped students on average did worse than control students (*ES* = -0.129, *p* = 0.034). The same treatment-gender interaction of comparable magnitude (*ES* = 0.233, *p* = 0.057) reappeared in Model 3.5. The treatment-year interaction was not statistically significant (shown in Model 3.6) most likely due to further reduced sample size, as only 20 sophomores and no

Table 4.4

Effect of Flipped Instruction on Motivation with OLS Models

	Model4.1	**Model4.2**	**Model4.3**	**Model4.4**
	Motivation4	**Motivation4**	**Motivation3**	**Motivation2**
(Intercept)	0.065	0.158	0.191^{+}	0.138
	(0.072)	(0.106)	(0.099)	(0.099)
Motivation (pre-survey)	0.548^{***}	0.524^{***}	0.530^{***}	0.558^{***}
	(0.043)	(0.044)	(0.041)	(0.041)
Treatment	-0.053	-0.245^{+}	-0.175	-0.147
	(0.082)	(0.134)	(0.125)	(0.125)
High School GPA	0.101^{*}	0.140^{**}	0.166^{***}	0.101^{*}
	(0.045)	(0.050)	(0.045)	(0.046)
Female		-0.187	-0.130	-0.091
		(0.122)	(0.113)	(0.113)
SATmath		-0.080^{+}	0.024	-0.046
		(0.047)	(0.044)	(0.043)
Treatment:Female		0.338^{*}	0.096	0.012
		(0.169)	(0.158)	(0.158)
GPA:SATmath		0.084^{+}	0.071^{+}	0.088^{*}
		(0.043)	(0.040)	(0.039)
STEM	0.161	0.175	-0.096	-0.198
	(0.153)	(0.164)	(0.152)	(0.154)
Non-STEM	-0.436^{*}	-0.530^{*}	-0.857^{***}	-0.096
	(0.216)	(0.220)	(0.205)	(0.196)
Undeclared	-0.248^{*}	-0.286^{**}	-0.227^{*}	-0.111
	(0.096)	(0.099)	(0.092)	(0.093)
Cases	422	403	396	411
Adj. R-squared	0.320	0.330	0.391	0.370
AIC	1048.80	994.31	913.98	966.31

Note. All estimates are standardized beta coefficients. Standard errors are in parentheses.

$^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

juniors or seniors enrolled into the subsequent course. The size of the coefficients, however, echoed the same trend revealed by Model 3.2.

Motivation. Shown in Table 4.4, Model 4.1 is the main effect model with motivation measured by the fourth post-survey as the dependent variable; non-significant demographic covariates were not shown. On average, flipped instruction did not change student motivation to any meaningful extent (*ES* = -0.031, *p* = 0.705). Model 4.2 shows significant treatment-female interaction (*ES* = 0.338, *p* = 0.047) and marginally significant GPA-SAT interaction (*ES* = 0.084, *p* = 0.050). However, the treatment-female interaction was not observed in the second (*ES* = 0.012, *p* = 0.940 from Model 4.4) and third (*ES* = 0.096, *p* = 0.544 from Model 4.3) post-surveys.

4.3.4 Perception and Implementation Issues

(3) Did flipped instruction impact perceived overall class quality? Were there further implementation issues?

Perception. Regardless of the introductory nature of this course, 51.55% and 38.92% of the students from the combined sample rated this course as "very" and "adequately" challenging, where the two sections differed little. Students' ratings agreed with exam outcomes, where the average raw scores were consistently less than 70% for both sections across exams. Moreover, in all four periods, flipped students rated the class to be of lower quality (*ES* range: -0.245 – -0.357, *p* value range: 0.009–0.0001).

From post-survey responses, we compared flipped students' ratings of the perceived effectiveness of different instructional avenues. Across periods, in-class problem solving was

ranked as the most effective means of learning, followed in order by learning before class, online videos, and in-class group discussion. The textbook and in-class lectures were rated as the least and second least effective means, which is not surprising considering that the textbook was not frequently used and lectures often took only a fraction of class time.

Implementation issues. Student comments from the standard campus-wide instructor evaluation provide additional insight. The positive comments echoed the benefits reported in our previous study, including (a) flexibility for learning at one's own pace, (b) availability of online videos for review before exams, (c) better preparation for class meetings, (d) more opportunities for demonstration and problem solving in class, and (e) more instructor-student interaction. Most importantly, we classified students' negative comments to identify weaknesses in our instruction. Two main sources of criticism emerged from the flipped classroom.

First, flipped students expressed strong frustration with the technology failures in class.

"Once Learning Catalytics stopped working, we started covering some material."

"I found the whole Learning Catalytics program to be really distracting. I feel like a lot of lecture time was wasted trying to get it running and I was always paranoid that my phone would be out of battery and I would not receive points, etc. Rather than make complicated answer questions (which were sometimes hard to input), a small clicker question here and there would be able to prove attendance and provide a general idea of the degree to which students understand those underlying concepts."

In addition, some flipped students criticized the active learning techniques involved, notably group discussion and peer instruction.

"She can have more examples of problems in class that she solves with the students before letting them solve other problems themselves. It's hard to apply what we don't know to try to answer the questions."

"Going through more problems together rather than allowing excess time for group discussion might be better because time is wasted and only a few problems are finished in 50 minutes where as more could be fit in. The idea of giving students time together to try a problem is a nice idea, but doesn't always execute the way intended."

"Explain the material much more thoroughly; answer questions by explaining the process to the student rather than making the student explain it."

"For a student with a very weak background in Chemistry, being asked questions that I don't know the answer to when seeking help only embarrassed me and makes me not want to ask questions."

4.4 Discussion

4.4.1 Compliance and study time

Giving assignments associated with each video and for-credit quizzes with each class effectively reduced pre-class study non-compliance. This finding agreed with reports from other studies (Foertsch, Moses, Strikwerda, & Litzkow, 2002; Mason et al., 2013; Narloch, Garbin, & Turnage, 2006). On the other hand, although only 16.29% students claimed that they often could not watch all the videos, this small fraction still translates into 50 students. In large undergraduate classes, non-compliance would affect a non-negligible number of students, even

though the fraction of students affected might be small. Flipped instructors, therefore, should consider monitoring non-compliance closely particularly when teaching a class comprised primarily of freshmen whose self-discipline and time-management skills are yet to be developed.

With regard to study effort, our current study reproduced what was observed in our prior study: Flipped instruction caused a shift in study time from post-class to pre-class without appreciably increasing students' overall workload. By implication, flipped students might benefit from spaced learning (Donovan & Radosevich, 1999). Given some students' opposition to the flipped pedagogy, it is advisable that flipped instructors should communicate this result to the students to dispel the concern that pre-class study would impose extra burden on them.

4.4.2 Exam performance and motivation

The presence of interaction effect regarding final exam outcome and post-course grade is an important finding. We believe interaction effect would most likely occur when the treatment conditions agree with the characteristics (e.g., motivation, intellectual capacity, and study habits) of a specific subgroup; others with characters departing from this niche group in varying degrees would thus benefit to lesser extents accordingly. In our case, second year females seemed to be the niche group. Flipped females consistently outperformed their control counterparts in both the final exam ($ES = 0.249$, $p = 0.055$) and post-course grade ($ES = 0.252$, $p = 0.041$), and showed higher end-of-course motivation ($ES = 0.338$, $p = 0.047$). In addition, females on average seem to spend more time outside the classroom ($ES = 0.149$, $p = 0.074$) than males did and flipped females relative to control females spent more time before class ($ES = 0.319$, $p = 0.069$) than

flipped males did relative to control males. Similarly, second year students did particularly well in the treatment condition. It is conceivable that sophomores were generally less reliant on instructor-initiated instruction and had stronger self-study, self-discipline, and time-management skills. They were hence more receptive to flipped instruction, as sophomores rated the class to be of higher quality particularly in the third ($ES = 0.577$, $p = 0.001$) and fourth ($ES = 0.400$, $p = 0.068$) post-surveys.

These results support the conjecture that flipped instruction might be more appropriate for students with strong drive, maturity, and skills. Our prior study suggests, without assignments and quizzes, it would take considerable drive, self-discipline, and self-directed learning skills for students to study before class (He, Holton, Farkas, & Warschauer, 2016). Although giving assignments and quizzes spurred students to complete pre-class learning assignments, the same set of attributes is still needed to ensure learning quality. Moreover, these attributes are also crucial for students to actively engage during class. When things go wrong in a flipped classroom, students with these qualities are arguably less vulnerable to suffer the consequences. Sophomores in our study, for example, might be more mentally mature, self-disciplined, active in self-directed learning, and emotionally less resistant to deviance from traditional lectures, which gave them an edge at every corner over the freshmen who were only high school seniors until recently.

4.4.3 Student perception and implementation issues

In this study, flipped students rated the class to be of lower quality. We therefore looked at students' comments for clues regarding implementation issues.

First, we believe massive technology failures in the flipped classroom were an important reason for the lower ratings. Both sections used LC (Learning Catalytics) instead of iClickers to facilitate peer instruction and real-time feedback. Each student was assigned a unique IP address and connected to the class via a smartphone or tablet. The control students took the class first and had little issue in this regard. In the treatment section, however, some students (random each time) could not get connected, because the control class had used up most of the IP addresses. This situation was not fully resolved until the sixth week. By that time, students were already weary of using the technology. Given the prolonged technology failures, communication with the students is critical in establishing confidence. In addition, the instructor should have changed back to iClicker while the issue of Learning Catatlytics was being diagnosed and resolved. Failures to communicate and adjust could have instilled negative feelings leading to undesirable consequences.

Second, some flipped students voiced criticisms against certain active learning techniques, notably group discussion and peer instruction. Supported by the ideas of constructivism and zone of proximal development, group work is highly valued by educational researchers and has become a key component in many active learning techniques. Our results suggest, however, having students work in groups might not be as effective as one would expect, as students often ranked group discussion in the bottom of the list of preferred teaching practices, a finding

reported by others as well (Enfield, 2013). Some students expressed frustration with their own limited skills for problem solving and regarded group discussion and peer instruction as ineffective use of class time. Some demanded the instructor to elaborate more on complex concepts and demonstrate solving some problems first before diving into group-based problem solving.

These results prompt us to reflect upon the benefits of flipped instruction and the associated active learning techniques as compared to traditional lectures. Although passive lecturing has its shortcomings, it is probably still the most widely used instructional technique regardless of the variety of novel instructional techniques invented over the past decades to supplant it. We believe the resilience of lecturing owes primarily to its simplicity. In contrast, flipped instruction is a promising, but complex, instructional technique that entails making multiple decisions on pre-class and in-class components. In a flipped classroom, for example, an instructor need to consider the number and length of videos, accompanying practice questions, pre-class quizzes, percentage of lectures retained in class, the number and kinds of in-class active learning activities to adopt, and different ways to conduct them. The more decisions to make, the more it is likely that some step might incur an implementation issue. As a result, we highly recommend that instructors new to the flipped pedagogy should choose fewer and simpler technologies to start with. Moreover, it is important to note that many active learning techniques frequently require students to work in groups. Staging group activities, however, entails making multiple decisions regarding, for example, the difficulty of the problems, group size, group forming tactic (e.g., getting the appropriate group heterogeneity in skills), and time allotment

(i.e., enough time for thorough discussion, but not too much to induce boredom and elicit off-topic conversation). While it is possible for instructors to monitor group work closely in small classes, in large classrooms where consistent and complete oversight is possible, student could sit out class time pointlessly, unwittingly reinforce each other's biases, and have their prior misconceptions strengthened.

4.5 Conclusions & Recommendations

Giving assignments associated with each video and for-credit quizzes with each class effectively reduced pre-class study non-compliance. However, non-compliance could still affect a non-negligible number of students, even though the proportion of students affected might be small. Flipped instructors should therefore consider monitoring non-compliance closely particularly in large introductory undergraduate classes.

Our current study reproduced what was observed in our prior study that flipped instruction did not appreciably increase the overall study time but only caused a shift in workload, which implies that flipped students might benefit from spaced learning. Flipped instructors could communicate this result to students to dispel the concern that flipped instruction exerts extra burden on them. Moreover, flipped researchers do not need to reduce class meetings to control for increase in required pre-class study time.

While flipped students on average underperformed their control counterparts in the cumulative final exam (ES = -0.204, p = 0.014 by two-sample t-test and ES = -0.107, p = 0.091 by OLS Model 3.1), strong interaction effects existed between treatment condition and gender as

well as year level. Females and sophomores benefited more in the flipped section. Similar trends were also observed with student letter grades in a subsequent chemistry course. The differentiated treatment effect lends support to the conjecture that flipped instruction is more appropriate for students with strong drive, maturity, and learning skills.

Flipped instruction did not increase student motivation throughout the course. The same treatment-gender interaction was observed with the final survey, where flipped females showed much stronger motivation ($ES = 0.338$, $p = 0.047$) compared to flipped males. However, this interaction effect was not shown with previous surveys. Therefore, the interaction effect might be either appearing gradually or due to random statistical noise. We are currently conducting more analysis on motivation to clarify this issue.

Throughout the course, flipped students rated the class to be of lower quality, as they raised complaints about technology failures in class and about the lack of efficiency with in-class group discussion and peer instruction. In the face of technology issues, it is recommended that the instructor should actively communicate with the students and consider changing technologies. Failures to communicate and adjust would lead to serious trust issues that negatively impact student motivation and satisfaction, which in turn could hurt student exam performance.

The variety of issues associated with our flipped classroom prompted us to reflect upon the resilience of traditional lectures, where its simplicity might be its greatest virtue. We caution against overreliance on complex technologies or teaching techniques. It is advisable that flipped instructors in first-year introductory courses should start simple and be cautious of deviating from traditional lectures too much too fast. For example, instead of diving directly into problem

solving, some review and elaboration of difficult concepts is necessary as a gentle warm-up. Rather than using open-ended questions with groups of several students, pairs of students working on a clear problem with timely formative feedback are much more tractable. In fact, for the first several lectures, a partially flipped classroom that retains some portions of lectures is highly recommended. Surveys can be delivered early in the second week to gauge student attitudes and identify problems. Once students have displayed favorable attitude towards the flipped pedagogy, instructors could consider gradually adopting a fully flipped classroom, using fancier technologies or teaching techniques in class, and working with increasingly challenging and open-ended problems. For any novel technology or technique employed, the promise to improve teaching is invariably accompanied by challenges. The most effective methods will depend on the instructor, the students, and the institutional climate; special consideration to each must be given.

References

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

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

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795-805.

Enfield, J. (2013). Looking at the impact of the flipped classroom model of instruction on undergraduate multimedia students at CSUN. *TechTrends*, *57*(6), 14-27.

Foertsch, J., Moses, G., Strikwerda, J., & Litzkow, M. (2002). Reversing the lecture/homework paradigm using eTEACH® web□based streaming video software. *Journal of Engineering Education, 91*(3), 267-274.

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. *Education, IEEE Transactions on, 56*(4), 430-435.

Michael, J. (2006). Where's the evidence that active learning works?. *Advances in Physiology Education, 30*(4), 159-167.

Moravec, M., Williams, A., Aguilar-Roca, N., & O'Dowd, D. K. (2010). Learn before lecture: a strategy that improves learning outcomes in a large introductory biology class. *CBE-Life Sciences Education, 9*(4), 473-481.

Narloch, R., Garbin, C. P., and Turnage, K. D. (2006). Benefits of prelecture quizzes. *Teaching Psychology, 33*(2), 109-112.

Pintrich, P. R., Smith, D. A., García, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and psychological measurement*, *53*(3), 801-813.

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

Quint, C. L. (2015). *A study of the efficacy of the flipped classroom model in a university mathematics class* (Doctoral dissertation, Teachers College, Columbia University).

Touchton, M. (2015). Flipping the classroom and student performance in advanced statistics: evidence from a quasi-experiment. *Journal of Political Science Education*, *11*(1), 28-44.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary educational psychology, 25*(1), 68-81.

Chapter 5 Third-year Implementation

5.1 Introduction

Our current study is a follow-up of two prior iterations of the flipped pedagogy (He, Holton, Farkas, & Warschauer, 2016; Study II). Given non-compliance with pre-class study as a serious implementation issue, our first flipped study found a small and statistically significant treatment effect on student final exam performance. Students' responses to the flipped pedagogy was distinctly lukewarm with one fifth displaying polarized feelings. Our second study of flipped instruction encountered massive technological failures. End-of-quarter student surveys showed strong negative ratings against the flipped classroom. The treatment effect was a small, but statistically significant, negative effect with strong interactions indicating that second-year females benefited from flipped instruction more than first-year males. Students' grades in a subsequent course showed exactly the same pattern, i.e., small negative overall impact with a strong interaction favoring second-year females. Our current study addressed the various implementation issues and changed the structure of the flipped course (see details below). As a result, the primary goal of this study is to look further into this issue by answering two questions:

(1) Did our current implementation of flipped instruction increase student final exam performance in the present class and overall grade in a subsequence class?

(2) Did this iteration positively impact student general motivation in and perceptions of the present class?

5.2 Methodology

5.2.1 Course Description

Data from the present study was collected from two sections of a first-year introductory chemistry course taught by the same instructor, where the fall 2014 class was the control section

and the fall 2015 class was the treatment. Before fall 2014, the instructor had taught the course using traditional lecture format seven times in three consecutive years, and taught in the flipped format twice in two preceding years before fall 2015. In each section, students met three times a week on Mondays, Wednesdays, and Fridays for ten weeks from 1:00 to 1:50 pm. Class attendance was mandatory and accounted for 5% of the final grade.

The way the control section was taught was described in the previous study. In previous studies, some students encountered difficulties adjusting to the flipped pedagogy and asked for more in-depth reviews before delving into problem solving. As a response, the instructor adopted a softer approach for introducing flipped instruction by including more lecturing component on Mondays and Wednesdays and more problem solving activities on Fridays. Although the treatment section was described to the students as having "Flipped Fridays", all class meetings were essentially flipped and differed only in the ratio of in-class lecturing to problem solving.

For each 50-minute class meeting, the treatment students were required to watch between one to three online videos before class. All videos made for the previous flipped class were reused. There were 53 videos in total with most within the range of 5–15 minutes ($M = 9.70$, $SD = 5.01$). While all videos were mandatory in our previous implementations, six videos that involves more difficult topics were made optional this time and the instructor spent time lecturing on these topics in class. Students were expected to spend about 30 to 50 minutes per week studying before class. To ensure compliance, each video was accompanied by an assignment and each class would begin with a quiz with straightforward questions testing on video material, where the quizzes accounted for 5% of the total grade. With these measures in place, our previous implementation showed reasonable compliance rate (Study II).

During class on Mondays and Wednesdays, students would take a two-minute open-note quiz to check for understanding. The quiz questions were on low-level knowledge to encourage student engagement and increase accountability for watching the pre-class videos. After the quiz, the course proceeded with a highly interactive lecture. The lecture briefly reviewed concepts from the videos. Time for review was adjusted depending on the results of the in-class quiz. A higher proportion of the time was devoted to more difficult concepts and problems than in the traditional lecture section. In each class meeting, approximately three problems were completed by students working in small ad-hoc groups. Students' responses were monitored in real time by the instructor using Learning Catalytics, an in-class response system. In a 50-minute class session, roughly 15 minutes were student centered activities, and 35 minutes were lecture.

On "Flipped Fridays", the class time was spent doing more in-depth problem solving. The class started out with a two-minute quiz as described above. A review of any problems identified by the quiz was completed. The rest of the class session focused on completion of questions and problems aimed to foster a deeper understanding of the material. In contrast to the problems on Mondays and Wednesdays, which were generally one- or two-step problems, Fridays' problems were often multi-stepped connecting multiple concepts together. Before each question was completed, a very brief introduction and review of the concepts was completed by the instructor. The students then solved the problems in ad-hoc groups. Learning Catalytics was used to monitor student progress in real time. The instructor, two TAs and three tutors roamed the classroom answering questions. Time given on each problem differed based on the difficulty of the problem and feedback from the response system. Further review was given if undesirable class performance necessitated it. On Fridays, in a 50-minute class session, approximately 35

minutes were student centered activities and 15 minutes were lecture relating specifically to these activities.

After class, students were required to complete homework administered via Mastering Chemistry, which is an online homework and assessment system developed by Pearson. The homework constituted 10% of the total grade.

5.2.2 Participants

The two quarters initially enrolled 516 students with 287 students in the control section and 229 students in the treatment section respectively. In the beginning of the quarters, students were informed of the study and all students were invited to participate. After excluding opt-outs and those who never took the final exam, the effective sample size of the participants was 277 students (i.e., 96.52%) in the control and 223 students (i.e., 97.38%) in the treatment section. Participants' demographics information was collected directly from the University's Registrar.

Student demographics were similar between sections, and a detailed comparison is shown in Table 5.1. In the combined sample, 46.68% (N = 232) were males and 53.32% (N = 265) females. They came from 28 different majors and 12 ethnic groups. For convenience, students' majors were regrouped into 51.00% (N = 255) Biology/Chemistry, 12.00% (N = 60) STEM (i.e., including all STEM majors except for biology and chemistry related ones), 6.40% (N = 32) Non-STEM, and 30.60% (N = 153) Undeclared. Similarly, ethnicity was regrouped into 11.60% (N = 58) White, 32.60% (N = 163) Black/Latino, 28.60% (N = 143) South Asia including Vietnamese, Thai, and Filipino, 24.80% (N = 124) East Asia including Korean, Chinese, and Japanese, and 2.40% (N = 12) Unstated. Freshmen constituted 89.00% (N = 445) of the students with 8.80% (N = 44) sophomores, and 2.20% (N = 11) juniors/seniors. The average SAT math score was 606.30 (SD = 73.87) and the average high school GPA was 2.85 (SD = 0.64). High school GPA, instead

college GPA, was requested, because the majority of the class were first-year students who took the course as one of their first college-level courses.

5.2.3 Measures

A number of measures, including exam and course performance, motivation, and perceived class quality, were collected from exams and surveys.

Examinations. In each quarter, a cumulative final exam was administered during the eleventh week, which accounted for 25% of the total grade. The two final exams were practically identical with only cosmetic changes. Raw scores were converted into percentages for the ease of comparison. We also collected students' letter grades from a subsequent chemistry course, where our current course is the first one in a three-course sequence. The letter grades were converted into numeric values in such a way that an A+ corresponds to 13 and an F to 1.

Surveys. A pre-survey and a post-survey were delivered to measure students' motivation and perceptions of the effectiveness of various learning avenues. The pre-survey was given immediately after the first class meeting and the post-survey was administered days before the final exam. To encourage participation, 0.5 extra credits were rewarded for completing each survey, leading up to one extra credit in total. All survey responses were kept separate from the instructor and the teaching assistants and not processed until after the quarter, except for counting reward credits. All survey items, were framed on 6-point scales with one being the most negatively keyed and six the most positively keyed responses. The survey response rate was 87.36% for the pre-survey and 66.37% for the post-survey in the control section, and 80.72% and 60.29% in the treatment section.

Motivation items on surveys were of primary interest. Based on the expectancy-value theory (Wigfield & Eccles, 2000), motivation was measured by eight items, two for each

construct regarding interest, utility, achievement value, and self-efficacy. The average reliability, measured by Cronbach's alpha, was .85 (*SD* = .04) for pre-survey and 0.86 (*SD* = .05) for the four constructs. The overall motivation measure was constructed by averaging the eight items with the mean Cronbach's alpha of 0.89 and 0.87 for the pre- and post-surveys.

Three items from standard university-wise, end-of-quarter, anonymous instructor evaluation were used to measure students' perception of the clarity of the instructor, the rating of the instructor, and the rating of the course. The response rate was 63.90% in the control and 55.16% in the treatment.

5.3 Results

5.3.1 Preliminary Comparisons

Group equivalence. Before examining treatment effect on final exam performance, group equivalence is checked first and descriptive statistics by section are presented in Table 5.1. Student demographics and pre-survey results suggest reasonable group equivalence on all measures except for the composite general motivation measure: Flipped students on average had lower motivation by -0.17 on a six-point scale, which is a small but significant effect (*ES* = -0.24, *p* = .03). Upon completion of the current course, 70.76% (*N* = 196) of the control students and 68.61% (*N* = 153) of the flipped students enrolled into the subsequence course, a difference that is not statistically significant under chi-squared test (*p* = .45). Descriptive statistics for students who were subsequently enrolled in the following quarter were also computed and the results presented in Table 5.2. Similar to results from Table 5.1, no systematic differences were identified except for general motivation measure, as flipped students reported lower motivation (*ES* = -0.30, *p* = .02).

Table 5.1

Descriptive Statistics of Demographics, Pre-Survey Results, and Final Exam Outcome by Group

Measure	Control ($N = 277$)	Treatment ($N = 223$)	t (p) or $\chi^2(p)$	Cohen's d
	M (SD) or Percentage (N)	M (SD) or Percentage (N)		
SAT Math	603.94 (71.81)	609.48 (76.62)	0.79 (0.43)	0.08
High School GPA	2.87 (0.62)	2.84 (0.67)	-0.51 (0.61)	-0.05
Chemistry/Biology	51.99% (144)	49.78% (111)	1.38 (0.71)	
STEM	11.91% (33)	12.11% (27)		
Non-STEM	7.22% (20)	5.38% (12)		
Undeclared	28.88% (80)	32.74% (73)		
Freshman	88.45% (245)	89.69% (200)	3.30 (0.19)	
Sophomore	8.30% (23)	9.42% (21)		
Junior/Senior	3.25% (9)	0.90% (2)		
Male	43.80% (120)	50.22% (112)	2.04 (0.15)	
Female	56.20% (154)	49.78% (111)		
White	11.19% (31)	12.11% (27)	2.04 (0.73)	
Black/Latino	31.40% (87)	34.08% (76)		
South Asia	27.80% (77)	29.60% (66)		
East Asia	26.72% (74)	22.42% (50)		
Unstated	2.89% (8)	1.79% (4)		
Interest	4.20 (0.93)	4.31 (1.09)	1.07 (0.29)	0.11
Utility	5.24 (0.84)	5.21 (0.98)	-0.37 (0.71)	-0.04
Importance	4.79 (0.92)	4.75 (1.10)	-0.39 (0.70)	-0.04
Self-efficacy	4.23 (0.87)	4.27 (1.09)	0.48 (0.63)	0.05
Motivation	4.80 (0.60)	4.63 (0.85)	-2.25 (0.03)	-0.24
Final Exam	67.98 (16.28)	68.00 (18.23)	0.008 (0.99)	0.001

Table 5.2

Descriptive Statistics of Demographics, Pre-Survey Results, and Post-Course Grade by Group

Measure	Control ($N = 196$)	Treatment ($N = 153$)	t (p) or $\chi^2(p)$	Cohen's d
	M (SD) or Percentage (N)	M (SD) or Percentage (N)		
SAT Math	605.82 (71.87)	609.8 (71.17)	0.49 (0.62)	0.06
High School GPA	2.96 (0.53)	2.99 (0.58)	0.51 (0.61)	0.05
Chemistry/Biology	62.24% (122)	66.01% (101)	5.78 (0.12)	
STEM	11.73% (23)	4.58% (7)		
Non-STEM	4.08% (8)	3.92% (6)		
Undeclared	21.94% (43)	25.49% (39)		
Freshman	91.84% (180)	94.77% (145)	1.25 (0.54)	
Sophomore	5.61% (11)	3.92% (6)		
Junior/Senior	2.55% (5)	1.31% (2)		
Male	41.45% (80)	43.79% (67)	0.19 (0.66)	
Female	58.55% (113)	56.21% (86)		
White	11.73% (23)	14.38% (22)	1.17 (0.88)	
Black/Latino	30.61% (60)	30.72% (47)		
South Asia	32.14% (63)	31.37% (48)		
East Asia	22.96% (45)	22.22% (34)		
Unstated	2.55% (5)	1.31% (2)		
Interest	4.24 (0.89)	4.24 (1.12)	0.07 (0.94)	0.00
Utility	5.37 (0.72)	5.32 (0.96)	-0.41 (0.69)	-0.06
Importance	4.83 (0.86)	4.8 (1.06)	-0.25 (0.81)	-0.03
Self-efficacy	4.22 (0.86)	4.19 (1.10)	-0.32 (0.75)	-0.03
Motivation	4.85 (0.57)	4.64 (0.88)	-2.34 (0.02)	-0.30
Final Exam	72.08 (12.35)	73.79 (13.09)	1.24 (0.22)	0.14
Post-Course Grade	7.01 (2.84)	8.14 (2.80)	3.71 (0.00)	0.40

In the following ordinary least squares (OLS) regression analyses, student demographics are included primarily to study potential interaction effects. Including demographic variables also helps to address minor imbalances between group (e.g., regarding general motivation) and reduces residual error, which in turn increases statistical power of the models for detecting small treatment effects.

5.3.2 Exam and Post-Course Performance

(1) Did our current implementation of flipped instruction increase student final exam performance in the present class and overall grade in a subsequence class?

Table 5.3

Effect of Flipped Instruction on Exam Performance with OLS Models

	Final Exam Score			**Post-course Grade**		
	Model3.1	**Model3.2**	**Model3.3**	**Model3.4**	**Model3.5**	**Model3.6**
(Intercept)	0.12^{+}	0.15^{*}	0.14^{*}	-0.11^{+}	-0.13^{*}	-0.17^{**}
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Treatment	0.12^{+}	0.05	0.05	0.33^{***}	0.35^{***}	0.34^{***}
	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
Prior Motivation	0.05^{+}					
	(0.03)					
High School GPA	0.70^{***}	0.68^{***}	0.67^{***}	0.80^{***}	0.88^{***}	0.75^{***}
	(0.03)	(0.03)	(0.03)	(0.04)	(0.06)	(0.07)
SATmath	0.13^{***}	0.13^{***}	0.13^{***}	0.15^{***}	0.15^{***}	0.12^{**}
	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Female	-0.14^{*}	-0.16^{*}	-0.15^{*}	-0.31^{***}	-0.30^{***}	-0.29^{***}
	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
Sophomore	-0.19	-0.13	-0.11			
	(0.13)	(0.11)	(0.11)			
Junior/Senior	0.41^{+}	0.43^{*}	0.52^{*}			
	(0.24)	(0.22)	(0.22)			
STEM	0.05	0.07	0.05			
	(0.11)	(0.10)	(0.10)			
Non-STEM	-0.44^{*}	-0.47^{**}	-0.49^{**}			
	(0.19)	(0.16)	(0.15)			
Undeclared	-0.12^{+}	-0.16^{*}	-0.15^{*}			
	(0.07)	(0.07)	(0.07)			
GPA:Sophomore			0.27^{*}			
			(0.12)			
GPA:Junior/Senior			-0.80^{+}			
			(0.41)			
Treatment:GPA					-0.15^{*}	-0.15^{+}
					(0.08)	(0.08)
Final Exam						0.23^{***}
						(0.07)
Cases	387	460	460	320	320	320
Adj. R-squared	0.62	0.59	0.59	0.65	0.65	0.66
AIC	723.6	905.3	900.3	581.7	579.7	570.2

Note. All continuous variables are standardized z-scores. All estimates are standardized beta coefficients. Standard errors are in parentheses.

$^{+}\ p < .10,\ ^{*}\ p < .05,\ ^{**}\ p < .01,\ ^{***}\ p < .001$

Final exam performance. As shown in Table 5.1, two-sample *t*-test suggests that flipped instruction had practically zero impact on the cumulative final exam (*ES* = 0.001, *p* = .99). Results from OLS regression are shown in Table 5.3, where all continuous variables are standardized z-scores. The regression coefficients are therefore beta coefficients and can be interpreted directly as effect sizes. The first three models used standardized z-scores from the cumulative final exam as the dependent variable.

Model 3.1 is the main effect model, where non-significant terms are not included. Flipped instruction showed a small, marginally significant, positive effect (*ES* = 0.12, *p* = .07). The effect of prior motivation measured by pre-survey is negligible small (*ES* = 0.05, *p* = .09). Model 3.2 shows the results with prior motivation removed. Most coefficients remain practically unchanged except for the treatment effect, which has become non-significant (*ES* = 0.05, *p* = .44). Potential interaction effects were studiously explored and results are shown in Model 3.3, which reveals interaction between high school GPA and year of enrollment. Sophomores with average high school GPA preformed slighted worse compared to their freshmen counterparts (*ES* = -0.11, *p* = .36). However, sophomores with GPA one standard deviation above average, would gain a statistically significant extra boost in final exam performance (*ES* = 0.27, *p* = .03). Juniors and seniors showed the opposite trend. However, due to the small presence of juniors and seniors (i.e., 2.20%, *N* = 11), we do not regard them as representative of the junior-senior population and choose not to interpret too much into the corresponding results.

Post-course performance. In the post-chemistry course, two-sample *t*-test shows that flipped students outperformed the control students by a half-letter grade (*ES* = 0.40, *p* < .001) and the corresponding results from OLS regression are presented as Model 3.4–3.6 in Table 5.3.

Model 3.4 is the main effect model without including interaction or non-significant terms. Flipped instruction had a statistically significant effect (*ES* = 0.33, $p < .001$) on subsequent course grade. Model 3.5 included one interaction term between treatment and high school GPA, which suggests that flipped students with average GPA outperformed their control counterparts (*ES* = 0.35, $p < .001$) and that students with GPA one standard deviation below average would gain an extra performance boost (*ES* = 0.15, $p = .046$). The estimated coefficients were consistent such that adding final exam scores as control did not meaningfully change the outcomes. Flipped students with average GPA still outperformed their control counterparts (*ES* = 0.34, $p < .001$) and the interaction term is also present (*ES* = 0.15, $p = .05$).

5.3.3 Motivation and Perceptions

(2) Did current implementation of flipped instruction positively impact student general motivation in and perceptions of the present class?

Motivation. From Table 5.1, two-sample *t*-test on prior motivation indicates that flipped students on average had lower motivation to start with (*ES* = -0.24, $p = .03$). The post-survey showed that flipped students had caught up with the control students (*ES* = 0.07, $p = .55$) by the end of the quarter. Therefore, OLS regression was applied to control for prior motivation and the results are shown in Table 5.4. Model 4.1 is the main effect model without interaction and non-significant terms. While prior motivation is a strong predicator of end-of-course motivation, flipped instruction had a small but statistically significant impact (*ES* = 0.22, $p = .047$). Moreover, strong interaction exists between treatment condition and prior motivation. Shown in Model 4.2, flipped students with average prior motivation to start with had higher post motivation than their control counterparts (*ES* = 0.22, $p = .04$) and flipped students with prior motivation one standard deviation below average would gain a significant extra increase in

motivation ($ES = 0.38$, $p < 0.001$). Model 4.3 shows one more interaction term between SAT math and prior motivation, which suggests that the effect of prior motivation on post motivation is stronger with students of lower SAT math scores.

Table 5.4

Effect of Flipped Instruction on Motivation with OLS Models

	Model4.1	**Model4.2**	**Model4.3**
(Intercept)	-0.11	-0.14^{+}	-0.15*
	(0.07)	(0.07)	(0.07)
Prior Motivation	0.52***	0.73***	0.73***
	(0.06)	(0.08)	(0.08)
Treatment	0.22*	0.22*	0.23*
	(0.11)	(0.11)	(0.11)
High School GPA	0.12*	0.12*	0.13*
	(0.06)	(0.06)	(0.06)
SATmath	0.13*	0.12*	0.1^{+}
	(0.06)	(0.06)	(0.06)
Treatment:Motivation		-0.38***	-0.24*
		(0.11)	(0.11)
Motivation:SATmath			-0.16***
			(0.04)
Cases	265	265	265
Adj. R-squared	0.25	0.28	0.31
AIC	688.5	678.9	665.6

Note. All continuous variables are standardized z-scores. All estimates are standardized beta coefficients. Standard errors are in parentheses.

$^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Perceptions. Campus-wide, anonymous instructor evaluation asked students to rate the clarity of the instructor, the overall teaching quality of the instructor, and the overall quality of the course. For all three measures, flipped students gave higher ratings ($ES = 0.53$, $p < 0.001$ for clarity, $ES = 0.53$, $p < 0.001$ for teaching quality, $ES = 0.47$, $p < 0.001$ for overall course quality). Student responses to open-ended questions regarding the strengths and weaknesses of the flipped course echoed the ratings. Some students singled out “Flipped Fridays” as their favorite sessions.

I really liked the flipped Friday's class. We got to do a fair amount of example problems with guidance and were able to ask specific questions pertaining to problem solving. I found that highly effective.

Love the Flipped Fridays and video lectures, and her attitude towards mostly first year college students is great. Also, she kept the class ahead of other classes, which I personally liked.

It was nice to have practice problems on Friday to help enforce what we learned and to go over them.

I also liked flipped Fridays. Doing questions in class helped me see what I needed help with.

The positive comments from the flipped section confirmed some proposed benefits of a flipped classroom: First, availability of online videos provides flexibility for learning at one's own pace and makes it convenient for review before exams.

I like the flipped class because I can take however much time I want to take notes on the videos.

I like the videos because if I am still unsure about a topic, I can go back and re-watch a video to relearn a topic.

I liked that the videos are always available so that I can look back and review them if I still don't understand the concept.

I really appreciated the videos that you made for us. I would not have gotten good grades on the midterm without them. Extremely helpful is an understatement!

Second, learning before class prepares the students for productive engagement during class meetings.

I liked how we were able to come to class with a basic understanding so we would not be completely lost in class.

I really like the flipped courses format of this class. Coming into class with a basic understanding obtained from the videos really helps cement the information during the lecture.

The flipped format was great because it gave students a taste of what to expect in addition to just doing homework. It made going to lecture a lot more productive than straight up lecturing and the participation via Learning Catalytics encouraged participation.

Third, having freed up time from lectures, a flipped classroom affords more opportunities for demonstration and problem solving in class and enables more instructor-student interaction.

The flipped format is good due to the fact that the lectures can focus primarily on the difficult areas and the videos beforehand can teach the basics.

I liked the flipped format of the course because it allowed for students to get the help they needed on practice problems.

I really enjoyed the flipped format of this course! I felt it was a meticulous, highly analyzed, and perfected system. I liked how I was able to practice the concepts with the guide of the professor (and help from the TAs). Very effective, interactive, and engaging.

The flipped format of this class is great in my opinion. It gives us time to do more hands-on activities and problems in class. Also, courses that are purely lecture make me fall asleep sometimes no matter how interested I am in the material. Getting help is also much easier in a flipped format course since the professor and TAs can recognize your mistakes immediately by walking around as you work. Overall, this is a great way to conduct a chemistry class.

Most importantly, compared with the criticisms raised against flipped instruction in our previous studies, the negative comments from the current flipped class were much less severe and critical. A few students uttered complaints against the amount of repetitive homework.

My major complaint was the amount of homework problems and I would say to make more of them optional because the repetitiveness of problems made the homework boring.

My biggest complaint is probably how long the homework is and how it occupies all my week.

A few others commented that the clarity and depth of some videos could be improved.

I strongly disliked some of the videos. Some videos were just ill-prepared in how she spoke. There were a lot of stutters, awkward pauses, and sentences that sounded like questions. Most times it turned me away from paying attention and I would just skim the video for the answers needed on MasteringChem

I am all for the method of teaching through video, as youtube videos are how I got through heavy-based math classes in high school (calculus, trig, physics, chemistry, etc) but I would like it if the videos we were more in depth on the topic. It seems as though the videos and questions that go with it are elementary compared to the understanding expected of us on the exam.

The only major complaint that resurfaced time and again is in fact demands for less lecturing and more practice.

I wish during class we went over more practice problems instead of just going over the video we watched at home.

My major complaint is that there should be more class time dedicated to doing practice problems and explaining how to do it instead of rushing through.

The time spent working on problems in class feels like time wasted. The problems are too easy and the instruction is too slow. The pre-example, like the one about making canoes, are childish and irritating. This course is too slow and the instructor babies the class. Hold the class to a higher standard. Spend less time in class reiterating the material covered at home.

Most days, class felt like any other lecture (not a flipped class). The lecture was repetitive; it seemed that the professor repeated the information in the video. Although there were many practice problems in the lecture, it still felt like a normal lecture. To improve this course (and take more advantage of having a flipped style class), it would be better to have more practice problems during class. Perhaps the first 15-20 minutes would be used to give a quick review of the videos and the rest of the time would be dedicated to problem-solving.

5.4 Discussion

5.4.1 Exam and Post-Course Performance

One seeming inconsistency shown in Table 5.3 relates to the treatment effect of flipped instruction. With prior motivation included as a covariate, Model 3.1 shows a small, marginally significant treatment effect ($ES = 0.12$, $p = 0.066$). Keeping prior motivation as a covariate while adding the interaction terms included in Model 3.2 and 3.3 produced estimates of treatment effect of similar magnitude (not shown in Table 5.3). Removing prior motivation, however, strongly reduced estimates of treatment effect without markedly changing other regression coefficients. The reduced size of treatment effect is comparable to that from two-sample *t*-test. We believe the inconsistency might arise owing to the reduced sample size due to missing data from survey non-responses associated with the prior motivation measure. While Model 3.2 and 3.3 include 92.0% ($N = 460$) of the students in the sample, Model 3.1 only retains 77.4% ($N = 387$) of students. Students who responded to the pre-survey might be different from non-

respondents. As a result, flipped instruction might have a small, marginally significant, positive effect on final exam performance for survey respondents, but the overall effect is small when non-respondents are included in the sample. Considering the larger sample size and result from two-sample *t*-test, we choose Model 3.3 to be final model, where the treatment effect of flipped instruction is negligibly small ($ES = 0.055$, $p = .36$).

While flipped instruction did not improve student final exam performance in the current course, for students subsequently enrolled, those who came from the flipped section with average high school GPA outperformed their control counterparts ($ES = 0.35$, $p < .001$) and the effect is stronger for students with lower GPA. Specifically, students with high school GPA one standard deviation below average would outperformed their control counterparts by 0.50 standard deviations, which is about a half letter grade increase. Similarly, students with GPA two standard deviations below average would gain an improvement of 0.65 standard deviations, which is close to a full letter grade difference. By contrast, students with high school GPA two standard deviation above average would perform on a par with the control students ($ES = 0.05$, $p = .77$).

Collectively speaking, these results suggest that our implementation of flipped instruction has some potential to bridge the achievement gap over time, since academically weaker students benefited much more from flipped instruction while academically stronger students were not performing significantly worse. This outcome is conceivable considering that the instructor had deliberately changed the class structure to cater to the needs of the majority of the class. Student with weaker academic skills would hence benefit more from the class. This conjecture can find some support from motivation results (as discussed below) and student responses to open-ended survey questions, since some students vocally demanded the instructor to move at a faster pace, include more challenging problems, and refrain from baby sitting the class, which indicates the

presence of a dissatisfied, less motivated, academically stronger group of students. In addition, similar results have been reported by others showing that flipped instruction selectively increased exam performance of academically disadvantaged students by about 0.60 standard deviations in a subsequent course (He & Link, 2015) and that flipped instruction benefits only students from the bottom tier while not affecting students from the middle and top tiers (Ryan & Reid, 2015).

5.4.2 Motivation and Perceptions

Although flipped instruction did not improve student final exam performance, it had positive impact on student motivation and perceptions of the course. The effect of flipped instruction on student motivation parallels the interaction effect between treatment condition and high school GPA on subsequent course grade. Specifically, flipped students with average prior motivation would have higher motivation than their control counterparts by the end of quarter ($ES = 0.23$, $p = .03$) and the increase is stronger for students with lower than average prior motivation. Flipped students with prior motivation two standard deviations below average would become more motivated relative to control students by 0.71 standard deviations. However, flipped students with prior motivation two standard deviations above average would have lower motivation by -0.25 standard deviations. This implies that students with high prior GPA were more likely to be less satisfied with the flipped course. Given the similar interaction patterns, we speculate that some causal link might exists such that end-of-course motivation could potentially influence post-course performance. However, an alternative explanation is that the current course structure caters more to the needs of academically weaker students, which results in students with lower prior GPA performing relatively better and students with higher GPA less satisfied and hence less motivated.

In general, students rated the flipped course higher in all three aspects regarding instructional clarity, instructor quality and overall course quality. By contrast, in our first study, students exhibited lukewarm feelings towards flipped instruction with about one fifth showing polarized responses. In the second study, flipped students had consistently rated the course to be of lower quality across surveys throughout the quarter. Compared with previous results, three factors might have contributed to the higher student ratings in this study.

First, the slightly reduced pre-class workload and the associated assignments and quizzes have ensured desirable pre-class study compliance, which is a precondition for a successful flipped classroom. Our first study has exposed non-compliance as a serious implementation issue due to absence of assignments and quizzes to hold students accountable for pre-class learning. As a result, students who failed to watch the videos regarded the class to be overly rushed and asked for more review and in-depth explanations, whereas those who adequately prepared for class claimed boredom and demanded for more problem solving activities. The differences in pre-class preparation mostly likely resulted in the overall lukewarm student reaction with one fifth of the students showing polarized responses. Our second study has shown that including assignments and quizzes effectively ensured pre-class study compliance. The current study used the same set of assignments and quizzes associated with each video, hence keeping non-compliance at bay.

Second, absence of technology failures contributes to the smooth delivery of the flipped pedagogy. Our second study used Learning Catalytics instead of iClickers as the class response system. The new technology had an unexpected problem. The control students took the class first and each student was assigned a unique IP address for connecting to the class via a smartphone or tablet. The flipped students came in the following session. However, about half of the students

could not get connected, because the control class had used up most of the IP addresses. This situation was not fully resolved until the six week. As a result, flipped students had voiced strong complaints regarding technology failures in the class, which we believe was the primary reason for the consistent lower ratings.

Third, gentle introduction of the flipped pedagogy is most likely a key determinant factor for the positive ratings. Our first two iterations of the flipped pedagogy suggest that first-year college students might lack the motivation, self-discipline and academic skills necessary for ensuring compliance with and the quality of pre-class study. We have therefore suggested that first-year introductory courses should start simple and be cautious of deviating from traditional lectures too much too fast. Our current study acted on this suggestion by retaining some lectures while periodically increasing and decreasing the problem solving component. As a result, we have observed much fewer complaints frequently voiced in previous studies. Interestingly, some students have singled out "Flipped Fridays" as their favorite sessions and would like to have more sessions flipped similarly. Incidentally, anecdotal evidence suggests that some of our flipped students who had disliked the flipped format quickly changed their minds when they were enrolled into a subsequent traditional course. We therefore conjecture that periodic contrast between the two instructional formats in the same weeks might actually contribute to the positive ratings of the flipped pedagogy.

5.5 Conclusions

Two sample *t*-test ($ES = 0.001$, $p = .99$) and OLS regression ($ES = 0.05$, $p = .44$) have shown that flipped instruction had little effect in improving student final exam performance in the current course. No marked interaction was identified with OLS regression, indicating that the treatment effect was uniform across subgroups of students in terms of demographics and prior

academic performance. Flipped instruction had an overall positive impact on student overall grade from a subsequent course ($ES = 0.33$, $p < .001$). Most importantly, interaction effect was identified between treatment condition and prior high school GPA. Students with lower GPA benefited more from flipped instruction while students with higher GPA were not performing significantly worse. By implication, these results suggest our flipped instruction has some potential in bridging the achievement gap over time.

The treatment effect on motivation shows a similar pattern to that on post-course performance. Flipped instruction improved student overall motivation relative to the control group ($ES = 0.22$, $p = .047$). Interaction effect existed such that students with lower prior motivation showed greater increase in motivation by the end of the course. Unlike post-course performance, however, students with much higher prior motivation, e.g., two standard deviations above average, would have lower motivation relative to control students. To account for the parallel outcomes regarding end-of-quarter motivation and post-course performance, two not mutually exclusive explanations are proposed. We speculate that end-of-course motivation might have a causal influence on post-course performance. Alternatively, the way current flipped classroom was conducted might cater more to the needs of academically weaker students, which eventually translates into the corresponding gains in end-of-course motivation and post-course overall grade.

Students rated the flipped course much more positively by all three measures regarding instructional clarity ($ES = 0.53$, $p < 0.001$), overall instructor quality ($ES = 0.53$, $p < 0.001$), and overall course quality ($ES = 0.47$, $p < 0.001$). Flipped students appreciated the flexibility due to the availability of video lectures, openly acknowledged that learning before class better prepared them for active engagement in class, and endorsed flipped instruction as way to introduce more

opportunities for demonstration, problem solving, and student-instructor interactions. Most importantly, compared to the criticisms raised against flipped instruction in our previous studies, negative comments were much less in scope and severity in the current study. In fact, the major complaint this time is the demand for less lecturing and more practice.

We believe three factors contribute to improved student ratings. First, assignments and quizzes associated with each video effectively reduced non-compliance with pre-class study, which is a necessary precondition for a successful flipped classroom. Second, absence of technology failures ensured smooth delivery of active learning activities in class. Third, the gentle introduction of flipped instruction might play a critical role, as it gave students time to adapt to the new learning scheme while reflecting upon the benefits of flipped instruction by repeatedly comparing against traditional lectures within the same quarter.

References

He, W., & Link, R. (2015). Bridging the achievement gap: A longitudinal study of the differential lingering effects of flipped instruction. AERA 2015 Annual Conference, Chicago, IL.

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.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary educational psychology*, *25*(1), 68-81.

Chapter 6 Conclusions and Implications

Our first-year and second-year implementations of flipped instruction have consistently shown that the flipped pedagogy did not increase students' overall study time. It only caused a shift in student workload, as an increase in pre-class study time was counterbalanced by a decrease in after-class study time. By implication, any positive impacts of flipped instruction should be attributed to factors other than mere increase in study effort. The shift in study time also implies that flipped students might in theory benefit from spaced learning, as they distributed their study time more evenly. Flipped instructors could communicate this result to students to dispel the concern that flipped instruction exerts extra burden on them. Moreover, to assess flipped instruction, it might be unnecessary to adjust additional pre-class study time by reducing the number of class meetings.

Our first implementation has suggested that without enough pre-class assignments and quizzes as precautions, non-compliance with pre-class study could be a serious issue. Habitual resistance, procrastination and distraction, and unintended lapses are three primary causes for non-compliance. The results also suggest that non-compliance seems to be more predominant among students with poor self-discipline, low motivation, and weak time-management and academic skills. The fallout of non-compliance affects the entire class, as under-prepared students tended to have difficulty following the class while well-prepared students reported boredom and did not benefit as much from in-class instruction. Most encouragingly, our second implementation suggests that providing adequate assignments and quizzes associated with each videos is an effective means to ensure pre-class study compliance. However, the quality of pre-class learning could still be affected by factors related to student maturity, motivation, and self-learning skills. Moreover, non-compliance could still affect a non-negligible number of students,

even though the proportion of students affected might be small. Flipped instructors should therefore consider monitoring non-compliance closely particularly in large introductory undergraduate classes.

It is argued that measuring student performance using long-term, high-stakes exams gives more practically meaningful results. With flipped instruction implemented for the first time, our OLS models showed a small and statistically significant treatment effect (ES = 0.192, p = .008) with the final exam. Most importantly, the overall treatment effect was more pronounced in the beginning, but diminished over time, which supports the previous claim that long-term high-stakes final exam should give a more realistic assessment.

Three years' results have shown that our flipped instruction had a small to non-existing impact on student final exam performance. Interaction effects were only detected with the second implementation, as second-year females were found to benefit most from the flipped pedagogy, which seems to support our previous claim that self-disciple and self-learning skills might be important factors contributing to the successful execution of flipped instruction.

Although our third and final implementation did not appreciably improve student performance in the current course, it significantly improved student performance in a subsequent course. Most encouragingly, student with weaker prior high school GPA benefited most from the flipped classroom, while stronger students were barely unaffected. The results hence confirm what others have found, that the flipped pedagogy has the potential to bridge the achievement gap between students.

Apart from possible pre-class non-compliance, technological issues could hinder a successful implementation. The variety of issues exposed during our second-year implementation prompted us to reflect upon the resilience of traditional lectures, where its

simplicity might be its greatest virtue. In contrast, flipped instruction is a relatively complex instructional technique that requires multiple decisions. The more decisions to make, the more likely that some step might incur an implementation issue.

In all three studies, student comments have consistently confirmed some proposed benefits of flipped instruction, including learning at one's own time and pace, better preparation for class, and more problem solving and teacher-student interaction. When various implementation issues exist, however, student motivation and perceptions towards the flipped pedagogy were not positive. With our first implementation, flipped instruction did not increase student motivation and perceived overall class quality. Treatment students' preference of flipped instruction over traditional lectures was lukewarm with about one fifth of the students displaying polarized feelings. For the second year, not only did student motivation not increase, students rated flipped instruction to be of lower quality. By contrast, with issues of non-compliance and technological issues resolved and flipped instruction introduced in a "softer" manner, student motivation and perceptions were consistently much higher in the third implementation.

Collectively speaking, it is advisable that flipped instructors in first-year introductory courses should start simple and be cautious of deviating from traditional lectures too much too fast. For example, instead of diving directly into problem solving, some review and elaboration of difficult concepts is necessary as a gentle warm-up. Rather than using open-ended questions with groups of several students, pairs of students working on a clear problem with timely formative feedback are much more tractable. In fact, for the first several lectures, a partially flipped classroom that retains some portions of lectures is highly recommended. Surveys can be delivered early in the second week to gauge student attitudes and identify problems. Once students have displayed favorable attitude towards the flipped pedagogy, instructors could

consider gradually adopting a fully flipped classroom, using fancier technologies or teaching techniques in class, and working with increasingly challenging and open-ended problems. For any novel technology or technique employed, the promise to improve teaching is invariably accompanied by challenges. The most effective methods will depend on the instructor, the students, and the institutional climate; special consideration must be given to each.