IMPACT OF EVIDENCE-BASED ACTIVE-ENGAGEMENT COURSES ON STUDENT PERFORMANCE AND GENDER GAP IN INTRODUCTORY PHYSICS

by

Nafis Imtiaz Karim

B.Sc., American International University-Bangladesh, 2007

M.Sc., Ecole Centrale de Nantes and Warsaw University of Technology, 2011

B.Sc., University of Leipzig, 2015

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

THE DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Nafis Imtiaz Karim

It was defended on

July 10, 2018

and approved by

Dr. Adam K. Leibovich, Professor, Department of Physics and Astronomy

Dr. Robert P. Devaty, Associate Professor, Department of Physics and Astronomy

Dr. Russell J. Clark, Senior Lecturer, Department of Physics and Astronomy

Dr. Larry J. Shuman, Professor, Department of Industrial Engineering

Dissertation Advisor: Dr. Chandralekha Singh, Professor, Department of Physics and Astronomy

IMPACT OF EVIDENCE-BASED ACTIVE-ENGAGEMENT COURSES ON STUDENT PERFORMANCE AND GENDER GAP IN INTRODUCTORY PHYSICS

Nafis Imtiaz Karim, Ph.D.

University of Pittsburgh, 2018

This thesis explores impact of evidence-based active-engagement courses (EBAE) on student achievement and gender gap in introductory physics. The first study is about the pedagogical content knowledge (PCK) of the physics teaching assistants (TAs) at identifying introductory students' difficulties using the Conceptual Survey of Electricity and Magnetism, which is important for implementing evidence-based pedagogy. The second study focuses on EBAE physics classes in which there are potential opportunities for instructors and TAs to apply their PCK and other research-based instructional strategies to improve student learning. We investigated whether EBAE classes improved student performance compared to traditional lecture-lased (LB) classes and whether EBAE classes helped improve student performance. We used the Force Concept Inventory (FCI) in physics I and the Conceptual Survey of Electricity and Magnetism (CSEM) in physics II as assessment instruments. Our findings suggest that, on average, students in EBAE classes outperformed students in LB classes in conceptual posttests although their scores on pretests were not statistically significantly different. Moreover, on average, both male and female students in EBAE classes outperformed those of the same gender in LB classes on posttests although there was no difference on the pretests. However, no reduction in the gender gap in EBAE classes was observed.

We also investigated the impact of stereotype threat in introductory physics classes. When students were asked to indicate their gender immediately before taking standardized physics tests, no deterioration in female students' performance on standardized test was observed compared to

the case when gender was not indicated. Moreover, we also investigated the extent to which agreeing with the stereotype that men to generally perform better in physics than women was correlated with students' performance and found that this type of belief is not very common (~10%). However, in some situations, female students who agreed with the stereotype performed worse than female students who did not agree with it. This effect appears to be stronger in the calculus-based courses compared to the algebra-based courses. Finally, we propose that implementing interventions to improve female students' social belongingness, self-efficacy and growth mindset may help reduce the gender gap in physics courses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First, I would like to express my thanks to my research advisor and supervisor, Dr. Chandralekha Singh for her continued guidance and support throughout my entire research. I feel privileged to have had the opportunity to work closely with her and have learned much about physics education research. I have also learned from her much about cognitive sciences, teaching physics and ways to interact with students in order to help them improve their attitudes towards learning physics.

I also express my gratitude to my research collaborator, Dr. Alexandru Maries from the University of Cincinnati for his help and continued support. I am very grateful to Dr. Maries for his valuable and insightful comments and suggestions that aided my research move forward.

Besides my advisor and collaborator, I would like to thank the thesis committee members: Dr. Robert Devaty, Dr. Adam Leibovich, Dr. Larry Shuman and Dr. Russell Clark for their comments and critiques on my research. I would like to express additional gratitude to Dr. Devaty for taking his precious time to read all my papers and provide innumerable fruitful comments which improved the quality of this dissertation. I am also thankful to Dr. Emily Marshman and Dr. Jeremy Levy for their feedback and helpful comments on my research.

My sincere thanks also goes to my academic advisor, Dr. Carlos Badenes and the departmental chair, Dr. Arthur Kosowsky for their helpful suggestions during my academic years.

I would also like to thank all the TAs, students and instructor who participated in my research and have helped me by providing valuable classroom data. This thesis was made possible by their willingness to participate.

Last but not the least, I would like to thank the National Science Foundation for its award.

1.0 INTRODUCTION

Human learning is a complex process. The research on human learning has been a part of psychology since the 1800s partly because the existence of a culture depends on the ability of its new members to learn set of skills, whether new or existing. The application of this research in teaching physics, and the development of Physics Education Research as an area of research is rather new. Today, there are many research-validated instructional approaches that attempt to improve student learning and problem solving skills of introductory and advanced physics students.

1.1 LEARNING

Since my thesis will focus on student learning and in particular, improving student learning of physics, we will define what we mean by 'learning'. One definition of learning, which is used in the book "Learning and Memory" by J. Anderson [1], is as follows: "Learning is the process by which long-lasting changes occur in behavioral potential as a result of experience." Anderson elaborates the key terms in this definition as follows [1]: In this definition, learning typically refers to the process of change. Also the change must be relatively long lasting to exclude certain transient changes that we will not call learning. Moreover, if a person learns something, but it does not affect that person's behavior because it is kept latent, there is no way to know what was learned.

In particular, not everything we learn has an impact on our behavior. However, learning can impact the potential for certain type of behavior, e.g., how we would perform on a physics test and one can device various tests to measure whether learning has taken place.

In the light of Anderson's definition of learning [1], I aspire to investigate several aspects of learning in introductory physics and how to improve students' learning and performances in introductory physics classes. My goal in this thesis is to evaluate students' learning of physics and their performances in traditional lecture-based (LB) introductory physics classes as well as in evidence-based active engagement (EBAE) classes. I also sought to investigate how students' gender and their beliefs related to certain stereotypes about gender impact their learning. In particular, my research includes an investigation of the impact of stereotype threats on male and female students' performance on standardized physics tests. I also studied the pedagogical content knowledge (PCK) of the teaching assistants (TAs) in introductory physics courses as it relates to electricity and magnetism concepts since this PCK can greatly impact student learning in those courses.

Many cognitive science researchers have proposed different learning frameworks to improve student learning. Therefore, before describing each component of my research investigations, I will first discuss these learning frameworks from cognitive science that have provided foundation for my investigations.

1.2 LEARNING FRAMEWORKS FROM COGNITIVE SCIENCE

Cognitive scientists have developed a number of learning frameworks in order to improve student learning. One overarching framework that can guide effective approaches to teaching and learning

is described by Collins et al. and is known as the Cognitive Apprenticeship model [2]. Moreover, in order to develop curricula and pedagogies to help students learn concepts and develop problem solving [3-20], reasoning and meta-cognitive skills, Vygotsky's Zone of Proximal Development (ZPD) framework, Piaget's optimal mismatch framework and Schwartz, Bransford and Sear's framework focusing on the Preparation For Future Learning (PFL) are also very useful.

1.2.1 Cognitive Apprenticeship Model

Cognitive apprenticeship model is a framework developed by Collins, Brown, and Newman in 1989 [2]. The model describes the process in which a master (or a teacher) teaches a skill to an apprentice (or a student). The authors developed six teaching steps for cognitive apprenticeship framework which can be condensed into three main steps. In particular, the three main steps, modeling, coaching and scaffolding, and weaning are at the core of cognitive apprenticeship model and help with cognitive and metacognitive development.

Modeling is the first step in which a master exemplifies or demonstrates his/her skills to the apprentice. In this step, the master carries out a task in front of the student so that the students can observe and build a conceptual model of the processes that are required to accomplish the task and perform well. In the second step, coaching and scaffolding, the instructor lets the student perform the task while observing the student do it. The instructor provides feedback and corrects the student's mistakes. This active real-time feedback to improve student's cognitive and meta-cognitive skills is central to helping students develop desired skills and learn useful concepts. This step of coaching and scaffolding continues until the student reaches to a certain level of expertise. At this point, the student can continue performing the task on his own for self-improvement while getting less frequent feedback as needed. The instructor gradually removes or weans the support

and lets the student develop self-reliance and continue the task on his/her own. This step is called weaning or fading.

1.2.2 Zone of Proximal Development (ZPD)

The Zone of Proximal Development (often called ZPD) is a concept attributed to the Soviet psychologist Lev Vygotsky in the early 1930s [21]. In this framework, ZPD is the difference between what a learner can do without an expert's help and what s/he can do with the help of an expert who is familiar with the learner's initial knowledge and skills and takes advantage of this knowledge to improve learning. The ZPD framework suggests that students should be presented with tasks that are within their ZPD, so that optimal learning can take place. As a child gradually learns concepts and develops the skills to perform certain tasks, their ZPD will stretch. In this sense, ZPD is a dynamic construct.

1.2.3 Optimal Mismatch

Piaget was a Swiss cognitive psychologist whose "optimal mismatch" framework [22] proposes that in order to optimize learning, one should provide optimal mismatch. A related framework is the theory of conceptual change put forth by Posner et al. [23]. In this framework, conceptual changes or "accommodations" can occur when the existing conceptual understanding of students is not sufficient for or is inconsistent with new phenomena observed. In particular, in order for conceptual change to occur, instructional design should provide opportunity for students to realize that there is some inconsistency between what their mental models are and what they are observing. Then after this cognitive conflict students are in a state of disequilibrium and providing optimal

mismatch would imply that students are giving appropriate guidance and support via the instructional design in order for desired assimilation and accommodation of knowledge to occur.

1.2.4 Preparation for Future Learning

Schwartz et al. [24] proposed that in order to prepare students for future learning and for helping them develop expertise in a particular domain so that they can transfer their learning from one context to another, two orthogonal dimensions should be considered carefully in any instructional design. In this two dimensional learning space, the two dimensions correspond to efficiency and innovation. According to Schwartz et al., efficiency-oriented practices or exercises alone do not require in-depth understanding or anything innovative. So, by repeating this type of task, a student or an apprentice can have very high efficiency but not be able to transfer the learning to new situation. In this case, the learner will become a routine expert. Innovation, on the other hand, requires an individual to think and understand the problem deeply. However, if the task is too innovative and does not have sufficient efficient component to it, novice students may not be able to accomplish the task and become frustrated novices. What Schwartz et al. proposed is that the instructional design should have a balance of innovation and efficiency so that the task is neither too easy nor too difficult and students focus on learning while engaged with the task. This type of task will create an optimal learning condition for the students to become adaptive experts so that they do not give up. In other words, tasks that have a sufficient balance of innovation and efficiency can prepare students for future learning and can help them transfer their learning from one situation to another.

We note that for our purposes, Vygotsky's ZPD framework, Piaget's optimal mismatch framework and Schwartz et al.'s PFL framework are very similar. In particular, all of these

frameworks emphasize the importance of knowing the prior knowledge and skills of the students in order to design instruction and employ pedagogies to help them. Knowing initial knowledge and skills of the students can guide an instructor to provide optimal mismatch, remain in the ZPD and give the instructor a good understanding of what is efficient or innovative for the students at a given point of time in the instructional sequence so that learning is optimized via appropriate curricula and pedagogies.

Now that we have reviewed the cognitive frameworks, below, we summarize the various research investigations discussed in the different chapters of the thesis.

1.3 PEDAGOGICAL CONTENT KNOWLEDGE OF TEACHING ASSISTANTS USING THE CONCEPTUAL SURVEY OF ELECTRICITY AND MAGNETISM

The Conceptual Survey of Electricity and Magnetism (CSEM) is a conceptual multiple-choice survey [25] commonly used to assess student learning in introductory electricity and magnetism courses. Pedagogical content knowledge, or PCK as defined by Shulman [26, 27], includes "Understanding of the conceptions and preconceptions that students bring with them to the learning of those most frequently taught topics and lessons." According to this definition, knowledge of students' common alternate conceptions is one aspect of PCK. The research presented in Chapter 2 uses the CSEM to explore this aspect of the PCK of physics graduate Teaching Assistants (TAs) in the context of electricity and magnetism. In particular, I explore the extent to which physics TAs are able to identify common alternate conceptions of introductory students on individual items on the CSEM. Knowledge of the common difficulties and of the types

of reasoning used by introductory physics students can be helpful in designing pedagogical strategies to improve student learning [26-33]. Findings will be presented in Chapter 2.

1.4 IMPACT OF EVIDENCE-BASED ACTIVE-ENGAGEMENT COURSES ON STUDENT PERFORMANCE IN INTRODUCTORY PHYSICS

In the study presented in Chapter 3, I used the Force Concept Inventory (FCI) [34] in the first semester courses and the Conceptual Survey of Electricity and Magnetism (CSEM) [25] in the second semester introductory physics courses to assess student learning. The FCI, CSEM and other standardized physics surveys [25, 34-40] have been used to assess introductory student understanding of physics concepts by a variety of educators and physics education researchers [41]. One reason for their extensive use is that many of the items on the survey have strong distractor choices which correspond to students' common difficulties so students are unlikely to answer the survey questions correctly without having good conceptual understanding. In the research discussed in Chapter 3, the performance of students in evidence-based active-engagement (EBAE) courses at a particular level is compared with primarily lecture-based (LB) courses in two situations: (I) the same instructor taught two courses, one of which was a flipped course involving EBAE methods and the other an LB course, while the homework and final exams were kept the same, (II) student performance in all of the EBAE courses taught by different instructors were averaged and compared with primarily LB courses of the same type also averaged over different instructors. Whenever differences between these two groups were observed (with students in EBAE courses performing better than students in the LB courses), we investigated which students were benefitting most from the EBAE courses, e.g., those who performed well or poorly on the

pretest given at the beginning of the course. Finally, we investigated the typical correlation between the performance of students on the validated conceptual surveys and their performance on the final exam, which typically places a heavy weight on quantitative physics problems. Findings will be presented in Chapter 3.

1.5 IMPACT OF EVIDENCE-BASED ACTIVE-ENGAGEMENT COURSES ON GENDER GAP IN INTRODUCTORY PHYSICS

In introductory physics, prior research has found that male students often outperform female students on conceptual assessments such as the Force Concept Inventory (FCI) [34, 42] and the Conceptual Survey of Electricity and Magnetism (CSEM) [25, 43], a phenomenon sometimes referred to as the "gender gap". The origins of gender gap on the FCI [34] both at the beginning and end of an introductory physics course have been a subject of debate with some researchers arguing that the test itself is gender-biased [44]. Some of the origins of the gender gap are related to societal gender stereotypes [45-48] that keep accumulating from an early age. In the study presented in Chapter 4, we used the FCI in the first semester introductory physics courses and the Conceptual Survey of Electricity and Magnetism (CSEM) [25] in the second semester courses to assess student learning. We also investigated any possible gender gap at the beginning of the course as well as the extent to which evidence-based pedagogies can help reduce it. Along with FCI and CSEM, other standardized physics surveys [35-40] have been used to assess introductory students' understanding of physics concepts by a variety of educators and physics education researchers. One reason for their extensive use is that many of the items on the survey have strong distractor choices which correspond to students' common difficulties so students are unlikely to answer the

survey questions correctly without having a good conceptual understanding. In addition to investigating the gender gap in the LB and EBAE courses on the standardized tests (FCI and CSEM), we also investigate the gender gap on the final exam performance. Findings will be presented in Chapter 4.

1.6 IMPACT OF STEREOTYPE THREAT ON STUDENT PERFORMANCE AND GENDER GAP IN INTRODUCTORY PHYSICS

Prior research has found that activation of a stereotype about a particular group in a test-taking situation, i.e., stereotype threat, can alter the performance of that group in a way consistent with the stereotype. Some researchers have found that subtle stimuli that can activate stereotype threat and result in deteriorated performance [49], e.g., asking students to indicate their ethnicity before taking a test [50]. In particular, prior research suggests that asking African American students to indicate their ethnicity before taking a difficult test on verbal ability resulted in decreased performance compared to students who were not [50]. Yet others have found that asking for gender or ethnicity before taking a test does not impact students' performance on standardized tests [51-53].

In a study presented in Chapter 5, I investigated whether asking introductory students to indicate their gender before taking the CSEM impacted their performance, both when it was administered as a pre-test (before instruction) and as a post-test (after instruction in relevant concepts). In the other study, I investigated the prevalence of the belief that men generally perform better in physics than women (a gender stereotype) among introductory students and the extent to which agreeing with this gender stereotype is correlated with the performance of female and male

students on the FCI. It was hypothesized that asking students for their beliefs about this gender stereotype may act as a stereotype threat, especially for female students who agree with the stereotype, and they may perform worse than female students who do not agree with it. The studies were conducted over several consecutive years. Findings will be presented in Chapter 5.

1.7 CHAPTER REFERENCES

1. J. R. Anderson, *Learning and Memory: An Integrated Approach*, New York, NY, John Wiley and Sons, Inc., 2nd Ed (2000).

2. A. Collins, J. S. Brown and S. E. Newman, Cognitive Apprenticeship: Teaching the crafts of reading, writing and apprenticeship in *Knowing, Learning and Instruction: Essays in Honor of Robert Glaser*, edited by L. Resnick, Hillsdale, NJ, Lawrence Erlbaum Associates, 453-494 (1989).

3. C. Singh, When physical intuition fails, Am. J. Phys. **70**, 1103 (2002).

4. C. Singh, Assessing student expertise in introductory physics with isomorphic problems. I. Performance on a non-intuitive problem pair from introductory physics, Phys. Rev. ST Phys. Educ. Res. **4**, 010104 (2008).

5. C. Singh, Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer, Phys. Rev. ST Phys. Educ. Res. **4**, 010105 (2008).

6. A. Mason and C. Singh, Assessing expertise in introductory physics using categorization task, Phys. Rev. ST Phys. Educ. Res. **7**, 020110 (2011).

7. A. Mason and C. Singh, Helping students learn effective problem solving strategies by reflecting with peers, Am. J. Phys. **78**, 748 (2010).

8. S. Y. Lin and C. Singh, Challenges in using analogies, Phys. Teach. **49**, 512 (2011).

9. S. Y. Lin and C. Singh, Using isomorphic problems to learn introductory physics, Phys. Rev. ST Phys. Educ. Res. **7**, 020104 (2011).

10. S. Y. Lin and C. Singh, Using an isomorphic problem pair to learn introductory physics: Transferring from a two-step problem to a three-step problem, Phys. Rev. ST Phys. Educ. Res. **9**, 020114 (2013).

11. A. Maries and C. Singh, To use or not to use diagrams: The effect of drawing a diagram in solving introductory physics problems, AIP Conf. Proc. 1513, 282 (2013). http://dx.doi.org/10.1063/1.4789707

12. A. Maries and C. Singh, A good diagram is valuable despite the choice of a mathematical approach to problem solving, Proceedings of the Physics Education Research Conference 2013, Portland, OR, edited by P. Engelhardt, A. Churukian, and D. Jones (2014), pp. 31–34. http://dx.doi.org/10.1119/perc.2013.inv.006

13. A. Maries, S. Y. Lin and C. Singh, Challenges in designing appropriate scaffolding to improve students' representational consistency: The case of a Gauss's law problem, Phys. Rev. Phys. Educ. Res. **13**, 020103 (2017).

14. A. Maries and C. Singh, Do students benefit from drawing productive diagrams themselves while solving introductory physics problems? The case of two electrostatics problems, Euro. J. Phys **39**, 015703 (2018).

15. A. Maries and C. Singh, Case of two electrostatics problems: Can providing a diagram adversely impact introductory physics students' problem solving performance? Phys Rev PER **14**, 010114 (2018).

16. A. Mason and C. Singh, Reflection and self-monitoring in quantum mechanics, Proceedings of the 2009 Phys. Ed. Res. Conference, Ann Arbor, MI, (M. Sabella, C. Henderson and C. Singh Eds.), AIP Conf. Proc., Melville, New York 1179, 197-200 (2009). http://dx.doi.org/10.1063/1.3266713

17. E. Yerushalmi, E. Cohen, A. Mason and C. Singh, What do students do when asked to diagnose their mistakes? Does it help them? I. An atypical quiz context, Phys. Rev. ST Phys. Educ. Res. **8** (2), 020109 (2012).

18. E. Yerushalmi, E. Cohen, A. Mason and C. Singh, What do students do when asked to diagnose their mistakes? Does it help them? II. A more typical quiz context, Phys. Rev. Special Topics Phys. Educ. Res. **8** (2), 020110 (2012).

19. C. Singh, Coupling conceptual and quantitative problems to develop expertise in introductory physics, Proceedings of the 2008 Phys. Ed. Res. Conference, Edmonton, Canada, (C. Henderson, M. Sabella, L. Hsu Eds.), AIP Conf. Proc., Melville New York 1064, 199-202 (2008). http://dx.doi.org/10.1063/1.3021253

20. A. Maries and C. Singh, Case of two electrostatics problems: Can providing a diagram adversely impact introductory physics students' problem solving performance? Phys Rev PER **14**, 010114 (2018).

21. L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, Cambridge, MA, Harvard University Press (1978).

22. H. Ginsberg and S. Opper, *Piaget's Theory of Intellectual Development*, Englewood Cliffs, NJ, Prentice Hall (1969).

23. G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 211-227 (1982).

24. D. Schwartz, J. Bransford and D. Sears, Efficiency and innovation in transfer in *Transfer of learning from a modern multidisciplinary perspective*, edited by J. Mestre, Information Age Publishing, Greenwhich, CT, (2005).

25. D. Maloney, T. O'Kuma, C. Hieggelke and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12-S23 (2001).

26. L. S. Shulman, Those who understand: Knowledge growth in teaching, Educ. Res. **15** (2), 4-31 (1986).

27. L. S. Shulman, Knowledge and teaching: Foundations of the new reform, Harv. Educ. Rev. **57** (1), 1-22 (1987).

28. P. M. Sadler, G. Sonnert, H. P. Coyle, N. Cook-Smith and J. L. Miller, The influence of teachers' knowledge on student learning in middle school physical science classrooms, Am. Educ. Res. J. **50** (5), 1020 (2013).

29. N. I. Karim, A. Maries and C. Singh, Exploring one aspect of pedagogical content knowledge of teaching assistants using the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res **14**, 010117 (2018).

30. N. I. Karim, A. Maries and C. Singh, Teaching assistants' performance at identifying common introductory student difficulties revealed by the Conceptual Survey of Electricity and Magnetism, 2017 Physics Education Research Conference Proceedings, 208-2011 (2018). http://dx.doi.org/10.1119/perc.2017.pr.047

31. A. Maries and C. Singh, Exploring one aspect of pedagogical content knowledge of teaching assistants using the TUG-K, Phys. Rev. ST PER **9**, 020120 (2013).

32. A. Maries and C. Singh, Performance of graduate students at identifying introductory students' difficulties related to kinematics graphs, Physics Education Research Conference Proceedings, p. 171 (2015). http://dx.doi.org/10.1119/perc.2014.pr.039

33. A. Maries and C. Singh, Teaching assistants' performance at identifying common introductory student difficulties in mechanics revealed by the Force Concept Inventory, Phys. Rev. ST PER **12**, 010131 (2016).

34. D. Hestenes, M. Wells and G. Swackhammer, Force Concept Inventory, Phys. Teach. **30** (3), 141-158 (1992).

35. C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys **71**(6), 607 (2013).

36. L. Rimoldini and C. Singh, Student understanding of rotational and rolling motion concepts, Phys. Rev. ST PER **1**, 010102 (2005).

37. L. Ding, R. Chabay, B. Sherwood and R. Beichner. Valuating an assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST PER **1**, 10105 (2006).

38. C. Singh and D. Rosengrant, Students' conceptual knowledge of energy and momentum, Proc. Phys. Educ. Res. Conf., Rochester, p. 123 (2001). http://dx.doi.org/10.1119/perc.2001.pr.018

39. J. Li and C. Singh, Developing and validating a conceptual survey to assess introductory physics students' understanding of magnetism, Euro. J. Phys. **38** (2), 025702 (2017).

40. C. Singh, Student understanding of symmetry and Gauss's law of electricity, Am. J. Phys. **74** (10), 923 (2006).

41. N. I. Karim, A. Maries and C. Singh, Impact of evidence-based flipped or active-engagement non-flipped courses on student performance in introductory physics, Can. J. Phys. **96** (4), 411-419 (2018).

42. A. Madsen S. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? Phys. Rev. ST PER **9**, 020121 (2013).

43. N. I. Karim, A. Maries and C. Singh, Do evidence-based active-engagement courses reduce the gender gap in introductory physics? Euro. J. Phys. **39**, 025701 (2018).

44. T. Majors and P. Engelhardt, Gender & LEAP Pedagogy: What does the Gender Force Concept Inventory have to say? PERC Proceedings p. 167 (2014) (available online).

45. G. C. Marchand and G. Taasoobshirazi, Stereotype threat and women's performance in physics International Journal of Science Education **35** (18), 3050 (2013).

46. M. Appel and N. Kronberger, Stereotypes and the achievement gap: Stereotype threat prior to test taking Educational Psychology Review **24** (4), 609 (2012).

47. C. McKown and R. S. Weinstein, The development and consequences of stereotype consciousness in middle childhood, Child Development **74**, 498 (2003).

48. L. Bian, S.-J. Leslie and A. Cimpian, Gender stereotypes about intellectual ability emerge early and influence children's interests, Science **355**, 6323 (2017).

49. S. Wheeler and R. Petty, The effects of stereotype activation on behavior: A review of possible mechanisms, Psychol. Bull. **127**, 797 (2001).

50. C. Steele and J. Aronson, Stereotype threat and the intellectual test performance of African Americans, J. Pers. Soc. Psychol. **69**, 797 (1995).

51. L. Strickler and C. Ward, Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance, J. Appl. Soc. Psychol. **34**, 665 (2004).

52. A. Maries and C. Singh, Stereotype threat? : Effects of inquiring about test takers' gender on conceptual test performance in physics, Proceedings of the 5th International Conference on Women in Physics, Waterloo, Canada, AIP Conf. Proc., **1697**, 120008-1 (2015). http://dx.doi.org/10.1063/1.4937713

53. A. Maries, N. I. Karim, and C. Singh, The impact of stereotype threat on gender gap in introductory physics, 2017 Physics Education Research Conference Proceedings, 256-259 (2018). http://dx.doi.org/10.1119/perc.2017.pr.059

2.0 PEDAGOGICAL CONTENT KNOWLEDGE OF TEACHING ASSISTANTS USING THE CONCEPTUAL SURVEY OF ELECTRICITY AND MAGNETISM

2.1 INTRODUCTION

2.1.1 Graduate Teaching Assistants

Graduate students in physics across the United States have been playing an important role in educating the next generation of students for a long time. In particular, in the US, it is quite common for physics graduate Teaching Assistants (TAs) to teach introductory physics recitation or lab sections which typically have lower enrollments than the "lecture" component of the course (20-40 compared to 100 or more in a lecture). In addition to the graduate TAs, in the last two decades, undergraduate TAs (sometimes referred to as Learning Assistants or LAs) have also played a role in educating students by, e.g., assisting faculty members in teaching large classes. Appropriate professional development of these TAs to help them perform their duties effectively is an important task. Physics education researchers have been involved in research on identifying common beliefs and practices among physics TAs that have implication for effective teaching [1-9]. For example, research suggests that sometimes graduate TAs struggle to understand the value of thinking about the difficulty of a problem from an introductory student's perspective and think that if they know the material and can explain it to their students in a clear manner, it will be sufficient to help their students learn [1, 3]. Also, while graduate TAs are able to recognize useful solution features and articulate why they are important when looking at sample introductory physics student solutions provided to them, they do not necessarily include those features in their

own solutions written for introductory physics courses [4-6]. Moreover, the TAs do not always engage in grading practices which are conducive to helping introductory physics students learn expert-like problem solving strategies and develop a coherent understanding of physics [7, 8].

It is also important to keep in mind that TAs may be given varying amounts of freedom regarding how to perform their teaching duties, depending on the instructor. However, discussions with the TAs who participated in this study and others from the University of Pittsburgh (Pitt) suggest that except for broader guidelines such as whether to discuss homework problems followed by a quiz or whether to have group problem solving [10-16] followed by a quiz in the recitation, the TAs often have considerable flexibility in how to perform their recitation duties. For example, many instructors meet with the TA only briefly at the beginning of the semester to outline general guidelines, e.g., answer student questions on the homework, solve problems on the board, and the TAs are left to their own devices for the rest of the semester except for some communication with the course instructor via email or during the grading of the exams. Thus, if TAs are knowledgeable about effective instructional approaches, they can make a significant contribution to introductory students' learning of physics in the recitations because they often have sufficient flexibility to lead the recitation in a manner that they think is conducive to student learning.

To help TAs learn about effective pedagogy, many institutions offer professional development programs which are sometimes discipline-specific [9, 17-19]. For more information about professional development programs and research on recruiting and educating future teachers, see Ref. [9] and references therein. The effectiveness of these professional development programs can be enhanced if those leading them are knowledgeable about TAs' conceptions regarding introductory physics students' difficulties [20]. For example, TAs may be largely unaware of certain introductory student alternate conceptions. If professional development

instructors preparing TAs discuss introductory students' alternate conceptions and engage the TAs in discussions about how to help introductory physics students learn, the TAs may be better prepared to conduct their teaching duties. It is even possible that in order to convince the TAs, the professional development instructors may have to share quantitative data on introductory physics student performance, which show that those alternate conceptions are common. This type of activity in TA professional development programs has the potential to enhance TAs' teaching effectiveness as they design, adopt and adapt activities to build on students' prior knowledge and help them develop a robust knowledge structure so that there is less room for those alternate conceptions. Similarly, if TA professional development instructors are aware that TAs know about certain student alternate conceptions, those can only be discussed briefly.

Thus, by focusing on what TAs know and do not know and gradually building their *pedagogical content knowledge*, or PCK for short [21-22] (more about PCK in the next section), they can be guided to learn and implement effective pedagogy. These considerations motivated us to carry out the research study discussed here using the Conceptual Survey on Electricity and Magnetism (CSEM), which is one of the many assessment tools often used to evaluate students' conceptual understanding of introductory concepts [23]. The goal of the present study was to evaluate TAs' knowledge of introductory student alternate conceptions in electricity and magnetism as revealed by the CSEM. For each item on the CSEM, the TAs were asked to identify the most common incorrect answer choice (MCI) selected by introductory physics students. This exercise was followed by a class discussion with the TAs related to this task, including the importance of knowing student difficulties and addressing them effectively in order for learning to be meaningful. We have found that this type of activity in a TA professional development course engenders a rich discussion about introductory student difficulties and promotes the importance of

thinking about their difficulties from their perspective in order to bridge the gap between teaching and learning. More information about potential uses of this type of activity in TA professional development is provided in the discussion and summary section.

2.1.2 Pedagogical Content Knowledge

There are several theoretical frameworks that inspire our research. These theoretical frameworks focus on the importance of the instructors familiarizing themselves with students' prior knowledge (including what students learn from traditional instruction) in order to scaffold their learning with appropriately designed curricula and pedagogies. In the context of this study, they point to the importance of being knowledgeable about student difficulties in order to help students learn better. For example, Piaget [24] emphasized "optimal mismatch" between what the student knows and where the instruction should be targeted in order for desired assimilation and accommodation of knowledge to occur. A related framework is the theory of conceptual change put forth by Posner et al. [25]. In this framework, conceptual changes or "accommodations" can occur when the existing conceptual understanding of students is not sufficient for or is inconsistent with new phenomena. They also suggest that these accommodations can be very difficult for students, particularly when students are firmly committed to their prior understanding, unless instructional design explicitly accounts for these difficulties. Within this framework, students can be motivated by an anomaly which provides a cognitive conflict and illustrates how their conceptions are inadequate for explaining a newly encountered physical situation. They can become dissatisfied with their current understanding of concepts and thereby make efforts to improve their understanding. But instructors must be aware of what conceptions students have and what

difficulties in learning physics these conceptions can lead to in order to design instruction that produces the desired cognitive conflict.

Being knowledgeable of what conceptions students have and the difficulties that these conceptions may lead to is one aspect of what Shulman defined as Pedagogical Content Knowledge (PCK) [21, 22]. Shulman defines PCK as the subject matter knowledge *for teaching*. In other words, PCK is a form of practical knowledge used by experts to guide their pedagogical practices in highly contextualized settings. Shulman writes "Within the category of pedagogical content knowledge, I include […] the most useful forms of representation of those ideas, the most powerful analogies, illustrations, examples, explanations, and demonstrations – in a word, the ways of representing and formulating the subject that make it comprehensible to others." In addition, according to Shulman, PCK also includes "an understanding of what makes the learning of specific topics difficult: the conceptions and preconceptions that students bring with them to the learning of those most frequently taught topics and lessons." [21]. Shulman developed the concept of PCK in response to the growing trend of proliferating general educational research in teacher preparation programs. The development of PCK was in part due to Shulman's previous research on the reasoning processes of physicians [26], which he found to be domain specific and contrary to the general assumption that certain physicians possess a general trait of diagnostic acumen which makes them better diagnosticians than others. Shulman generalized this observation to conclude that good teachers not only possess domain specific knowledge, but also possess more practical knowledge about teaching that is domain specific (i.e., PCK). Shulman therefore encouraged research on teachers' PCK and the types of teacher preparation programs that are likely to improve and/or develop teachers' PCK. Since Shulman introduced the concept of PCK, much has been written about it [27-42]. For example, Grossman [29] includes PCK as one of the "four

general areas of teacher knowledge [which are] the cornerstones of the emerging work on professional knowledge for teaching: general pedagogical knowledge, subject matter knowledge, pedagogical content knowledge, and knowledge of context" and argues that PCK (as opposed to their subject matter knowledge) generally has the greatest impact on teachers' classroom activities. Others have also stressed the importance of PCK in shaping instructional practice and discuss professional development programs which take PCK into account [36, 37]. For example, Borko and Putman [37] describe the Cognitively Guided Instruction Project, a multi-year program of curriculum development, professional development and research which has shown "powerful evidence that experienced teachers' pedagogical content knowledge and pedagogical content beliefs can be affected by professional development programmes."

Given the importance of PCK in shaping instructional practices, it is not surprising that researchers have attempted to document teachers' PCK [31, 33, 34] and others have attempted to document the development of teachers' PCK [35, 38]. However, these tasks are challenging to carry out for multiple reasons such as the fact that much of the knowledge teachers have of their practice is tacit [39, 40], or the fact that although there is a general consensus among researchers on PCK as a construct, its boundaries are not clearly delineated [41]. Also, extended observations are needed in order to recognize when teachers' PCK is instantiated in their practice [31]. To overcome some of these challenges, researchers have often used multi-method approaches to investigate teachers' PCK. For example, observational data are not sufficient because a teacher may use only a small portion of the representations he/she has at his/her disposal. In addition, observations do not provide insight into teachers' instructional decisions – we see what they are doing, but do not know why. Partly due to these issues, Loughran et al. [31] used both classroom observations and follow-up interviewing of teachers. The interviews encouraged teachers to

articulate their knowledge and explored alternative representations that the teachers did not use during the teaching sessions. This investigative approach is quite time-consuming both to carry out and analyze since both the observations and interviews provide lengthy qualitative data which require coding and analysis. Baxter and Lederman [42] provide a review of methods and techniques for studying PCK and the subject matter knowledge of teachers.

Partly due to all of the difficulties in carrying out an involved investigation of PCK, we developed a relatively straightforward method for delving into one particular aspect of PCK, namely knowledge of student difficulties with particular topics. This method makes use of standardized multiple choice tests developed by physics education researchers and quantitative data from students taking these tests. Teachers are provided with a copy of a particular test (e.g., CSEM), and for each item on the test they are asked to select what they expect would be the MCI selected by introductory students after being instructed in the relevant topic. Then, quantitative student data are used to quantify the extent to which teachers are knowledgeable about common student difficulties which are revealed by the incorrect answer choices students commonly select. Previous research with K-12 teachers [20] has found that on items which have a strong distractor (i.e., common student alternate conception), there is a large difference in learning gains between students taught by teachers who could identify the alternate conception and students taught by teachers who could not. It is therefore valuable to explore the extent to which teachers are knowledgeable about student alternate conceptions on items drawn from carefully designed standardized tests.

Two prior research studies conducted using the method described in the preceding paragraph used the Force Concept Inventory (FCI) [43, 44] and Test of Understanding Graphs in Kinematics (TUG-K) [45, 46]. The main findings from these studies are as follows:

- TAs were able to identify common student alternate conceptions in certain contexts, but struggled to identify them in other contexts.
- TAs sometimes expected certain answer choices to be the MCIs, when instead, those answer choices were selected by very few students.
- Think aloud interviews with TAs engaged in the task of determining the MCIs of introductory students suggested that the TAs were reflective and often had reasonable thoughts regarding how introductory students may be thinking about the questions. Interviews also suggested that the TAs were sometimes distracted by certain answer choices that were not common among introductory students, and reasoned that those answer choices would be common.
- TAs performed better in identifying common student alternate conceptions when working in groups compared to when working individually.

In this study, we extend our previous work and use the CSEM to investigate the extent to which physics graduate students enrolled in a semester-long course for teaching assistants are knowledgeable about introductory students' alternate conceptions related to electricity and magnetism. Knowledge of the introductory student alternate conceptions that graduate students are and are not aware of can be especially useful in designing effective professional development programs and inform future research on identifying and documenting the pedagogical content knowledge of TAs.

2.2 METHODOLOGY

2.2.1 Participants

The participants in this study were 81 first year graduate students (three separate cohorts) enrolled in a semester long mandatory pedagogy oriented TA training course at Pitt, which meets once a week for two hours. The graduate student population at Pitt is consistent with that of a typical research-based state university. The TAs teach introductory recitations and labs, typically in a traditional manner. In the recitations, the TAs primarily answer student questions, solve problems on the board and give students a quiz in the last 10-20 minutes. In the labs, the TAs start by demonstrating the procedures needed for that lab and the students closely follow the detailed procedures written in the lab manual.

Since this is the first and last pedagogy-oriented semester long course most physics graduate students at Pitt will ever take, it is designed to help graduate students become more effective teachers in general. During the course, graduate students get a general overview of cognitive research and PER during one two hour session and discuss their instructional implications. The graduate students are also introduced to curricula and pedagogies based on PER which emphasize the importance of being knowledgeable about introductory students' difficulties in order to help them develop expertise in physics. Each week, students complete various reflective exercises designed to help them perform their TA duties in a student-centered manner. For example, in one class, they discuss how to write effective problem solutions for introductory physics classes and what features should be included in solutions they hand out to students [4, 5, 6]. In another class, they are given sample student solutions and asked to grade them individually and in groups, followed by a discussion about how to grade students to help them learn better [7,

8]. In the second half of the semester each graduate student also leads an interactive discussion of the solution of a physics problem in the class in the manner in which he/she would lead a discussion if teaching introductory students and receive feedback from the other graduate students in the course (who are asked to pretend to be introductory students and ask questions) and the instructor. Thus, the TA training course (which is required of all first-year graduate students) is not focused on helping the TAs implement physics education research (PER) based curricula in specific recitations or labs (e.g., University of Washington tutorials [47]), but is a general introduction to pedagogical issues in physics teaching and learning.

This study focuses on issues related to the professional development of TAs who teach recitations and labs for introductory physics courses and typically have a closer association with introductory students than the course instructors and thus, they may be in an even better position (compared to the course instructors) to help introductory students learn if the TAs are versed in effective pedagogy. At Pitt, the TAs generally hold regular office hours and interact with introductory students in the physics resource room where they help introductory students with any questions related to their introductory physics courses. In addition, recitation class sizes are usually much smaller than the sizes of lecture classes taught by instructors. Therefore, TAs who are knowledgeable about introductory student difficulties related to electricity and magnetism concepts can play a significant role in improving introductory student understanding of these concepts and they can address introductory students' difficulties directly in their interactions with students.

In addition to the quantitative study, we conducted think-aloud interviews [48] with 11 TAs. Due to the availability of the TAs for individual interviews, some of the interviewed TAs participated in the quantitative study (they were in the TA professional development course in

which the quantitative study was carried out) but others were not. We also note that for the TAs who participated in the quantitative study, at least one year had passed before they were interviewed. Thus, the CSEM PCK task carried out in the TA professional development course was not fresh in their mind by the time of the interviews. Each of the 11 TAs had at least one semester of teaching experience in recitations. We did not find any qualitative differences in the reasoning of the TAs whether they had participated in the quantitative study earlier or not. More details about the interviews are provided in the methods section below.

2.2.2 Materials

The materials used in this study are the CSEM, which was given to the TAs in the TA training course as explained below, the post-instruction introductory students' data that were collected over a period of four years from an average of 388 algebra-based introductory physics students at 30 different institutions across the United States [23], the quantitative data obtained from the TAs in the TA training course, and the follow-up interview data. These data were used to determine introductory students' common alternate conceptions on each item on the CSEM, to assess the knowledge physics TAs have of introductory student alternate conceptions and to understand the reasoning TAs use when selecting certain incorrect answers as the most common.

2.2.3 Methods

In the quantitative study, the TAs were provided with the CSEM and, for each item on the CSEM, they were asked to identify what they expected to be the MCI of introductory students if students did not know the correct answer in a posttest (after traditional instruction in relevant concepts).

We refer to this task as the CSEM-related PCK task. In years two and three of the study, the researchers also asked TAs to predict the percentage of introductory physics students who answer each question on the CSEM correctly in a post-test (after traditional instruction in relevant concepts). We investigated data from each year separately and found very few differences between the different years. Therefore all the data were combined (for TAs' predictions on the percentage of students answering each question correctly, only years two and three were combined because this question was not asked during the first year). Each year, after the TAs completed the CSEM-related PCK task, there was a full class discussion about the tasks and why knowledge of student difficulties is critical for teaching and learning to be effective in general. The TAs were not prompted to explain their reasoning for their choices, but in the class discussion certain items on the CSEM were discussed in detail and the TAs mentioned their reasoning about why they expected certain incorrect answer choices to be most common among introductory students.

In order to obtain an in-depth account of TAs' reasoning (related to why they expect certain answer choices to be most common among introductory students), think-aloud interviews were conducted with 11 TAs. Certain questions were selected from the CSEM based on the research questions of the study (described in the next section). The main goal of the interviews was to identify possible reasons why TAs expected that certain answer choices would be common among introductory students when in fact those answer choices were not common. Thus, the quantitative data collected was used to identify questions in which this may be occurring and the interviews focused on those questions. For example, Q2 on the CSEM on which roughly half the TAs expected that answer choice D would be most common among introductory students, but this answer choice was only selected by 11% of introductory students (see Table 2.1). During the interviews, the TAs were given a copy of the CSEM and for each question selected to be discussed

in the interview, the TAs were asked to first identify the correct answer after which they were asked to identify the MCI while thinking aloud. They were not disturbed during this time unless they became quiet for a long time in which case they were asked to keep talking. After discussing all of the questions selected by the interviewer, the TA was sometimes asked to look back at some of the questions and provide more details about why he/she expected a particular incorrect answer choice to be the MCI if his/her reasoning was not clear enough when they were thinking aloud without being disturbed.

We note that the task given to TAs was framed such that they had to identify the MCI for each multiple choice question that introductory physics students would select *after* instruction if they did not know the correct answer (rather than *before* instruction), because individual discussions with some faculty members (who had taught introductory physics) indicated that they felt that they (and the TAs) had no way of knowing the "pre-conceptions" of introductory physics students at the beginning of the course. Their reluctance to contemplate introductory physics students' difficulties about electricity and magnetism before instruction motivated us to ask them to identify the MCI for each question if the introductory student did not know the correct answer *after* traditional instruction in relevant concepts. We note that it does not make a significant difference whether the question is phrased about introductory physics students' difficulties with each question in the post-test or pre-test because the common alternate conceptions of introductory physics students rarely changed after traditional instruction. An analysis of the pre- and post-test data in Ref. [23] for each item on the CSEM suggests that the percentage of students who had a certain alternate conception either decreased after instruction or remained roughly the same and the most common difficulties remained the same in the pre-test and post-test. Therefore, the performance of TAs at identifying the most common alternate conceptions after traditional

instruction also provides an indication of their understanding of the initial knowledge state of introductory physics students related to CSEM content. Since we asked the TAs to identify the alternate conceptions after traditional instruction, we performed our data analysis using the post-test data printed in Ref. [23].

In order to quantify TAs' performance at identifying the alternate conceptions of introductory students, scores were assigned to each TA. A TA who selected a particular incorrect answer choice as the MCI in a particular question received a PCK score which was equal to the fraction of introductory physics students who selected that particular incorrect answer choice. If a TA selected the correct answer choice as the MCI (a rare occurrence), his/her data was removed *only* for that specific question because he/she was explicitly asked to indicate the *incorrect* answer choice which is most commonly selected by introductory students if they did not know the correct answer after traditional instruction in relevant concepts.

For example, on question 1, the percentages of algebra-based students who selected A, B, C, D and E are 4%, 63%, 23%, 7% and 3%, respectively (as shown in Table 2.1). Answer choice B is correct, thus, the PCK score assigned to TAs for each answer choice if they selected it as the MCI would be 0.04, 0, 0.23, 0.07 and 0.03 (A, B, C, D and E). The total PCK score a TA would obtain on the task for the entire CSEM can be obtained by summing over all of the questions (this is referred to as "CSEM-related PCK score"). These scores can be used to determine if TAs performed better than if they were randomly guessing. More details on how this was done are provided in the supplementary material.

We note that the approach used to determine the CSEM-related PCK score weighs the responses of TAs by the fraction of introductory physics students who selected a particular incorrect response. This weighting scheme was chosen because the more prevalent an introductory

student difficulty is, the more important it is for a TA to be aware of it and take it into account in their instruction. Furthermore, this approach also provides a reasonable PCK score when there is more than one common alternate conception. For example, if a question has two incorrect answer choices that are commonly selected by introductory students, e.g., Q29 in which 26% of introductory students selected A and 23% selected B (both incorrect). If all the TAs selected answer choice A as most common, their PCK score would be 100%, but if half the TAs selected A and half select B, their PCK score would be 92.5%.

The researchers jointly determined a heuristic that if a particular alternate conception is held by more than 20% of introductory students in a particular context, it is important for TAs to be aware of this alternate conception. Therefore, some of the discussion here is focused on the questions on the CSEM in which the introductory physics students' national data showed that at least 20% of them selected an incorrect answer choice in the post-test (after instruction). Table 2.1 shows that there were 26 such questions (out of 32 total questions) and the other 6 questions did not have such distractors.

2.2.4 Research Questions and Approach for Investigation

The following two research questions (RQ1, RQ2) were developed for the purpose of investigating the CSEM-related PCK of the TAs. For each research question, we provide motivation for investigating it as well as details about the methods used to investigate it.

RQ1:(i) *Are there situations in which a significant fraction of TAs select answer choices that very few introductory students select? What are some common examples of reasoning that TAs use to select these answer choices?*

(ii) *What alternate conceptions do TAs struggle to identify?*

(iii) *What alternate conceptions can TAs identify?*

As noted, knowledge of introductory student alternate conceptions can be helpful in determining the pedagogical approaches that may be effective in helping students learn better. Therefore, TAs' knowledge of introductory student difficulties can play an important role in improving introductory student learning. In addition, TAs should also have reasonable expectations regarding how many introductory students have certain alternate conceptions. If a TA significantly overestimates the prevalence of a certain type of alternate conception (e.g., he/she thinks that 50% of students have the alternate conception, whereas the percentage of introductory students with that difficulty is less than 10%), the TA may spend considerable time and effort attempting to help students with something the majority of them already know, and thus not use class time effectively in addressing student difficulties which are more common. Similarly, if a TA underestimates the prevalence of a certain type of alternate conception, he/she is unlikely to consider instructional strategies to address it. This prompted us to investigate instances in which a significant fraction of the TAs select answer choices on the CSEM which they think are most commonly selected by introductory students, while those answer choices are actually selected by very few introductory students, and instances in which few TAs identify common student alternate conceptions. Knowledge of what TAs think are the common alternate conceptions of introductory students, but which are instead not common, and also what alternate conceptions are common among introductory physics students, but the TAs are not aware of them, can be valuable for developers of professional development programs because it can provide them with an understanding of TAs' prior knowledge regarding introductory students' common reasoning patterns.

Table 2.1 provides detailed data on TAs' performance in identifying the alternate conceptions of introductory students that are discussed in detail in the results section. In particular, we discuss the following:

- Alternate conceptions which many TAs expected to be most common among introductory students, which were instead not common among introductory students,
- Alternate conceptions which were common among introductory students which were not identified very well by TAs as the most common,
- Alternate conceptions that were common among introductory students which the majority of TAs were able to identify, and
- Qualitative results from detailed think-aloud interviews with 11 TAs which focused on what common reasoning TAs used to select certain answer choices as most common (e.g., answer choices which were not common among introductory students).

We note the following about the interviews: in general, during the interviews, the TAs were reflective and sometimes thought back to when they were teaching introductory physics in recitation themselves. In some of the questions, they were able to identify the MCI and had good ideas about the common difficulties of introductory students. However, an important goal of the interviews was to identify the reasoning the TAs commonly use when they select answer choices which were not very common among introductory students. Therefore, the discussion focusing on this aspect in a particular question should not be taken as an indication that the interviewed TAs did a poor job at identifying common alternate conceptions of introductory students on those questions.

RQ2:(i) *To what extent are TAs able to predict the difficulty of the questions?*

(ii) *To what extent is TAs' ability to identify introductory students' alternate conceptions correlated with their ability to predict the difficulty of a question?*

If a TA has good knowledge of introductory students' alternate conceptions, he/she may be able to also predict how difficult a question can be for introductory students. As mentioned earlier, in two out of three years of study, we asked a cohort of 56 TAs to also predict the difficulty of each question by estimating the percentage of introductory students who answered each question correctly after traditional instruction in relevant concepts. We then ran a correlation analysis between TAs' accuracy in estimating the difficulty of a question and their average PCK score. If the two are correlated, we would expect to see a negative correlation. In other words, the better the TAs are in estimating the difficulty of a question (i.e., the difference between TAs' estimated percentage and actual percentage is closer to zero), the larger the PCK score.

2.3 RESULTS

We note that the common incorrect answer choices of introductory students are similar for both algebra-based and calculus-based classes (see Ref. [23]). Therefore, the researchers performed the analysis of the CSEM-related PCK performance with the student data from algebra-based classes in Ref. [23] as discussed below:

2.3.1 Performance of TAs in Identifying Introductory Physics Students' Alternate Conceptions Related to the CSEM

There are 26 questions (out of 32) on the CSEM which reveal alternate conceptions held by 20% or more of introductory students. Analysis of the CSEM-related PCK score of the TAs was conducted on each of these questions and the results are displayed in Table 2.1.

Table 2.1 shows all CSEM items, the percentages of introductory physics students who answered each question correctly, the percentages of introductory students who selected each incorrect answer choice ranked from most to least common, the incorrect answer choices most commonly selected by TAs (as most common among introductory students), and the percentages of TAs who selected these answer choices. Correct answers are indicated by the green shading in Table 2.1, and incorrect answer choices selected by 20% or more introductory students are indicated by the red font. We note that Table 2.1 lists the averages while Figs. 2.15 – 2.18 in the chapter appendix can be consulted for the distribution of TA's predictions for the most common incorrect answer choices of introductory students for each of the CSEM questions. In addition, the second column (titled ">RG") in Table 2.1 indicates whether TAs performed better than random guessing (RG) in identifying the MCI for a particular question: "Yes"/blank field indicate that TAs performed/did not perform better than random guessing. Table 2.1 also shows the normalized average CSEM-related PCK scores of the TAs. Their scores were normalized on a scale from zero to 100 because for each question on the CSEM, there is a minimum and a maximum possible score, which correspond to the smallest and largest fractions of introductory physics students who selected a particular incorrect answer choice among the four incorrect answer choices. The normalization was done in the following manner:

$$\text{Normalized PCK Score} = 100 \times \frac{\text{Average PCK Score} - \text{Minimum Possible PCK Score}}{\text{Maximum Possible PCK Score} - \text{Minimum Possible PCK Score}}$$

The normalized CSEM-related PCK score is therefore zero if the TAs obtained the minimum possible score and 100 if they obtained the maximum possible score. The normalized score also provided a means to compare TAs' CSEM-related PCK performance for different questions which have different minimum and maximum possible CSEM-related PCK scores. The researchers jointly determined a heuristic that the performance of the TAs was 'good' (and shaded green in Table 2.1) if their normalized CSEM-related PCK score was more than 2/3 of the maximum possible score, 'average' (and shaded yellow) if their normalized score was between 1/2 and 2/3 of the maximum possible score and 'poor' (red shading) if their normalized score was less than 1/2 of the maximum possible score.

2.3.2 Results Relevant to Each Research Question

RQ1:(i) *Are there situations in which a significant fraction of TAs select answer choices that very few introductory students select? What are some common examples of reasoning that TAs use to select these answer choices?*

Table 2.1 shows that out of the 26 CSEM questions in which students had an alternate conception, in 18 of those questions, TAs' average normalized PCK scores were not good (less than 2/3 of the maximum possible). Out of those 18 questions, for 12 of them, TAs' PCK scores were average (between 1/2 and 2/3 of maximum possible normalized score) and for the other 6 they were poor (less than 1/2 maximum possible normalized score). We now turn to discussing questions in which TAs' normalized PCK score was either poor or average and focus on the questions which revealed that many TAs expected introductory students to have certain alternate conceptions, while few

introductory students actually had those alternate conceptions. In this section, we group questions together based on the concepts involved.

2.3.2.1 Charge distribution on conductors/insulators (Q1, Q2)

Q1 and Q2 ask about what happens to an excess charge placed at some point P on a conducting (Q1) or insulating hollow sphere (Q2). For Q1, introductory students' most common alternate conception (23% of introductory students) was that the charge distributes everywhere on the inside and outside of the metal sphere. On Q2, introductory students had two alternate conceptions: that the charge distributes itself everywhere on the outside of the sphere (i.e., not distinguishing between insulating and conducting – answer choice B selected by 21% of introductory students) and that there will be no charge left (answer choice E, selected by 19% of introductory students). On both Q1 and Q2, TAs' average PCK score is moderate (59% and 57%, respectively), in large part because many TAs expected that the MCI is choice D for both questions, namely that most of the charge is at point P, but some of it will spread over the sphere (19% and 49% of the TAs selected choice D, but only 7% and 11% of introductory students selected this choice in Q1 and Q2, respectively).

On Q1, some of the TAs reasoned that D would be the MCI because students would expect that the charges would move, but that there isn't enough force to move all the charges everywhere around the sphere, or that it takes more than a few seconds for the charge to spread everywhere and therefore some will remain at point P. For example, one interviewed TA stated: "They [students] don't expect that for a metal [there is] enough push in order to move all the charges from that point [P]." Another interviewed TA motivated his selection of choice D as the MCI by stating: "Most people probably think it's D […] because they might not recognize that it has to be an instantaneous distribution of charge. So they recognize that the charge will have to spread over

the surface, and since we know it's metal, I'm assuming they understand a conductor won't have charge on the inside. It [charge] is all gonna be on the surface, but they might assume that the majority of the charge hasn't fully distributed yet."

On Q2, the TAs' most common reasoning for selecting answer choice D was that it was the incorrect answer choice that is most similar to the correct one and that introductory physics students may have some understanding that an insulating sphere is different from a conducting sphere, but would not fully understand it. For example, one TA said: "If they understand this is insulating material [i.e., they do not miss this information when reading the question], they will choose D […] because they know something about insulating that it is not like the conducting, but they [may not know] that the charge will stay at the position [P]." This certainly seems reasonable, however, it appears that few introductory students selected this answer choice.

Table 2.1. Questions on the CSEM, percentages of introductory algebra-based physics students who answered the questions correctly in a post-test, percentages of introductory students who selected each incorrect answer choice ranked from most to least common, the percentage of TAs who selected each incorrect answer choice as most common among introductory students, and normalized average PCK score. The first column of the table lists the CSEM question numbers and the second column titled "> RG" shows a "Yes" when the TAs on average performed better than random guessing (RG). The details of how this analysis was carried out is described in the appendix.

CSEM Item #	>RG	Correct Answer	Intro Student Choices				TA Choices (Individually)				Normalized Average PCK Score
			1st	2nd	3rd	4th	1st	2nd	3rd	4th	
1	Yes	63% B	23% C	7% D	4% A	3% E	54% C	19% D	15% E	11% A	58.9%
2		42% A	21% B	19% E	11% D	5% C	49% D	25% B	16% E	10% C	57.4%
3	Yes	76% B	9% C	8% D	5% A	0% E	54% A	26% D	19% C	1% E	72.0%
4	Yes	40% B	32% C	21% D	5% A	2% E	57% C	22% A	21% D	0% E	72.4%

5	Yes	32% C	22% D	20% B	14% A	11% E	50% D	32% A	16% B	1% E	72.3%
6		67% E	13% B	10% C	7% A	4% D	51% A	34% B	12% C	3% D	58.8%
7	Yes	31% B	42% C	19% A	5% D	2% E	45% A	43% C	6% D	6% E	62.3%
8	Yes	53% B	21% D	10% C	8% E	5% A	43% D	27% E	16% C	14% A	53.1%
9		52% B	16% D	12% C	10% A	5% E	36% D	23% E	22% C	19% A	58.4%
10		35% C	25% E	20% B	12% D	6% A	47% B	25% A	20% D	9% E	49.4%
11	Yes	33% E	30% A	14% B	13% C	9% D	45% A	26% C	22% D	6% B	51.9%
12		67% D	13% C	9% A	8% B	2% E	31% C	29% B	23% A	17% E	61.5%
13	Yes	51% E	27% A	20% B	1% C	0% D	56% A	42% B	1% C	1% D	87.1%
14	Yes	16% D	54% A	13% E	9% B	4% C	46% A	24% E	18% B	11% C	52.1%
15	Yes	24% A	34% C	24% B	9% D	8% E	74% C	17% B	5% E	4% D	84.1%
16		32% E	22% B	17% D	13% A	13% C	35% A	25% D	24% C	16% B	27.3%
17	Yes	51% E	23% C	16% B	6% D	2% A	63% C	18% B	17% D	2% A	77.7%
18	Yes	47% D	28% E	17% C	4% B	2% A	52% E	37% C	7% B	3% A	74.2%
19	Yes	34% A	25% B	14% C	11% D	10% E	61% B	18% E	11% D	10% C	63.9%
20	Yes	17% D	32% C	20% B	18% A	8% E	44% C	25% A	20% B	11% E	64.3%
21		44% E	21% C	15% A	8% B	8% D	41% C	31% B	17% A	12% D	49.9%
22	Yes	32% D	28% C	22% A	11% B	4% E	59% C	22% A	12% E	7% B	77.6%

23	Yes	45% A	15% B	13% C	11% E	9% D	48% D	20% E	17% B	16% C	33.8%
24	Yes	25% C	45% B	19% D	8% E	2% A	48% D	45% B	4% A	3% E	64.5%
25		48% D	20% C	12% B	11% A	5% E	35% E	25% C	25% A	16% B	42.1%
26		49% A	21% D	11% B	6% C	6% E	47% C	29% D	17% B	7% E	34.9%
27		40% E	23% D	19% A	8% C	5% B	46% A	24% C	18% B	12% D	51.6%
28	Yes	40% C	35% E	12% B	8% A	3% D	55% E	32% A	11% B	2% D	63.1%
29	Yes	23% C	26% A	23% B	19% D	6% E	51% B	26% A	16% D	7% E	80.2%
30	Yes	48% A	15% C	14% D	9% E	7% B	66% D	17% C	10% E	7% B	77.2%
31		26% E	25% C	18% A	17% D	15% B	47% D	20% C	17% B	16% A	34.3%
32		18% D	40% B	23% A	16% C	1% E	42% A	31% E	16% C	11% B	41.1%

x%	TAs' CSEM-related PCK score is less than 50%
x%	TAs' CSEM-related PCK score is between 50% and 67%
x%	TAs' CSEM-related PCK score is more than 67%

2.3.2.2 Coulomb's force law (Q3, Q4, Q5, Q6, Q7, Q8)

On Q3 and Q6 there are no strong alternate conceptions and on Q4 and Q5 (shown in Fig. 1), the TAs identified the most common alternate conceptions quite well (PCK score is over 70% on both of these questions). Moreover, on Q7, there is one strong distractor (answer choice C selected by 42% of students) and another answer choice (A) is selected by 19% of the students. Nearly identical percentages of TAs selected these two answer choices (43% and 45% of the TAs selected choices C and A, respectively) as the MCI. Since choice C is much more common than choice A, this resulted in only a moderate PCK score on this question (62%).

For questions 3 -5:
Two small objects each with a net charge of +Q exert a force of magnitude F on each other.

We replace one of the objects with another whose net charge is +4Q:

3. The original magnitude of the force on the +Q charge was F; what is the magnitude of the force on the +Q now?

(a) 16F (b) 4F (c) F (d) F/4 (e) other

4. What is the magnitude of the force on the +4Q charge?

(a) 16F (b) 4F (c) F (d) F/4 (e) other

Next we move the +Q and +4Q charges to be 3 times as far apart as they were:

5. Now what is the magnitude of the force on the +4Q?

(a) F/9 (b) F/3 (c) 4F/9 (d) 4F/3 (e) other

Figure 2.1. Questions 3, 4 and 5 on the CSEM.

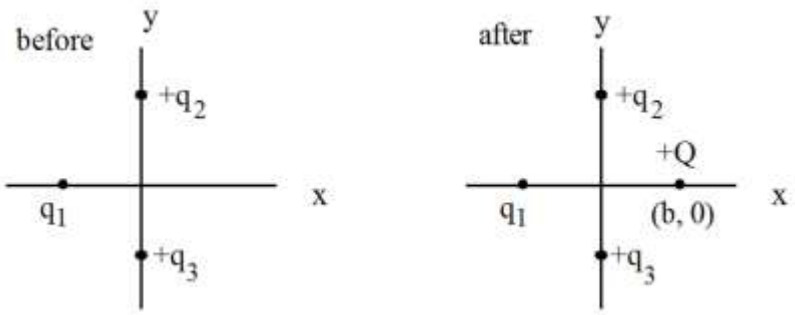

Figure 2.2. Figure provided for Q8 on the CSEM.

Q8 provides students with the two situations depicted in Fig. 2.2 and states that in the configuration on the left, charges q_2 and q_3 are positive and that the net force acting on q_1 as the result of its interaction with the two charges points in the positive x direction (to the right). The question asks what happens to the force acting on q_1 when another positive charge ($+Q$) is placed at the location shown in the configuration on the right. The MCI (choice D) selected by 21% of

introductory students is that the force will increase and its direction may change due to the interaction between Q and charges q_2 and q_3. While almost half the TAs (43%) selected this as the MCI, nearly one third (27%) of the TAs select answer choice E, which states that the answer cannot be determined without knowing the magnitude of q_1 and/or Q. However, this answer choice was selected by only 8% of introductory students. In interviews, some of the TAs also selected choice E as the MCI. One interviewed TA, for example, motivated selecting choice E by stating: "I think most of them [introductory students] will go with E […] because they might think that F is kq_1q_2 divided by r [squared] and then they think, 'ok, nothing is [given], q is not [given], r is not [given]', then they cannot decide [what happens to] the force." It appears that some of the TAs think that students may remember the equation for the electric force acting between two charges, but since none of the information is explicitly given (i.e., by providing values for the charges and distances), the electric force cannot be calculated so the question cannot be answered. However, it appears that very few introductory students may be reasoning this way since only 8% of them selected this answer choice.

2.3.2.3 Relation between electric field and force (Q10, Q12, Q15)

Q10 on the CSEM states that a positive charge is released from rest in a uniform electric field and asks about its subsequent motion. The two most common alternate conceptions are that the charge remains at rest (answer choice E selected by 25% of introductory students and 9% of TAs) and that it will move at constant velocity (answer choice B selected by 20% of introductory students). Answer choice A is similar to choice B except that it says that the charge moves at constant speed instead of constant velocity and only 6% of the introductory students selected this answer choice. However, 25% of the TAs selected this answer choice, which partly accounts for their poor average PCK score (49%). During the interviews, some of the TAs who selected this answer choice did

not seem to consider B very carefully. For example, one TA stated: “They might think it will go at constant speed because the field is uniform so the effect is constant throughout the path.” However, a more common occurrence in the interviews was for TAs to consider both choices A and B as the most common and either say they are not sure which one is more common or that introductory students would select among these two answer choices equally. It is possible that in the quantitative study conducted in the TA professional development course, TAs had similar considerations and some TAs opted for choice A while others opted for choice B as the most common. However, as shown in Table 2.1, much fewer introductory students selected choice A compared to choice B. For the other two questions in this grouping, there were either no alternate conceptions or the TAs performed well at identifying them (on Q15, TAs exhibited the second highest normalized average PCK score).

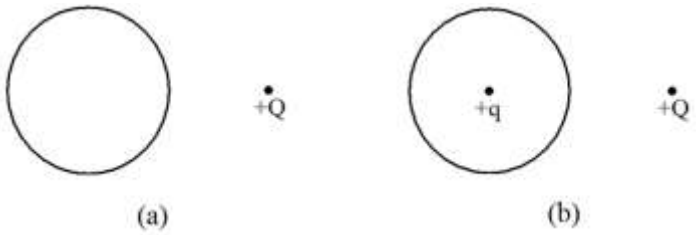

Figure 2.3. Diagrams provided for Q13 (a) and Q14 (b) on the CSEM.

2.3.2.4 Induced charge and electric field/force (Q13, Q14)

Q13 and Q14 provide students with the diagrams shown in Fig. 2.3. In Q13, the sphere is hollow and conducting and has an excess positive charge on its surface. The question asks for the direction of the electric field at the center of the sphere. In Q14, the sphere is also hollow and conducting, but it has no excess charge, and the question asks about the forces acting on the two charges. On both of these questions, the most common difficulty of introductory students is to not recognize that the conducting sphere alters the electric field/forces. Thus, on Q13, 27% of them selected

choice A on which the electric field is to the left (as though the sphere does not affect it) and on Q14, 54% of them also selected choice A for which the forces that the two charges feel are the same (once again, as though the sphere does not affect the forces). The TAs' PCK performance is very good on Q13 (87% – the highest of all questions on the CSEM), but only moderate on Q14 (52%), on which 53% of the TAs selected other answer choices (B, C, and E), which combined were selected by only 26% of introductory students. In the interviews, TAs who selected answer choice A as the MCI on Q13 usually did so because they expected that introductory students would only think about the electric field caused by the +Q charge and ignore the metal sphere. For example, one TA who selected A said: "Maybe someone would say leftward because they think of the positive being the source so they think of it making a [field] line and the [field] line is going outward from the charge, and they think it's just going to go straight through the sphere." On the other hand, on Q14, this same TA said that students would select choice E most commonly because they may think that the charge distribution on the sphere affects the forces that the two charges are experiencing. "They might think that little q at the center of the sphere […] is feeling forces from the charges that are distributed along the surface [of the sphere], and big Q here might feel force from this guy [q] and all the surface charges [on the sphere]." Other interviewed TAs cited similar reasoning for selecting choice E in Q14.

On Q13, another TA selected choice A as the MCI and stated that introductory physics students may ignore the effect of the sphere. When looking at Q14, this TA explicitly mentioned his previous answer and stated: "My thought is similar to the last one, to kind of just ignore the sphere. So A, maybe." In other words, students may ignore the effect of the sphere and select A. But after noticing answer choice E, he changed his mind and went on to say the following: "I think E [may be most common] because they might realize that the sphere does do something to change

things, so they think ‘ok, I know [the forces would normally be] equal and opposite, but now there’s a sphere here, so I don’t know exactly how that works’ [i.e., what the effect of the sphere is] so they’ll just throw in something [i.e., include some effect due to the sphere], so E is that something.” It appears that this TA was aware that introductory students may be guided by similar incorrect thinking (conducting sphere will not have an effect) on Q14 as on Q13, but on Q14, selected the answer choice which incorporates a correct idea (conducting sphere has an effect), but is missing another idea in order to be fully correct. In many other questions, TAs often selected answer choices which fit this category. For example, as mentioned earlier, on Q1, some TAs thought that introductory students would select answer choice D, which states that some of the charge does spread over the sphere – a partially correct answer. Similarly, on Q2, some TAs selected the same answer, which is partially correct because some of the charge does remain at point P. They also sometimes explicitly noted that they were selecting this answer choice as the MCI because it is the one that is most similar to the correct answer. On Q10, many interviewed TAs considered answer choices A and B, stating they expected that introductory physics students would be aware that the charge should move, but they may not know that it moves with a constant acceleration (more examples will be discussed below). While sometimes using this strategy to identify the MCI may provide a reasonable answer choice (i.e., one that is fairly common among introductory students), it often misled the TAs into selecting an answer choice that was not very common – as was the case on Q1, Q2, and Q14 (and other questions that are discussed below). On Q14, for example, this reasoning led some TAs to select choice E as the MCI. However, this answer choice was only selected by 13% of introductory students.

On Q14, 18% of the TAs in the quantitative study selected answer choice B as the most common among introductory students, but only 9% of introductory students selected it. One

interviewed TA who selected choice B as the MCI to Q14 noted that introductory students may reason in the following way: "Inside the conductor there is no field. But they might think the sphere is shielding the field due to the inside charge also. So, everything is shielded and there is no force [i.e., neither +q nor +Q experience a force]." Other TAs who selected choice B used very similar reasoning. Similar to TAs' reasoning for selecting choice E discussed earlier, answer choice B also incorporates a partially correct idea: the metal sphere "protects" the charge inside from the effect of outside charges, which is partly why many interviewed TAs selected it as the MCI.

2.3.2.5 Relation between electric potential and electric field/force (Q16, Q18, Q19, Q20)

On Q16, introductory students' responses are spread over the four incorrect choices almost evenly. On Q18, the TAs performed well in identifying the alternate conception. Moreover, on Q19 and Q20, the TAs appear to be able to identify the alternate conceptions.

2.3.2.6 Work/Electric potential energy (Q11, Q17)

Q11 asks what happens to the electric potential energy of a positive charge after being released from rest in a uniform electric field. The most common alternate conception of introductory students is that it remains constant because the electric field is uniform (answer choice A selected by 30% of introductory students). A much less common answer choice is choice C, namely that the electric potential energy will increase because the charge will move in the direction of the electric field (selected by only 13% of introductory students). However, 26% of the TAs selected this answer choice, and during an interview, one TA reasoned that perhaps students are thinking about total energy instead of electric potential energy (or perhaps they are confusing kinetic and electric potential energy): "It (the charge) has an acceleration and velocity is increasing right? So they [students] may think that the potential [energy] should increase because velocity is

increasing." Another interviewed TA who selected choice C as the MCI selected it for a very similar reason. On Q17, TAs' average PCK was good (78%).

2.3.2.7 Force on/motion of charged particle in a magnetic field (Q21, Q22, Q25, Q27)

Q21 asks what happens to a positive charge that is placed at rest in a magnetic field. The most common alternate conception of introductory students is that the charge moves in a circle at constant speed (answer choice C selected by 21% of introductory students). On this question, many TAs thought that introductory students may confuse electric and magnetic field and thereby conclude that the charge moves with constant acceleration (answer choice B selected by 31% of TAs), but this answer choice is very rarely selected by students (only 8% of them selected this choice) and thus, TAs' average PCK score on this question is poor (just below 50%). For example, one interviewed TA stated: "I can see people confusing or essentially just ignoring that it's a magnetic field thinking that it should do the same thing as it does in an electric field, so, constant acceleration. Yea, that would be my guess – B, they would think that it would do the same thing it does in an electric field."

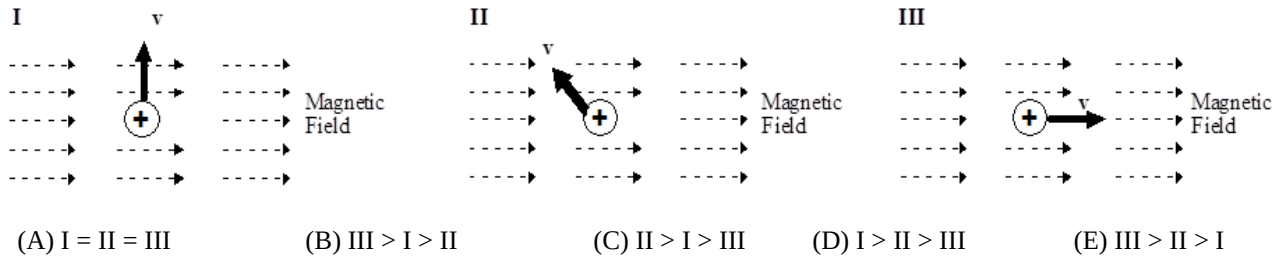

Figure 2.4. Three situations and answer choices provided in Q25 on the CSEM.

Q25 provides the three situations shown in Fig. 2.4 of a positive charge moving in an external magnetic field and asks students to rank them according to the magnitude of the magnetic force. Interviews suggest that TAs struggled to identify the MCI which is that the force is largest in situation II (where the charge moves 'against' the magnetic field) and least in situation III (where

the charge moves 'with' the electric field), and situation II is in between – answer choice C selected by 20% of introductory students. TAs' selections however are quite varied, with a significant percentage of them opting for each incorrect answer choice, which resulted in poor PCK performance (42%).

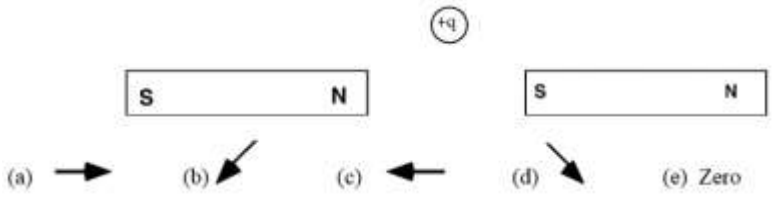

Figure 2.5. Physical situation and answers provided for Q27 on the CSEM.

Interviews suggest that the TAs had difficulty determining how introductory physics students may reason about this question incorrectly. The TAs sometimes opted for choice A (same force in all situations) because they expected that some students may only recall *qvB* as the magnetic force on a charge moving in a magnetic field and thus conclude that the forces are equal in the three situations. If they did not select this answer choice, they usually started by stating that when the velocity and magnetic field are in the same direction, students may think that this leads to the largest force. For example, one TA stated: "They [students] are thinking 'oh, the magnetic field is pushing it along in this direction and it's already moving in that direction' so that's just compounding the effect [i.e., force is largest in situation III]." Other interviewed TAs reasoned in a similar way, but after concluding that students may think the force is largest in situation III, they had difficulty applying the same reasoning to situations I and II. They sometimes stated that for situation II, students may think that the acceleration is least because the charge is moving in a direction (partly) opposite to the magnetic field, and conclude that the force is least in situation II (and select B). Other TAs stated that perhaps introductory students are somehow thinking of the dot product instead of the cross product and conclude that choice E is the most common answer.

Yet other TAs, after considering situation II, changed their minds because they thought that since the charge is moving 'against' the magnetic field, students may think that the field is exerting the largest force. This was one of the questions on the CSEM which took the TAs the most time to answer (i.e., determine what they expected would be the MCI). One TA, after trying to figure it out for a while, just gave up and said that maybe introductory students would just rank the situations in the opposite order (i.e., not read the question correctly and think the situations should be ranked from least to greatest).

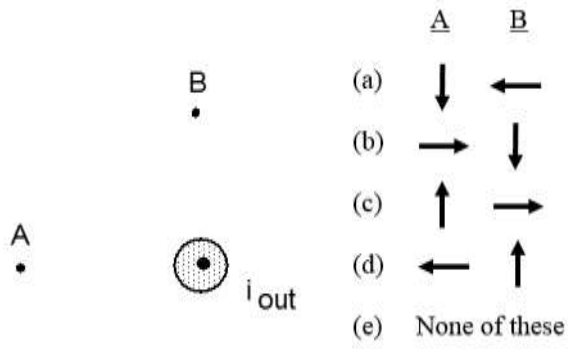

Figure 2.6. Diagram and answer choices for Q26 on the CSEM

Q27 shows a positive charge placed at rest near two magnets, the one on the left being three times stronger than the one on the right (see Fig. 2.5). It asks for the magnetic force acting on the charge and provides the answer choices shown in Fig. 2.5. On this question, the MCIs are choice A (19%) and choice D (23%). Only 12% of the TAs selected choice D, and 24% of them selected choice C – an answer choice selected by only 8% of introductory students, which resulted in a moderate PCK score (52%) on this question. One interviewed TA selected choice C because he expected students to think that the magnet on the left is pushing the charge towards the right and the magnet on the right is pushing the charge towards the left. When asked why he expected students to think this way he stated that he did not know how to explain it, it was just his gut feeling based on his experience teaching recitations.

2.3.2.8 Magnetic field caused by a current (Q23, Q26, Q28)

Q26 is shown in Fig. 2.6. On this question, the most common alternate conception of introductory students is that the magnetic field is radially outward from the wire (answer choice D selected by 21% of introductory students). On this question, 47% of the TAs selected answer choice C in which the direction of the magnetic field is opposite to the correct direction (i.e., clockwise instead of counterclockwise), but only 6% of introductory students selected this answer choice. This resulted in a low average PCK score on this question (35%). All the interviewed TAs who selected this answer choice essentially said that students may either use their left hand or use the right hand rule incorrectly; however, the choices selected by many introductory students do not suggest this as a major difficulty.

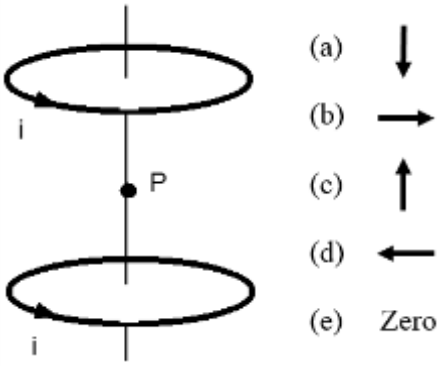

Figure 2.7. Diagram and answer choices for Q28 on the CSEM

On Q26, some interviewed TAs used similar reasoning as some of the TAs who selected choice E on Q14 – students have some correct ideas (try to use the right hand rule), but are not fully correct (obtain the incorrect direction). It is important to point out that after recognizing that students may be answering the question incorrectly for this reason (which does not seem to be common), the interviewed TAs did not consider all the other answer choices carefully, and did not realize that students may have other alternate conceptions, namely that the magnetic field would be radially outward from the wire (i.e., confusion between electric and magnetic field). After the

TAs answered all the other questions in the interview, they were often asked to return to this question and think about whether they expected that any introductory students would select answer choice D (radially outward magnetic field). After being asked to consider this answer choice explicitly, they were often able to recognize the alternate conception guiding introductory students to select choice D and some interviewed TAs wanted to change their original answer. Similarly to Q14, some TAs attempted to identify common alternate conceptions on Q26 by arguing that introductory students may have some correct ideas, but miss something that causes them to not have the fully correct answer. However, it appears that for this question (and others mentioned earlier), this type of reasoning from the TAs often steered them in the wrong direction and caused them to identify an answer choice that is not common among introductory students while missing the most common alternate conception.

Q28 on the CSEM provides the diagram and answer choices shown in Fig. 2.7. The loops shown in Fig. 2.7 carry currents of equal magnitude and the question asks for the direction of the magnetic field at point P. The MCI is that the two magnetic fields created by the two wires cancel out (answer choice E, selected by 35% of introductory students). Here, the majority of TAs (55%) selected this answer choice, but 32% of them selected choice A (an answer choice selected by only 8% of introductory students), which resulted in a moderate PCK score (63%). Similarly to Q26 discussed above, all of the TAs who selected this answer choice during interviews claimed that introductory physics students may use the right hand rule incorrectly and obtain the incorrect direction, however, it appears that very few students do this.

2.3.2.9 Faraday's law (Q29, Q30, Q31, Q32)

On Q29 and Q30, TAs' performance is quite good, and on Q31, introductory students seem to be randomly selecting from the four incorrect answer choices.

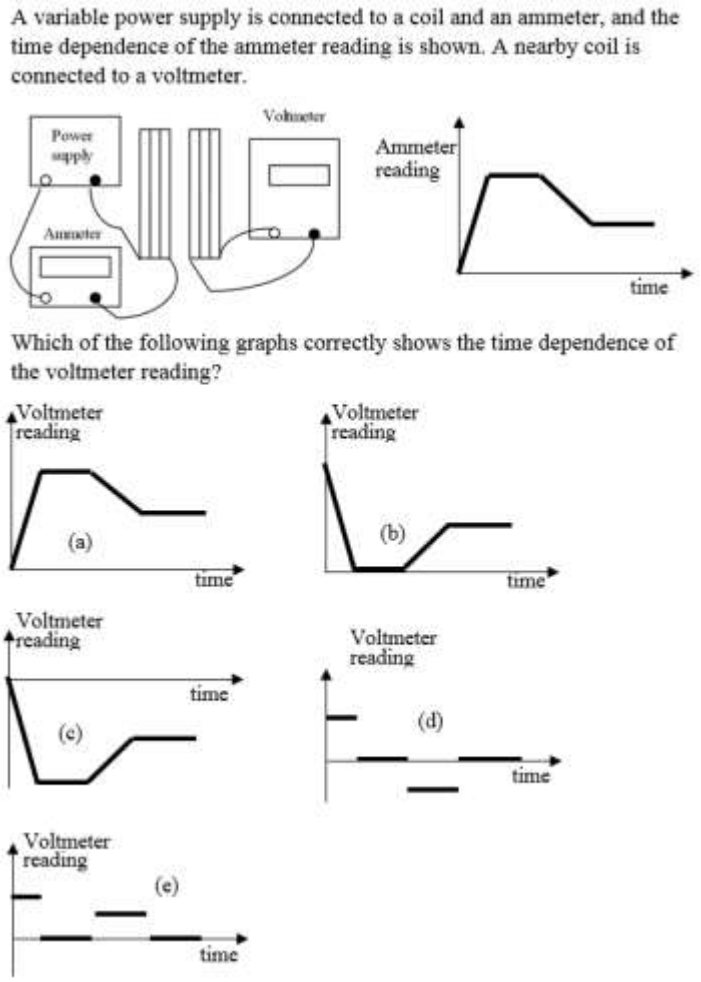

Figure 2.8. Q32 on the CSEM

Q32 is one of the most challenging questions on the CSEM (only 18% of students answered it correctly). The question and answers are shown in Fig. 2.8. On this question, the MCI is choice B (selected by 40% of them) and only 11% of TAs selected this answer choice as the most common among introductory students. Also, 31% of the TAs selected answer choice E, but only 1% of introductory students selected this choice. Therefore, TAs' average PCK on this question is quite low (41%). In the interviews, one TA selected this choice, explaining that it is possible that introductory students only think of the magnitude of the emf (once again, the TA combined a correct idea, i.e., only a changing flux induces an emf, with an incorrect one, i.e., introductory students do not recognize that the induced emf changes direction). It therefore appears that this

question is very challenging for introductory students (only 18% of them answer it correctly in a post-test), indicating that they have a lot of difficulty recognizing that the induced emf in the secondary coil is only non-zero when the current in the primary coil is changing. However, it appears that many TAs are unaware of this difficulty.

RQ1. (ii) *What alternate conceptions do TAs struggle to identify?*

We will now focus on questions in which less than 20% of the TAs identified a common student alternate conception. Q3, Q4 and Q5 are related. They are all shown in Fig. 1. On Q3, 76% of introductory students realize that the force on the $+Q$ charge should increase by a factor of 4. On Q4 however, many introductory students think that after increasing the charge on the right from $+Q$ to $+4Q$, the magnitude of the force on it remains the same, F (instead of increasing by a factor of 4 to $4F$). This alternate conception was selected by 57% of the TAs. Q5 asks students what happens to the magnitude of the force when the charges are moved to be 3 times as far apart. Many students who selected choice C on Q4 thought that the force will now decrease by a factor of 3 and selected choice B on Q5 (20%), while a smaller percentage (14%) thought that the force will decrease by a factor of 9 (correct thinking, but incorrect conclusion because the force on the +4Q charge is initially $4F$ not F). In other words, the most common alternate conception of students is that when the two charges are moved three times as far apart, the force on them decreases by a factor of 3. If the TAs are aware that this is the most common alternate conception, then among the TAs who selected answer choice C on Q4, many of them should select answer choice B on Q5. However, while 57% of the TAs selected choice C on Q4, only 16% the TAs identified choice B as the MCI on Q5, and 32% selected choice A, possibly because answer choice A is a combination of a correct idea (force decreases by a factor of 9) and an incorrect one (force on +4Q charge before increasing the distance between the two charges is F).

On Q10, 25% of introductory students thought that a charged particle at rest in an electric field will stay at rest. Only 9% of the TAs identified this alternate conception. In the interviews, they were explicitly asked whether they expected that introductory students would harbor this alternate conception. Nearly all the interviewed TAs said that it is unlikely that students do not know that charges placed in an electric field would move and thus, the interviews highlighted how challenging it is for TAs to identify this alternate conception.

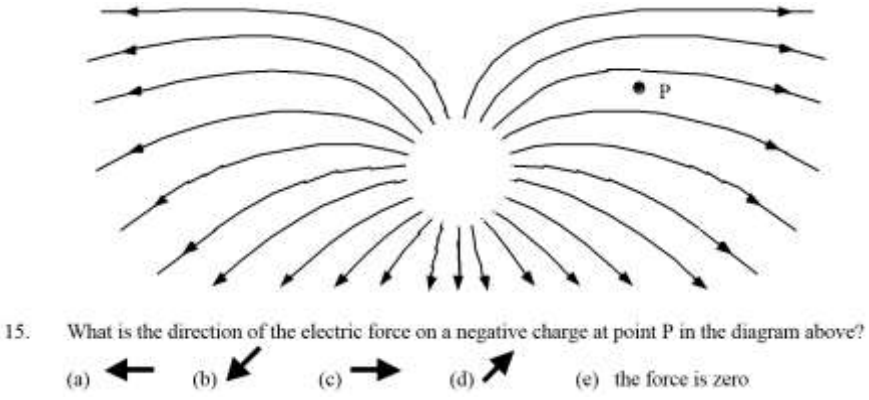

Figure 2.9. Q15 on the CSEM

Q15 is shown in Fig. 2.9. On this question, 24% of introductory students expected that the electric force points directly towards the positive charge from which all the field lines originate (answer choice B), but only 17% of the TAs identified this as the MCI. This is likely due to another alternate conception common amongst more introductory students (34%), namely that the electric force points to the right. The vast majority of the TAs identified this more common alternate conception, thus leading to a good average PCK score on this question (84%, the second largest among all CSEM questions). Q16 states that an electron is placed at a position on the *x* axis where the electric potential is equal to +10V and asks about the subsequent motion of the electron. On this question, 22% of introductory students thought that the electron would move towards the right (the most commonly selected incorrect answer), but this answer choice was the one least likely to be selected by the TAs (only 16% of them selected it). One interviewed TA thought that the

introductory students will place the electron on the positive *x* axis and a positive charge at the origin of the coordinate axis (to give concreteness to the situation) and claim that the electron would move to the left.

On Q27 (shown in Fig. 2.5), 23% of introductory students selected choice D, but only 12% of the TAs selected it. In interviews too, the TAs sometimes considered it, but usually selected either choice A or C (selected by 19% and 8% of introductory students and 46% and 24% of the TAs, respectively). On Q32, shown in Fig. 2.8, the most common alternate conception, held by 40% of introductory students, is that the reading on the voltmeter opposes the reading in the ammeter (i.e., reading on the ammeter increases, therefore reading on the voltmeter decreases and vice versa). The introductory students may be trying to apply Lenz's law, but may not realize that the induced emf opposes the *change* in flux rather than the flux itself. Only 11% of TAs identified this alternate conception.

RQ1 (iii) *What alternate conceptions can TAs identify?*

In this section, we will focus on the questions in which at least 50% of the TAs identified a common student alternate conception (i.e., incorrect answer choice selected by 20% or more introductory physics students). On Q1 (discussed earlier), 23% of introductory students selected answer choice C which states that the excess charge spreads everywhere over the inside and outside surface of the sphere. This implies that students may be thinking that the positive charges spread as far from each other as possible [23]. This alternate conception was correctly identified by 54% of the TAs, which indicates that TAs are aware that students do not know that charges on a metallic sphere are distributed only on the outer surface.

On Q4 (discussed earlier), 32% of introductory students selected answer choice C, which suggests that they might have the alternate conception that the electric force on a charge is only

proportional to the charge that is applying the force. Students may also not recognize that Newton's 3rd law applies (i.e., the electric force exerted on the $+Q$ charge by the $+4Q$ charge has the same magnitude as the electric force exerted on the $+4Q$ charge by the $+Q$ charge). This difficulty was identified by 57% of the TAs.

On Q5 (shown in Fig. 1), 22% and 20% of introductory students selected option D and option B, respectively. The introductory students who selected either of these two options are likely to think that the electric force is inversely proportional to the distance (instead of distance squared), so that when the separation between two charges is tripled, the force between them decreases by a factor of 3. So if an introductory student answers 4F/3 on Q5, he/she probably thought that the force decreased by a factor of 3, and the original force was 4F (Q4). If instead, a student answers F/3 on Q5, that student probably thought that the original force was F. Half of the TAs identified option D as the MCI, whereas only 16% selected option B. This suggests that many of the TAs expected that most introductory students would answer Q4 correctly.

On Q13 (discussed earlier) in which students were asked to find the direction of the electric field inside a hollow metal sphere due to the presence of an external positive charge, 27% of introductory students selected option A, which neglects to incorporate the effect of the metal sphere on the electric field. Roughly half of the TAs (56%) selected option A as the MCI, thus suggesting that they are aware that introductory students have difficulty recognizing how conducting objects respond to the external electric field (i.e., free charge moves in order to make the electric field inside the conductor zero).

On Q15 (shown in Fig. 2.9), 34% introductory students selected option C, thus neglecting to incorporate the sign of the charge. This difficulty was identified by 74% of TAs.

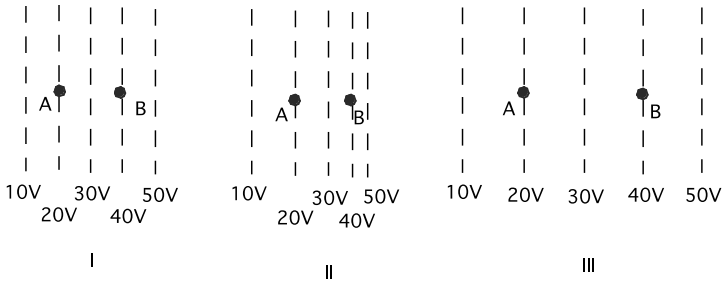

Figure 2.10. Three situations provided for Q17-Q19 on the CSEM

On Q17, students are asked to compare the work needed to move a positive charge from point A to point B in three different situations (shown in Fig. 2.10). 23% of introductory students answered that the most work is done when moving the charge in situation III. These students likely thought that the work is maximum in situation III because the distance over which the charge is moved is largest, and did not consider the potential difference between the two points. Many TAs (63%) identified this alternate conception.

Q18 also relates to the three situations shown in Fig. 2.10 and asks introductory students to compare the magnitude of the electric field at point B in all three cases. Here, 28% of introductory students selected E which states that the electric fields are equal. These students only considered that the equipotential line on which B lies is at 40V and did not recognize that it is the *change* in electric potential (i.e. gradient) that is related to the magnitude of the electric field rather than the electric potential itself. Just over half the TAs (52%) identified this difficulty.

Q19 asks students for the direction of the electric force acting on a positive charge if placed at point A or B in situation III. One quarter of the students selected answer choice B (right at point A and right at point B), possibly because "right" is the direction in which the electric potential increases and they expected that a positive charge would be pushed in that direction. This alternate conception was identified by 61% of the TAs.

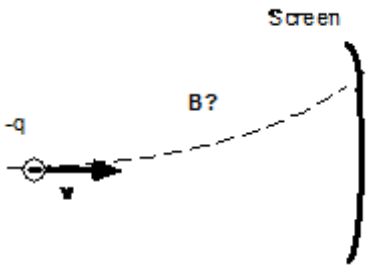

Figure 2.11. Diagram provided for Q22 on the CSEM

Q22 provides the diagram shown in Fig. 2.11 and asks for the direction of the magnetic field responsible for making the electron path curve in the way shown. 28% of introductory students selected "into the page" which would be correct if the electron was positively charged and 59% of the TAs identified this difficulty. Also, 22% of introductory students selected upward, suggesting that they may think that the direction of the magnetic force is the same as the direction of the magnetic field, significantly fewer TAs identified this alternate conception (22%).

On Q28 (shown in Fig. 2.7), 35% of introductory students selected answer choice E which states that the magnetic field at point P is zero. These students likely thought that the magnetic fields created by the two loops are in opposite directions and they therefore cancel. This alternate conception was identified by 55% of the TAs.

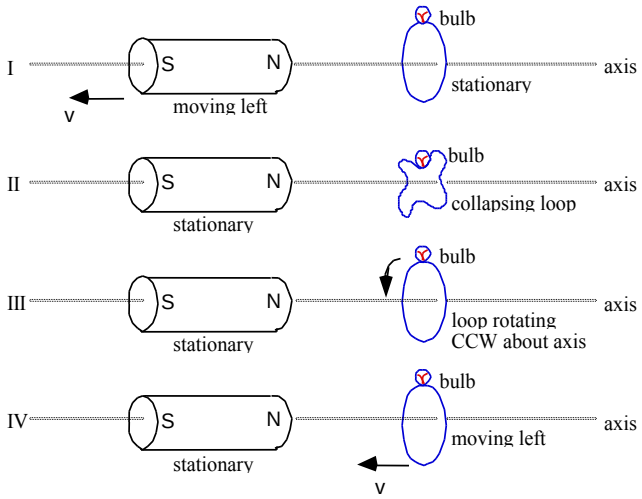

Figure 2.12. Diagram provided for Q29 on the CSEM

Q29 asks students to identify all of the situations shown in Fig. 2.12 in which the light bulb is glowing. 23% of introductory students only selected situations I and IV in which there is relative motion between the magnet and the loop. These introductory physics students did not recognize that in situation II, the electric flux is changing (because the area of the loop is changing) and therefore there will be an induced emf in the loop (light bulb glows). Roughly half the TAs identified this alternate conception. Furthermore, some introductory students (26%) also selected situation III (i.e., answered that the light bulb glows in situations I, III and IV, answer choice A), sometimes due to overgeneralizing that there is an induced emf in any situation in which the loop is moving, while much fewer TAs (26%) identified this alternate conception.

RQ2: (i) *To what extent are TAs able to predict the difficulty of the questions?*

Fig. 2.13 shows TAs' average predictions of the difficulty of each question on the CSEM, i.e., the percentage of introductory students who answered each question correctly (TAs' Predictions) as well as the actual difficulty of each question (National Data in Ref. [23]). Fig. 2.13 shows that the TAs underestimated the average difficulty of the majority of the questions on the CSEM. The distribution of correct and most common incorrect predictions of TAs for CSEM items along with the averages (connected by a red line) is shown in Figs. 2.15 – 2.18 in the chapter appendix. In the same figures, the average of the introductory students' choices (National Data) is shown (connected by a blue line) for comparison.

The discrepancy between TAs' predicted difficulty and the actual difficulty in Fig. 2.13 is quite large for some questions, in particular, the questions that were most difficult for students (e.g., 14, 20, 24, 29, 31, 32). Fig. 2.13 also shows that TAs' predicted difficulty does not fluctuate very much: with the exception of only five questions, the TAs' predicted difficulty is between 45% and 65% for all the questions on the CSEM, thus indicating that the TAs did not have a good sense

of how difficult the questions are from the perspective of introductory students. This conclusion is further supported by averaging TAs' predictions over all questions and comparing them to the actual average difficulty: TAs over-predicted introductory students' performance on the CSEM by 15% on average.

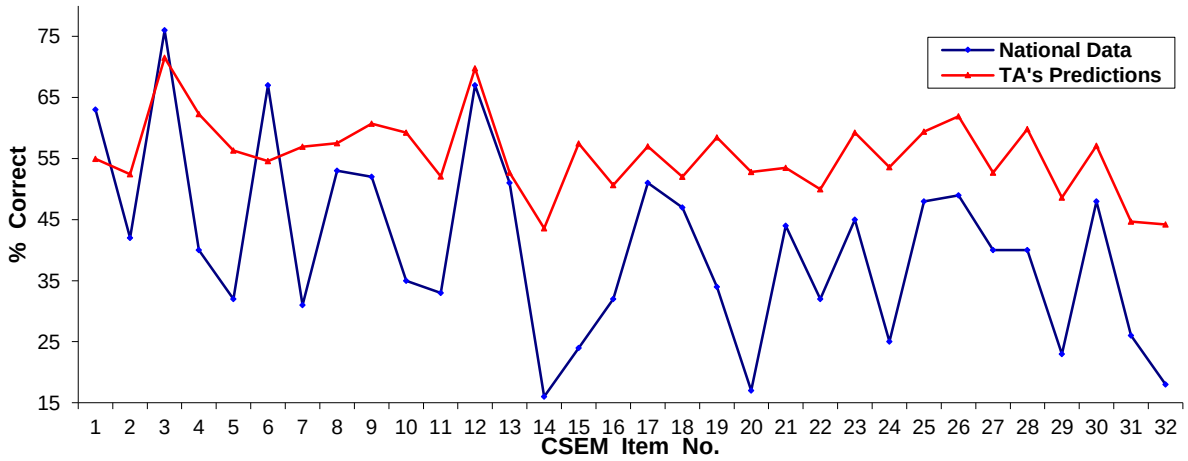

Figure 2.13. Comparison of percentages of correct answers predicted by TAs with algebra-based introductory physics students' actual performance after traditional instruction as obtained from [23]. Standard Deviations range between 17.7 and 24.6 and are not shown for clarity.

RQ2: (ii) *To what extent is TAs' ability to identify introductory students' alternate conceptions correlated with their ability to predict the difficulty of a question?*

The accuracy of TA predictions of the difficulty of a question can be quantified by taking the difference between TAs' average prediction for the percentage of students who answer the question correctly and the actual percentage of students who answer the question correctly. Any value above zero implies that the TAs are underestimating the difficulty of the question, and a value below zero indicates that they are overestimating the difficulty of the question. TAs' ability to identify the alternate conceptions is reflected in their average normalized PCK score for each question. These data are plotted in Fig. 2.14. If the TAs are more likely to identify the common student alternate conceptions when they are accurate in predicting the difficulty of a question, a

negative trend should be observed because better accuracy would correspond to a lower value for the difference between TAs' predicted difficulty and actual difficulty (note that this is only true if the TAs typically underestimate the difficulty of a question, which can easily be seen in Fig. 2.14 – there are only three questions in which the TAs overestimate the difficulty of a question). Two trends can be observed in Fig. 2.14:

i. The TAs typically underestimate the difficulty of the questions on the CSEM.

ii. There is nearly no correlation (correlation coefficient is 0.028) between TAs' ability to predict the difficulty of a question and their ability to identify the most common alternate conceptions.

The last point implies that sometimes the TAs are able to accurately predict the difficulty of a question, but not able to identify common student alternate conceptions, and vice versa, which suggests that these two abilities are different facets of pedagogical content knowledge.

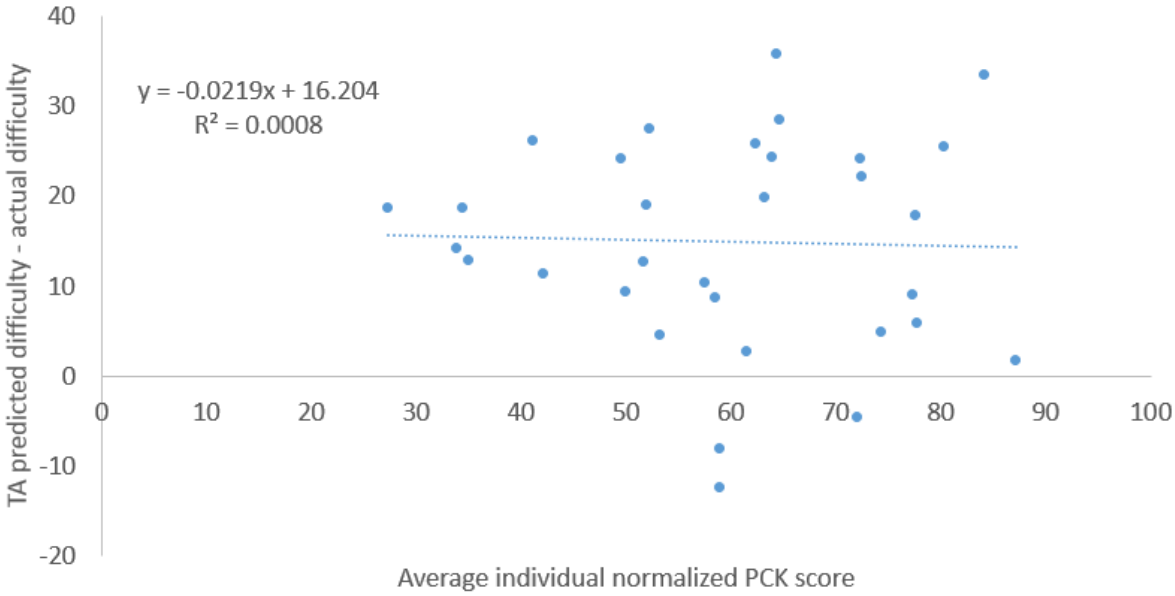

Figure 2.14. Scatter plot of TAs' ability to predict the difficulty of a question (measured as the difference between TAs' predicted difficulty and the actual difficulty) and their average normalized PCK score. The line shows almost zero correlation between these two factors.

2.4 DISCUSSION AND SUMMARY

Awareness of introductory physics students' common difficulties and being able to understand how challenging certain concepts are for introductory students are important aspects of pedagogical content knowledge. One can take advantage of introductory students' initial knowledge and design effective pedagogical approaches which account for these difficulties and help students learn better [44, 46, 49]. Our investigation used the CSEM to evaluate this aspect of pedagogical content knowledge in the context of introductory electricity and magnetism for 81 TAs who were all first-year physics graduate students enrolled in a TA training course. For each item on the CSEM, the TAs were asked to identify what they expect is the MCI of introductory physics students. Additionally, in years two and three of the study, the TAs were also asked to estimate the difficulty of each question on the CSEM. In all three years there was an in class discussion with the TAs related to the PCK task. Additionally, think-aloud interviews were conducted to obtain an in-depth account of what reasoning TAs use to arrive at the conclusion that certain alternate conceptions may be common.

General approach often used by the TAs to identify common incorrect answer choices of introductory students

When trying to decide what answer choices would be common among introductory students, TAs often selected answer choices which incorporate both correct and incorrect ideas. While this approach was sometimes productive in helping them identify the MCIs, it often led to TAs selecting answer choices that were not common at all. More importantly, after TAs identified a particular answer choice which incorporated a correct and incorrect idea, they often neglected to consider other answer choices carefully or think about what alternate conceptions could lead to students selecting them. There are many examples, for instance, Q26 (shown in Fig. 2.6) in which

TAs often selected the answer choice which has the direction opposite to the correct direction. They often stated that they were motivated to select this choice because students may try to use the right hand rule (correct idea), but do so incorrectly. However, the TAs did not always consider the other choices carefully and did not necessarily think about what alternate conceptions may lead introductory students to select those choices. We note that after they were explicitly asked to think about a particular answer choice (e.g., choice D for Q26), the TAs sometimes predicted the alternate conception, in this case that the magnetic field points radially outward from the wire (i.e., making a generalization from electric field due to a positive charge, e.g., point charge or line of charge), and said that they expected this answer choice to be more common than the one they originally selected. This suggests that a productive approach to helping TAs identify common incorrect answer choices of students is to explicitly ask them to first identify what alternate conceptions may lead introductory students to select each incorrect answer choice (for a particular question) and only after that ask them to decide which one they expected to be most common.

TAs struggled to identify alternate conceptions regarding how charge distributes on conductors/insulators

There are two questions on the CSEM which ask what happens to a charge placed at a particular point on a conducting/insulating sphere. For both questions, many TAs selected answer choices that were not common among introductory students. On the question in which the sphere is insulating, nearly half the TAs expected that students would think that most of the charge remains where it was placed, but some does spread over the sphere. Interviews suggested that the TAs selected this answer choice because it is the choice which is most similar to the correct answer (charge remains where it was placed), i.e., the TAs used the same strategy we described above in other contexts.

TAs struggled to identify alternate conceptions regarding the magnetic field caused by a current

On both questions related to magnetic field caused by a current for which there was a common alternate conception, the TAs selected answer choices which are not at all common among introductory students. On both questions TAs' often selected answer choices in which the right hand rule was used incorrectly, but very few introductory students selected those answer choices.

TAs struggled to identify alternate conceptions regarding the motion of/force on a charged particle in a magnetic field.

Out of the four questions dealing with the concept of Lorenz force (Q21, Q22, Q25, Q27), only on one of them (Q22) did the majority of TAs identify the MCIs. On the other ones, the TAs often selected answer choices that were not common. Also, Q25 was one of the most challenging questions for the TAs; in interviews they often spent a considerable amount of time trying to figure out how introductory students may answer the question and sometimes even ended up essentially guessing, or committing to an answer only after being asked to select one.

Alternate conceptions held by very few students which TAs expected would be the most common

There were multiple instances in which TAs selected certain incorrect answer choices which they thought would be most common among introductory students, but those answer choices were very rarely selected by introductory students. Three such examples are presented in the preceding paragraphs and there are many others. We will mention two more: on Q21, 31% of the TAs expected that introductory students would confuse the magnetic field with an electric field and think that the charge will move at constant acceleration, but only 8% of introductory students selected this answer choice and on Q32, 31% of the TAs selected answer choice E, but only 1% of the introductory students selected this incorrect option.

Alternate conceptions that the TAs were able to identify

The TAs performed reasonably well at identifying alternate conceptions related to Coulomb's force law (Q3-8), although, there is room for improvement, especially on Q8. On Q3 and Q6, there are no strong alternate conceptions, and on Q4 and Q5, TAs' average PCK score was high. On Q7, the majority of the TAs identified the alternate conceptions, and their performance was close to being considered good according to our heuristics. Q8 is the only one on which the TAs could improve significantly, and this is the only question of the group that has a complicated setup and asks students to compare two configurations side by side, one with three charges and the other with four. It is possible that TAs' lower PCK performance on this question was due to the setup being more complicated than those used in the other questions.

TAs performed reasonably well in identifying the alternate conception that the electric field inside a hollow metallic sphere due to an external charge is the same as it would be without the hollow metal sphere. In other words, TAs were aware that introductory students have difficulty understanding that the inside of a metallic sphere is shielded from outside electric fields. The TAs' normalized PCK score (87%) was the highest on this question (Q 13) among all the questions on the CSEM.

On two other questions involving Faraday's Law/Lenz's law (Q29 and Q30), TAs performed well in identifying introductory students' alternate conception that an emf is induced in a loop whenever there is ***any*** type of relative motion between a magnet/current carrying wire and a loop of wire.

TAs' ability to predict the difficulty of the questions on the CSEM

Our results also suggest that the TAs typically underestimated the difficulty of the questions on the CSEM, especially on the challenging questions. For all but five questions on the

CSEM, TAs' average predictions for the percentage of introductory students who answer the questions correctly were between 45% and 65%, while the actual percentages varied much more widely. This strongly suggests that the TAs struggled to think about the difficulty of the questions from a student's perspective. Furthermore, we found that TAs' ability to predict the difficulty of a question was uncorrelated with their ability to identify the MCIs, thus suggesting that the two are separate facets of pedagogical content knowledge.

Using a PCK task as a pedagogical tool

Many TAs explicitly noted that the CSEM-related PCK task was challenging and it was difficult for them to think about physics questions from an introductory physics student's perspective. In the think-aloud interviews, graduate students sometimes made comments which indicated that they found the task challenging (e.g., explicitly commenting "I don't know introductory students well enough…"). However, many TAs noted that the CSEM-related PCK task was worthwhile and helped them think about the importance of putting themselves in their students' shoes in order for teaching and learning to be effective, especially after receiving introductory student data on how students actually performed and discussing particular student alternate conceptions.

If such a task is used in professional development (for example for teaching assistants), our interviews suggest that teaching assistants should be explicitly told to first try to identify (and perhaps write down) what alternate conceptions or incorrect reasoning may lead introductory students to select each of the incorrect answer choices *before* deciding which one is most common. In interviews, we found that TAs often identified one possible alternate conception, selected it, and moved on to the next question without carefully considering other answer choices. When prompted to consider another answer choice, they were sometimes able to identify the most

common (or a more common) alternate conception and even asked if they could change their original answer.

We note that the authors have been using tasks similar to the one described here in the professional development of TAs at their institutions and have found them to be very useful in setting the stage for a discussion on the importance of being aware of introductory students' difficulties and alternate conceptions in order to design instruction to help students learn. The TAs discuss questions which have been carefully selected to engender productive discussions among TAs, e.g., a question on which interviews and quantitative data suggest that the TAs will consider multiple answer choices but the introductory students are likely to select only one of them; or a question which has multiple common incorrect answer choices, etc. The TAs are explicitly asked to identify and discuss with each other what reasoning introductory students may use to select each incorrect answer choice before making a decision about which one is most common. Additionally, they are asked to predict the difficulty of each question. After the TAs complete the task, they are shown data from students, and some TAs explicitly express that it is very valuable for them to learn about the common student difficulties in concrete contexts. We found that TAs tend to trust student data more than statements like "research has found that…" The discussion is then focused on how TAs can identify common student difficulties related to various physics concepts, e.g., by listening to students when reasoning about physics and coming up with guiding questions in real time to develop a grasp of how students are thinking in specific contexts. At one of the institutions (A.M.), the rest of the professional development program (which meets once a week for a semester) is focused on the tutorials students work on and their common difficulties on specific questions on the tutorials, as well as effective approaches the TAs can use to help students develop a coherent knowledge structure of those introductory physics concepts. Using such tasks with actual data from

introductory students in TA professional development courses can be effective at other institutions as well.

Comparison to prior studies related to TAs' PCK for multiple choice assessments

We note that in this study, we also found that discussions between TAs did not significantly improve their PCK performance on the CSEM-related PCK task; overall, working in groups only improved their average PCK performance on the whole test by 2%, an improvement that was not statistically significant.

Our earlier studies of PCK [44, 46, 50] on both the FCI and TUG-K found that the group PCK performance was significantly better than individual PCK performance. Also, the benefits of group or collaborative work have been a consistent finding of PER in particular and educational research in general [51]. The fact that we found a different result here with the CSEM survey, especially when contrasting it with our earlier results using a PCK task with other assessments (FCI and TUG-K), suggests that the PCK task may be more challenging when the assessment used is the CSEM compared to other assessments related to force or kinematics. One potential reason for this is the crucial difference between the topics of mechanics (including kinematics) and electricity and magnetism: our daily experience with the real world leads to a relatively predictable (Aristotelian) world view and TAs could more easily reason their way to common misconceptions held by students. Electricity and magnetism, on the other hand, deals with concepts that are not primarily learned experientially (e.g., charges, fields and currents), which likely makes it more difficult to predict the most common difficulties of students. We note however, that whether the context is electricity and magnetism, force and motion, kinematics, or quantum mechanics, whether intuitive or not, student difficulties can be classified in a few categories [52, 53]. Knowing

the types of incorrect reasoning students engage in for a particular context can help in designing instruction to help students develop a robust knowledge structure [53].

Despite the difference mentioned above, there are many commonalities in the three PCK studies. In all of these three studies, there are questions for which the TAs' performance at identifying common student difficulties is good, while there are also questions in which TAs struggled to identify student difficulties. Both interviews and the quantitative data show that it was often the case that TAs selected answer choices that are not very common among introductory students. In interviews, they sometimes considered different answer choices and struggled to select the most common one, sometimes only doing so after being reminded that they should try to identify the answer choice that is most common.

Our earlier studies using the TUG-K and FCI showed that the ability to identify common introductory students' alternate conceptions was not dependent on familiarity with US teaching practices and that TAs exhibited comparable performance in identifying introductory students' alternate conceptions for the FCI or TUG-K regardless of whether they obtained their undergraduate degree in the US or elsewhere. Therefore, we did not explicitly compare the PCK performance of TAs with different institutional backgrounds in detail for the present study. However, informal observations during the TA training course as well as interviews suggest that the CSEM related PCK performance of TAs of different backgrounds, e.g., Chinese vs. American, appears to be comparable.

2.5 CHAPTER REFERENCES

1. C. Singh, Categorization of problems to assess and improve proficiency as teacher and learner, Am. J. Phys. **77**, 73 (2009).

2. D. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores, Am. J. Phys. **70**, 1259 (2002).

   A. J. Mason and C. Singh, Assessing Expertise in Introductory Physics Using Categorization Task, Phys. Rev. ST PER **7**, 020110 (2011).

3. C. Singh, Rethinking tools for training teaching assistants, Proceedings of the 2009 Phys. Ed. Res. Conference, Ann Arbor, MI, (M. Sabella, C. Henderson, C. Singh Eds.), AIP Conf. Proc., Melville, New York **1179**, p. 59 (2009). http://dx.doi.org/10.1063/1.3266754

4. S. Y. Lin, C. Henderson, W. Mamudi, E. Yerushalmi, and C. Singh, Teaching assistants' beliefs regarding example solutions in introductory physics, Phys. Rev. ST PER **9**, 010120 (2013).

5. E. Yerushalmi, C. Henderson. W. Mamudi, C. Singh, and S. Y. Lin, The group administered interactive questionnaire: An alternative to individual interviews, Proceedings of the 2011 Phys. Educ. Res. Conference, Omaha, NE, (S. Rebello, C. Singh, P. Engelhardt Eds.), AIP Conf. Proc., Melville, New York **1413**, p. 97 (2012). http://dx.doi.org/10.1063/1.3680003

6. S. Y. Lin, C. Singh, W. Mamudi, C. Henderson, and E. Yerushalmi, TA-designed vs. research-oriented problem solutions, Proceedings of the 2011 Phys. Educ. Res. Conference, Omaha, NE, (S. Rebello, C. Singh, P. Engelhardt Eds.), AIP Conf. Proc., Melville, New York **1413**, p. 255 (2012). http://dx.doi.org/10.1063/1.3680043

7. E. Yerushalmi, E. Marshman, A. Maries, C. Henderson, and C. Singh, Grading practices and considerations of graduate students at the beginning of their teaching assignment, Proceedings of the 2014 Phys. Educ. Res. Conference, Minneapolis, MN, (P. Engelhardt, A. Churukian, D. Jones Eds.) p. 287 (2015). http://dx.doi.org/10.1119/perc.2014.pr.068

8. C. Henderson, E. Marshman, A. Maries, E. Yerushalmi, and C. Singh, Instructional goals and grading practices of graduate students after one semester of teaching experience, Proceedings of the 2014 Phys. Ed. Res. Conference, Minneapolis, MN, (P. Engelhardt, A. Churukian, D. Jones Eds.) p. 111 (2015). http://dx.doi.org/10.1119/perc.2014.pr.024

9. F. Lawrenz, P. Heller, and R. Keith, Training the teaching assistant: Matching TA strengths and capabilities to meet specific program goals, J. Col. Sci. Teach. **22**, 106 (1992).

10. P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. 1. Group vs individual problem solving, Am. J. Phys. **60**, 627 (1992).

11. P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. 2. Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).

12. C. Singh, Impact of peer interaction on conceptual test performance, Am. J. Phys. **73**(5), 446 (2005).

13. C. Singh, Effectiveness of group interaction on conceptual standardized test performance, Proceedings of the 2002 Phys. Ed. Res. Conference, Boise (Eds. S. Franklin, K. Cummings and J. Marx), p. 67 (2002). http://dx.doi.org/10.1119/perc.2002.pr.017

14. A. J. Mason and C. Singh, Helping students learn effective problem solving strategies by reflecting with peers, Am. J. Phys. **78**, 748 (2010).

15. A. Mason and C. Singh, Impact of guided reflection with peers on the development of effective problem solving strategies and physics learning, Phys. Teach. **54**, 295 (2016).

16. A. J. Mason and C. Singh, Using reflection with peers to help students learn effective problem solving strategies, Proceedings of the 2010 Physics Education Research Conference, Portland, OR, (C. Singh, M. Sabella, S. Rebello Eds.), AIP Conf. Proc., Melville New York **1289**, 41-44, (2010). http://dx.doi.org/10.1063/1.3515243

17. C. Sandifer and E. Brewe (Eds.), Recruiting and Educating Future Physics Teachers: Case Studies and Effective Practices, American Physical Society, PhysTEC, 2015.

18. C. Henderson, E. Marshman, R. Sayer, C. Singh, and E. Yerushalmi, Graduate teaching assistants use different criteria when grading introductory physics vs. quantum mechanics problems, Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA p. 140 (2016). http://dx.doi.org/10.1119/perc.2016.pr.030

19. E. Yerushalmi, R. Sayer, E. Marshman, C. Henderson and C. Singh, Physics graduate teaching assistants' beliefs about a grading rubric: Lessons learned, Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA p. 408 (2016). http://dx.doi.org/10.1119/perc.2016.pr.097

20. P. Sadler et al., The influence of teachers' knowledge on student learning in middle school physical science classrooms, Am. Educ. Res. J. **50**, 1020 (2013).

21. L. S. Shulman, Those who understand: Knowledge growth in teaching, Educ. Res. **15**, 4 (1986).

22. L. S. Shulman, Knowledge and teaching: Foundations of the new reform, Harv. Educ. Rev. **57**, 1 (1987).

23. D. Maloney, T. O'Kuma, C. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. supplement, **69**, s12 (2001).

    L. Ding, R. Chabay, B. Sherwood and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST PER **2**, 010105 (2006).

C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys. **71**, 607 (2003).

24. H. Ginsberg and S. Opper, *Piaget's Theory of Intellectual Development* (Prentice Hall, Englewood Cliffs, 1969).

25. G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accomodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 211-227 (1982).

26. A. S. Elstein, L. S. Shulman, and S. Sprafka, *Medical Problem Solving: The Analysis of Clinical Reasoning* (Harvard University Press, Cambridge, 1978).

27. J. H. van Driel, N. Verloop, and W. de Vos, Developing science teachers' pedagogical content knowledge, J. Res. Sci. Teach. **35**, 673 (1998).

28. P. L. Grossman, *The Making of a Teacher: Teacher Knowledge and Teacher Education* (Teachers College Press, New York 1990).

29. P. L. Grossman, What are we talking about anyhow: Subject matter knowledge for secondary English teachers, in *Advances in Research on Teaching, Vol. 2: Subject Matter Knowledge*, edited by J. Brophy (JAI Press, Greenwich, CT, 1991), pp. 245–264.

30. J. Gess-Newsome and N. G. Lederman, *Examining Pedagogical Content Knowledge*, (Kluwer Academic Publishers, Boston, 2001).

31. J. Loughran, P. Mulhall, and A. Berry, In search of Pedagogical Content Knowledge in science: Developing ways of articulating and documenting professional practice, J. Res. Sci. Teach. **41**, 370 (2004).

32. H. Borko and R. T. Putnam, Expanding a teacher's knowledge base: A cognitive psychological perspective on professional development, in T. R. Guskey and M. Huberman, *Professional Development in Education: New Paradigms and Practices* (New York: Teachers College Press 1995).

33. C. L. Ebert, An assessment of prospective secondary teachers' pedagogical content knowledge about functions and graphs, paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA, USA (1993).

34. A. N. Geddis, B. Onslow, C. Beynon, and L. Oesch, Transforming content knowledge: Learning to teach about isotopes, Science Education **77**, 575 (1993).

35. J. H. Van Driel and O. De Jong, Investigating the development of preservice teachers' pedagogical content knowledge, Paper presented at the annual meeting of the National Association for Research in Science Teaching, St. Louis, MO, USA (2001).

36. G. Zavala, H. Alarcón, and J. Benegas, Innovative training of in-service teachers for active learning: A short teacher development course based on Physics Education Research, J. Sci. Teach. Prep. **18**, 559 (2007);

N. G. Lederman, J. Gess-Newsome, and M. S. Latz, The nature and development of preservice science teachers' conceptions of subject matter and pedagogy, J. Res. Sci. Teach **31**, 129 (1994).

37. D. Zollman, Preparing future science teachers: the physics component of a new programme, Phys. Educ. **29**, 271 (1994); H. Borko and R. Putnam, Learning to teach, in D. Berliner and R. Calfee (Eds.), *Handbook of Educational Psychology* (673-708), (Macmillan, New York, 1996).

38. H. Akkoç and S. Yeşildere, Investigating development of pre-service elementary mathematics teachers' pedagogical content knowledge through a school practicum course, Procedia – Social and Behavioral Sciences **2**, 1410 (2010).

39. K. Carter, The place of story in the study of teaching and teacher education, Educational Researcher **22**, 5 (1993).

40. D. M. Kagan, Ways of evaluating teacher cognition: Inferences concerning the Goldilocks Principle, Review of Educational Research **60**, 419(1990).

41. J. J. Loughran, R. F. Gunstone, A. Berry, P. Milroy, and P. Mulhall, Science cases in action: Developing an understanding of science teachers' pedagogical content knowledge, paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans, LA, USA (2000).

42. J. A. Baxter and N. G. Lederman, Assessment and measurement of Pedagogical Content Knowledge, in J. Gess-Newsome and N. G. Lederman (Eds.), *Examining Pedagogical Content Knowledge: The Construct and its Implications for Science Education* (Kluwer Academic Publishers 1999).

43. D. Hestenes, M. Wells and G. Swackhamer, Force Conceptual Inventory, Physics Teacher 30 141 (1992).

44. A. Maries and C. Singh, Teaching assistants' performance at identifying common introductory student difficulties in mechanics revealed by the Force Concept Inventory, Phys. Rev. PER **12**, 010131 (2016).

45. R. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

46. A. Maries and C. Singh, Exploring one aspect of pedagogical content knowledge of teaching assistants using the test of understanding graphs in kinematics, Phys. Rev. ST PER **9**, 020120 (2013).

47. L. C. McDermott and P. S. Schaffer, *Tutorials in Introductory Physics*, (Upper Saddle River, Prentice Hall, NJ 1998).

48. K. A. Ericsson, and H. Simon, Verbal reports as data, Psychological Review 87, 215 (1980).

49. J. R. Thompson, W. M. Christensen, and M. C. Wittmann, Preparing future teachers to anticipate student difficulties in physics in a graduate-level course in physics, pedagogy, and education research, Phys. Rev. ST PER **7**, 010108 (2011).

50. A. Maries and C. Singh, Performance of graduate students at identifying introductory students' difficulties with kinematics graphs, Proceedings of the 2014 Phys. Ed. Res. Conference, Minneapolis, MN (A. Churukian, P. Engelhardt, D. Jones Eds.), p. 171 (2015). http://dx.doi.org/10.1119/perc.2014.pr.039

51. D. W. Johnson and R. T. Johnson, An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning, Educational Researcher, 38(5), 365 (2009).

52. C. Singh and E. Marshman, Review of student difficulties in upper-level quantum mechanics, Phys. Rev. ST PER **11**, 020117 (2015).

53. E. Marshman and C. Singh, Framework for understanding the patterns of student difficulties in quantum mechanics, Phys. Rev. ST PER **11**, 020119 (2015).

2.6 CHAPTER APPENDIX

2.6.1 Mathematical Description of CSEM-related PCK Score Calculation

We define indices *i*, *j* and *k* that correspond to the following:

- *i*: index of TAs (81 TAs; it takes values from 1 to 81);
- *j*: CSEM question number (32 questions; it takes values from 1 to 32);
- *k*: incorrect answer choice number for each question (4 incorrect answer choices; it takes values from 1 to 4).

Then, we let F_{jk} be the fraction of introductory physics students who selected incorrect answer choice *k* on item *j* (e.g. F_{21} = 0.04, F_{22} = 0.23, F_{23} = 0.07, F_{24} = 0.03). We let TA_{ijk} correspond to whether TA_i chose incorrect answer choice *k* on item *j* (for a given *i* and *j*, TA_{ijk}=1 only for the incorrect answer choice *k*, selected by TA_i on item *j*, otherwise TA_{ijk}=0). Then, the

PCK score of the *i*-th TA on item *j* (referred to TA_{ij}) is: $TA_{ij} = \sum_{k=1}^{4}(TA_{ijk} \cdot F_{jk})$. Then, the total PCK score of the *i*-th TA (TA_i) on the whole survey can be obtained by summing over all of the questions:

$$TA_i = \sum_{j=1}^{32} TA_{ij} = \sum_{j=1}^{32}\left[\sum_{k=1}^{4}(TA_{ijk} \cdot F_{jk})\right].$$

Also, the average PCK score of all of the TAs on item *j* (referred to as $\overline{TA_j}$) can be obtained by taking an average over the TA scores on that particular question:

$$\overline{TA_j} = \sum_{i=1}^{81} TA_{ij} = \frac{1}{81}\sum_{i=1}^{81}\left[\sum_{k=1}^{4}(TA_{ijk} \cdot F_{jk})\right].$$

These can be converted to percentages by multiplying by 100. A similar approach can also be adopted for the groups (G_{ij} = PCK score of the *i*th group on item *j*; G_i = PCK score of the *i*-th group on the whole survey; $\overline{G_j}$ = average PCK score of all groups on item *j*) and for random guessers (RG_{ij} = PCK score of *i*th random guesser on item *j*; RG_i = PCK score of *i*th random guesser; $\overline{RG_j}$ = average PCK score of random guessers on item *j*). The PCK scores of each TA/group/random guesser (GS_i, G_i, RG_i as described above) were used to obtain averages and standard deviations in order to perform *t*-tests to compare the CSEM-related PCK performance of TAs with that of the groups and random guessers on the whole survey. In order to compare the PCK performance of these different groups on individual items, the averages and standard deviations of the PCK scores on that particular question (e.g., for question *j* on the CSEM: TA_{ij}, G_{ij}, RG_{ij}) were used to perform *t*-tests.

2.6.2 Comparison of TA Performance with Random Guessing

Random guessing on this task would correspond to choosing one of the four incorrect answer choices for each question with equal probability (25%). Therefore, one quarter of the random

guessers always selected the first incorrect answer choice, one quarter selected the second incorrect answer choice, etc. Therefore, each individual random guesser obtains a score for each question and these scores were used to perform a comparison with the TA scores via t-test. We note that, with 80 ‘random guessers’, the TAs predictions are better than random guessing for 20 questions as described in Table 2.1 whereas with 40 ‘random guessers’, the TAs predictions are better than random guessing for 19 questions (Q3 was worse in this case).

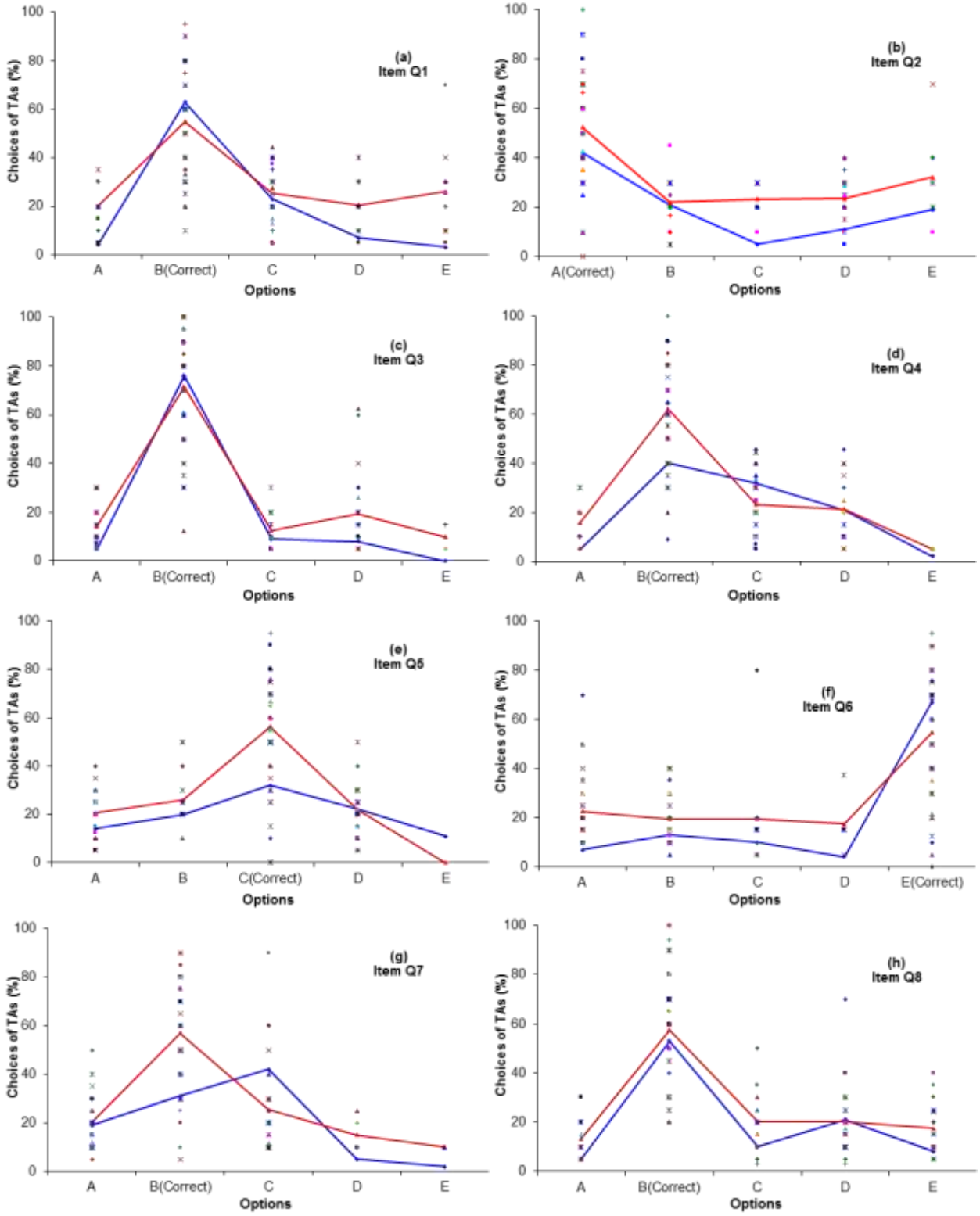

Figure 2.15. The distribution of correct and most common incorrect predictions of TAs for CSEM items Q1–Q8 (a–h) along with the averages (connected by a red line). The average of the introductory students' choices (National Data) is shown (connected by a blue line) for comparison.

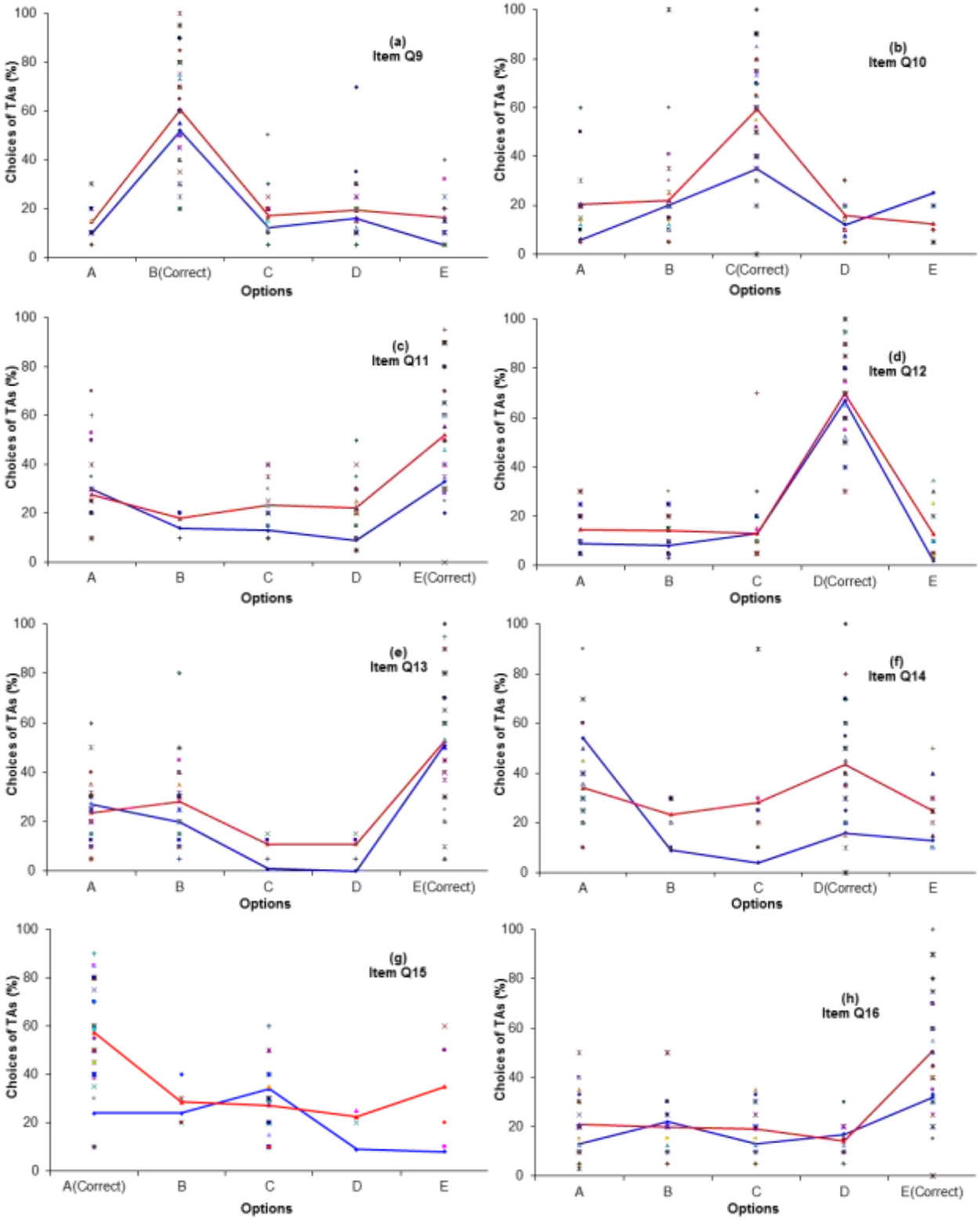

Figure 2.16. The distribution of correct and most common incorrect predictions of TAs for CSEM items Q9–Q16 (a–h) along with the averages (connected by a red line). The average of the introductory students' choices (National Data) is shown (connected by a blue line) for comparison.

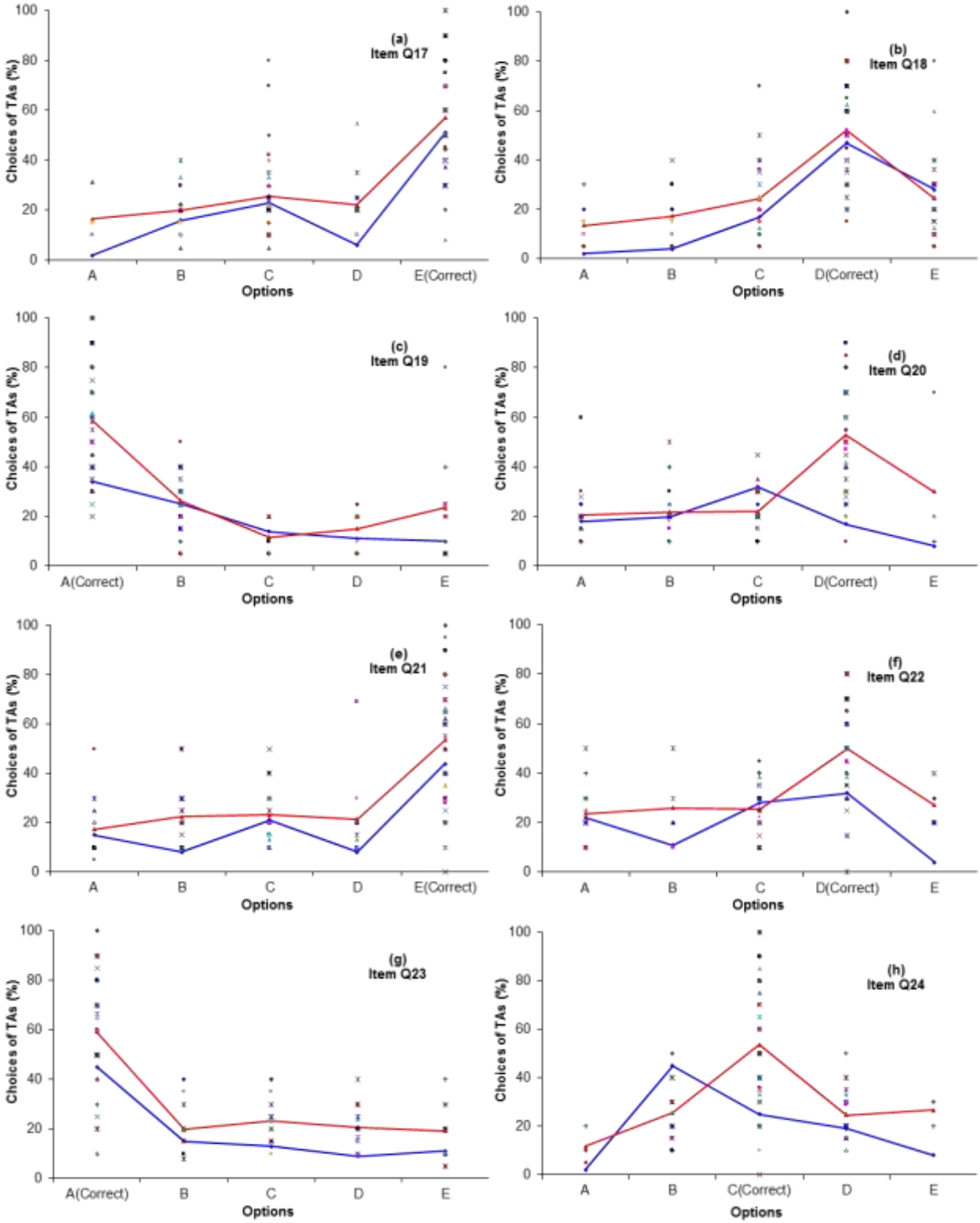

Figure 2.17. The distribution of correct and most common incorrect predictions of TAs for CSEM items Q17–Q24 (a–h) along with the averages (connected by a red line). The average of the introductory students' choices (National Data) is shown (connected by a blue line) for comparison.

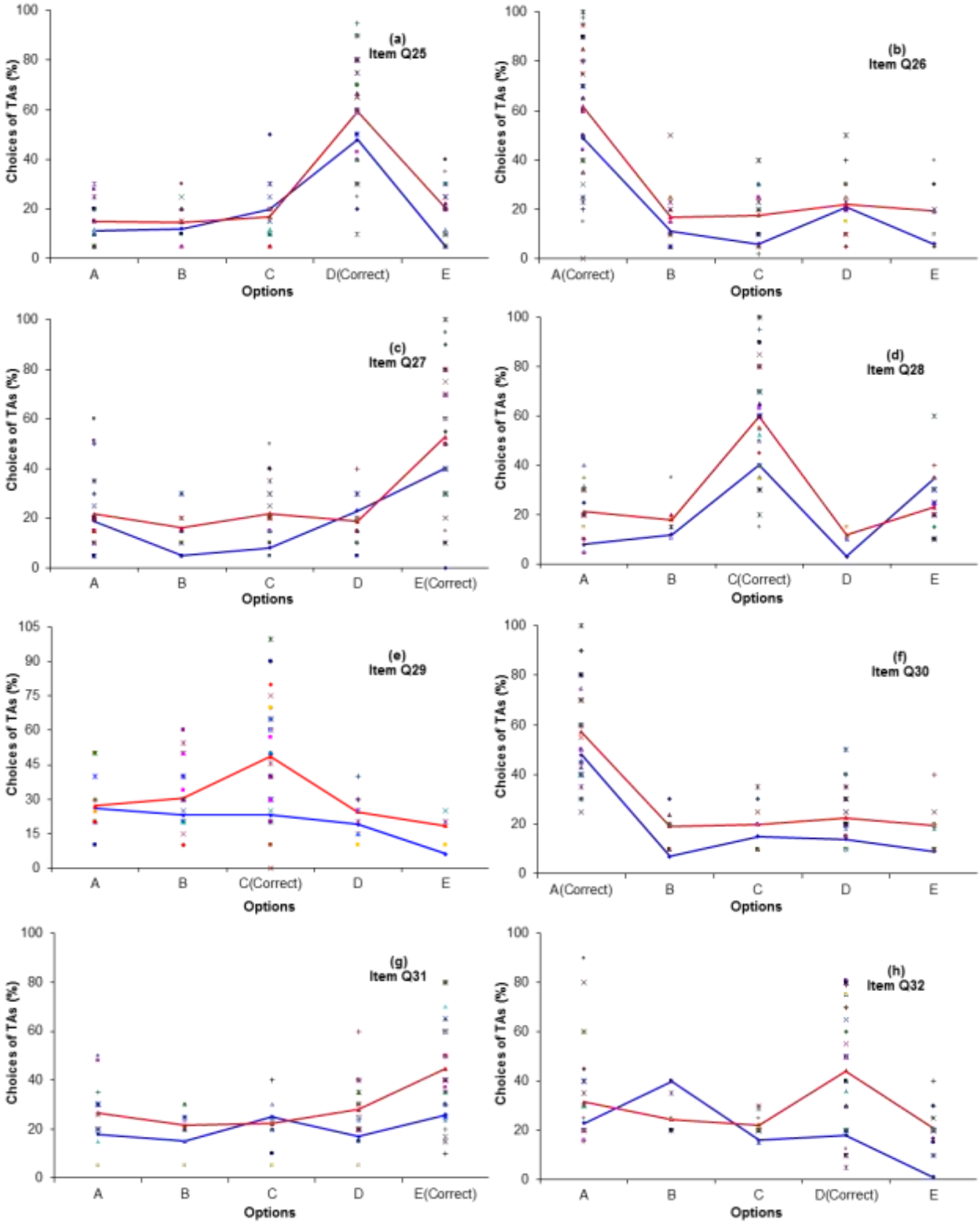

Figure 2.18. The distribution of correct and most common incorrect predictions of TAs for CSEM items Q25–Q32 (a–h) along with the averages (connected by a red line). The average of the introductory students' choices (National Data) is shown (connected by a blue line) for comparison.

3.0 IMPACT OF EVIDENCE-BASED ACTIVE-ENGAGEMENT COURSES ON STUDENT PERFORMANCE IN INTRODUCTORY PHYSICS

3.1 INTRODUCTION

3.1.1 Physics Education Research-based Active Engagement Methods

In the past two decades, physics education research has identified the challenges that students encounter in learning physics at all levels of instruction [1-15]. Building on these investigations, researchers are developing, implementing and evaluating evidence-based curricula and pedagogies to reduce these challenges to help students develop a coherent understanding of physics concepts and enhance their problem solving, reasoning and meta-cognitive skills [16-27]. In evidence-based curricula and pedagogies, the learning goals and objectives, instructional design, and assessment of learning are aligned with each other and there is focus on evaluating whether the pedagogical approaches employed have been successful in meeting the goals and enhancing student learning.

One highly successful model of learning is the field-tested cognitive apprenticeship model [28]. According to this model, students can learn effectively if the instructional design involves three essential components: "modeling", "coaching and scaffolding", and "weaning". In this approach, "modeling" means that the instructional approaches demonstrate and exemplify the criteria for good performance and the skills that students should learn (e.g., how to solve physics problems systematically). "Coaching and scaffolding" means that students receive appropriate guidance and support as they actively engage in learning the content and skills necessary for good performance. "Weaning" means reducing the support and feedback gradually to help students

develop self-reliance [28]. In traditional physics instruction, especially at the college level, there is often a lack of coaching and scaffolding: students come to class where the instructor lectures and does some example problems, then students are left on their own to work through homework with little or no feedback. This is akin to a piano instructor demonstrating for the students how to play the piano and then asking students to go home and practice. This lack of prompt feedback and scaffolding can be detrimental to learning.

Some of the commonly used evidence-based active-engagement (EBAE) approaches implemented in physics include peer instruction with clickers popularized by Eric Mazur from Harvard University [29-32], tutorial-based instruction in introductory and advanced courses [33-35] and collaborative group problem solving [36-39], e.g., using context-rich problems [11-12]. In all of these evidence-based approaches, formative assessment plays a critical role in student learning [40]. Formative assessment tasks are frequent, low-stakes assessment activities which give feedback both to students as well as instructors about what students have learned at a given point. Using frequent formative assessments helps make the learning goals of the course concrete to students, as well as provides them with a way to track their progress in the course with respect to these learning goals. When formative assessment tasks such as concept-tests, tutorials and collaborative group problem solving are interspersed throughout the course, the distinction between teaching and learning is blurred [40-41].

Moreover, technology is increasingly being exploited for pedagogical purposes to improve student learning. For example, Just-in-Time Teaching (JiTT) is an instructional approach in which instructors receive feedback from students before class and use that feedback to tailor in-class instruction [42-44]. Typically, students complete an electronic pre-lecture assignment in which they give feedback to the instructor regarding any difficulties they have had with the assigned

reading material, lecture videos, and/or other self-paced instructional tools. The instructor then reviews student feedback before class and makes adjustments to the in-class activities. For example, Eric Mazur's Perusall system [45] allows students to read the textbook and ask questions electronically and the system uses their questions to draft a "confusion report" which distills their questions to three most common difficulties. Then, during class, students may engage in discussions with the instructor and with their classmates, and the instructor may then adjust the next pre-lecture assignment based on the progress made during class. It has been hypothesized that JiTT may help students learn better because out-of-class activities cause students to engage with and reflect on the parts of the instructional material they find challenging. In particular, when the instructor focuses on student difficulties in lecture which were found via electronic feedback before class, it may create a "time for telling" [46] especially because students may be "primed to learn" better when they come to class if they have struggled with the material during pre-lecture activities. The JiTT approach is often used with peer discussion and/or collaborative group problem solving inter-dispersed with lectures in the classroom.

In addition, in the last decade, the JiTT pedagogy has been extended a step further with the maturing of technology [47-66] and "flipped" classes with no in-class lectures have become common with instructors asking students to engage with short lecture videos and concept questions associated with each video outside of the class and using the entire class-time for active-engagement. The effectiveness of flipped classes in enhancing student learning can depend on many factors including the degree to which evidence-based pedagogies that build on students' prior knowledge and actively engage them in the learning process are used, whether there is sufficient buy-in from students and the incentives that are used to get students engaged with the learning tools both inside and, equally importantly, outside the classroom.

Moreover, research suggests that effective use of peer collaboration can enhance student learning in many instructional settings in physics classes including in JiTT and flipped environments, and with various types and levels of student populations. Although the details of implementation vary, students can learn from each other in many different environments. Integration of peer interaction with lectures has been popularized in the physics community by Mazur. In Mazur's approach [67], the instructor poses concrete conceptual problems in the form of conceptual multiple-choice clicker questions to students throughout the lecture and students discuss their responses with their peers. Heller et al. have shown that collaborative problem solving with peers in the context of quantitative "context-rich" problems [11-12] can be valuable both for learning physics and for developing effective problem solving strategies.

Cognitive apprenticeship [28] is one framework that can be used to understand why the EBAE instructional strategies that take advantage of peer discussion and collaboration may be successful in helping students learn. The EBAE pedagogies provide instructors with an opportunity to receive feedback on common student difficulties. The instructors often use this feedback to adjust their in-class activities to effectively build on students' prior knowledge, thus providing students with the necessary coaching and scaffolding to help them learn. Peer discussion also provides students with an opportunity to be coached by their peers who may be able to discern their difficulties even better than the instructor, and carefully designed targeted feedback from the instructor after the peer discussion can provide appropriate scaffolding.

3.1.2 Focus of the Research: Comparing Introductory Physics Student Performance in EBAE (Flipped and Non-flipped) Courses with LB Courses

In this study, we used the Force Concept Inventory (FCI) [71] in the first semester courses and the Conceptual Survey of Electricity and Magnetism (CSEM) [72] in the second semester courses to assess student learning. The FCI, CSEM and other standardized physics surveys [71-78] have been used to assess introductory student understanding of physics concepts by a variety of educators and physics education researchers. One reason for their extensive use is that many of the items on the survey have strong distractor choices which correspond to students' common difficulties so students are unlikely to answer the survey questions correctly without having good conceptual understanding. In the research discussed here, the performance of students in EBAE courses at a particular level is compared with primarily LB courses in two situations: (I) the same instructor taught two courses, one of which was a flipped course involving EBAE methods and the other an LB course, while the homework and final exams were kept the same, (II) student performance in all of the EBAE courses taught by different instructors were averaged and compared with primarily LB courses of the same type also averaged over different instructors. Whenever differences between these two groups were observed (with students in EBAE courses performing better than students in the LB courses), we investigated which students were benefitting most from the EBAE courses, e.g., those who performed well or poorly on the pretest given at the beginning of the course. Finally, we were also interested in the typical correlation between the performance of students on the validated conceptual surveys and their performance on the final exam, which typically places a heavy weight on quantitative physics problems.

3.1.3 Framework for Exploring the Effectiveness of EBAE Pedagogies

We compare introductory physics student performance in EBAE flipped and active-engagement non-flipped courses with LB courses with inspiration from several theoretical frameworks. The overarching framework that is used for the instructional design of all of the EBAE courses in this study (whether flipped or active-engagement non-flipped) was the cognitive apprenticeship model [28, 79, 80]. This framework focuses on providing opportunities to coach students and scaffold their learning. All of the EBAE classes were designed to give students similar coaching and scaffolding to develop their problem solving and reasoning skills. The EBAE courses focused on the cognitive approach to instructional design for various learning units and building on students' prior knowledge in order to help them learn better. For example, Piaget's framework [81], which emphasizes "optimal mismatch" between what a student knows and where the instruction should be targeted in order for desired assimilation and accommodation of knowledge to occur, was helpful in developing the instructional design. A related framework is the theory of conceptual change put forth by Posner et al. [82]. In this framework, conceptual changes or "accommodations" can occur when the existing conceptual understanding of students is not sufficient for or is inconsistent with new phenomena they are learning about. These frameworks also suggest that these accommodations can be very difficult for students, particularly when students are firmly committed to their prior understanding, unless instructional design explicitly accounts for these difficulties. The model suggests that it is important for instructors to be knowledgeable about student ideas, e.g., which they may apply in inappropriate contexts to make incorrect inferences while solving physics problems. Within this framework, students can be motivated by an anomaly which provides a cognitive conflict and illustrates how their conceptions are inadequate for explaining a newly encountered physical situation, so they become dissatisfied with their current

understanding of concepts and improve their understanding. Taking inspiration from these frameworks, EBAE instructors tried to focus on student conceptions and their difficulties in learning physics in order to design instruction that produces the desired cognitive conflict and learning.

3.2 METHODOLOGY

3.2.1 Courses and Participants

The participants in this study were students in 16 different algebra-based and calculus-based introductory physics courses (more than 1500 students in first semester courses and more than 1200 students in the second semester courses) at a typical large research university in the US (University of Pittsburgh). The courses fall into three categories:

1) A lecture-based (or LB) course is one in which the primary mode of instruction was via lecture. In addition to the three or four weekly hours for lectures, students attended an hour long recitation section taught by a graduate TA. In recitation, the TA typically answered student questions (mainly about their homework problems which were mostly textbook style quantitative problems), solved problems on the board and gave students a quiz in the last 10-20 minutes.
2) A flipped course is one in which the class was broken up into two almost equal size groups with each group meeting with the instructor for half the regular class time. For example, for a 200 student class scheduled to meet for four hours each week (on two different days), the instructor met with half the class (100 students) on the first day and the other half on

the second day. This was possible in the flipped classes since the total contact hours for each instructor each week with the students was the same as in the corresponding LB courses. Students watched the lecture videos before coming to class and answered some conceptual questions which were based upon the lecture video content. They uploaded the answers to those conceptual questions before class onto the course website and were graded for a small percentage of their grade (typically 4-8%). Although students had to watch several videos outside of class in preparation for each class, each video was typically 5-10 minutes long, followed by concept questions. On average, students in a flipped class had to watch recorded videos which took a little less than half the allotted weekly time for class (e.g., for the courses scheduled for four hours each week, students watched on average 1.5 hours of videos each week, and in the courses scheduled for three hours each week, students watched around one hour of videos). These video times do not include the time that students would take to rewind the video, stop and think about the concepts and answer the concept questions embedded after the videos that counted for their course grade. In the spirit of JiTT, the instructors of the flipped courses adjusted the in-class activities based upon student responses to online concept questions which were supposed to be submitted the night before the class. About 90% of the students submitted their answers to the concept questions that followed the videos to the course website before coming to the class. The web-platforms used for managing, hosting and sharing these videos and for having online discussions with students about them asynchronously (in which students and the instructor participated) were Classroom Salon or Panopto. In-class time was used for clicker questions involving peer discussion and then a whole class discussion of the concept-tests, collaborative group problem solving involving quantitative problems in which 2-3 students

worked in a group (followed by a clicker question about the order of magnitude for the answer to the quantitative problem on which students worked collaboratively) and lecture-demonstrations with preceding clicker questions on the same concepts. In addition to the regular class times, students attended an hour long recitation section which was taught the same way as for students in the LB courses.

It is important to note that the instructors who taught the flipped courses also taught LB courses at the same time (usually teaching two courses in a particular semester: one flipped and one LB). Students in both flipped and LB courses completed the same homework and took the same final exam. For the calculus-based flipped courses, the students also took the same midterm exams. This was not possible for the algebra-based courses because the exams were scheduled at different times. However, in the algebra-based courses they took the same final exam and had the same homework. Additionally, the instructors attempted to make the actual delivery of content (done via videos in the flipped courses and via in-class lecture in the lecture-based courses) very similar. Essentially, the content of the videos was delivered in-class in the lecture based courses.

3) EBAE interactive non-flipped course. In this course, the instructor combined lectures with research-based pedagogies including clicker questions with peer discussion, conceptual tutorials, collaborative group problem solving, and lecture demonstrations with preceding clicker questions on the same concepts similar to the flipped courses. In addition, students attended a reformed recitation which primarily used context-rich problems to get students to engage in group problem solving or worked on research-based tutorials while being guided by a TA. The instructor ensured that the problems students solved each week in the recitation activities were closely related to what happened in class. Students also worked

on some research-based tutorials during class in small groups, but if they did not finish them in the allotted time, they were asked to complete them as homework.

3.2.2 Materials

The materials used in this study are the conceptual FCI and CSEM multiple-choice (five choices for each question) standardized surveys which were administered in the first week of classes before instruction in relevant concepts (pretest) and the after instruction in relevant concepts (posttest). Apart from the data on these surveys that the researchers collected from all of these courses, each instructor administered his/her own final exam which was mostly quantitative (60%-90% of the questions were quantitative although some instructors had either the entire final exam or part of it in a multiple-choice format with five options for each question to make grading easier). Ten course instructors (who also provided the FCI or CSEM data from their classes) provided their students' final exam scores and most of them also provided a copy of their final exam.

3.2.3 Methods

Our main goal in this investigation was to compare the average performance of students in introductory physics courses that used EBAE pedagogies with the average performance of students in LB courses by using standardized conceptual surveys, the FCI (for physics I) and CSEM (for physics II) as pre-/posttests. We not only calculated the average gain (posttest – pretest scores) for each group but also calculated the average normalized gain, which is commonly used to determine how much the students learned from pretest to posttest taking into account their initial scores on

the pretest. It is defined as $\langle g \rangle = \frac{\%\langle S_f \rangle - \%\langle S_i \rangle}{100 - \%\langle S_i \rangle}$, in which $\langle S_f \rangle$ and $\langle S_i \rangle$ are the final (post) and initial (pre) class averages, respectively. Then, $\text{Norm}\, g = 100\langle g \rangle$ in percent [16]. This normalized gain provides valuable information about how much students have learned by taking into account what they already know based on the pretest. We wanted to investigate whether the normalized gain is higher in one course compared to another.

In order to compare EBAE courses with LB courses, we performed *t*-tests [83] on FCI or CSEM pre- and posttest data. We also calculated the effect size in the form of Cohen's *d* defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma_{pooled}}$, where μ_1 and μ_2 are the averages of the two groups being compared (e.g., EBAE vs. LB) and $\sigma_{pooled} = \sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}$; here σ_1 and σ_2 are the standard deviations of the two groups being compared.

Moreover, although we did not have control over the type of final exam each instructor used in his/her courses, we wanted to look for correlation between the FCI/CSEM posttest performance and the final exam performance for different instructors in the algebra-based and calculus-based EBAE or LB courses. Including both the algebra-based and calculus-based courses, 10 instructors provided the final exam scores for their classes. We used these data to obtain linear regression plots between the posttest and the final exam performance for each instructor and computed the correlation coefficient between the performance of students on the validated conceptual surveys and their performance on the final exam for different instructors. These correlation coefficients between the conceptual surveys and the final exam (with strong focus on quantitative problem solving) can provide an indication of the strength of the correlation between conceptual and quantitative problem solving in introductory physics courses.

Out of all introductory physics courses (algebra-based or calculus-based physics I or II) included in this study, there were four EBAE courses: two completely flipped classes in algebra-based introductory physics I and one completely flipped and one interactive active engagement class in calculus-based introductory physics II.

3.3 RESULTS

Table 3.1 shows the intra-group pre-/posttest data (pooled data for the same type of courses) on the FCI survey for the calculus-based and algebra-based physics I courses. For the algebra-based courses, some were EBAE courses while others were LB courses, whereas all the calculus-based courses were LB. We find statistically significant improvements from the pretest to the posttest for each group but the normalized gain (Norm g) is largest (30%) for the EBAE courses.

Table 3.1. Intra-group FCI pre-/posttest averages (Mean) and standard deviations (SD) for first-semester introductory physics in calculus-based LB courses, and algebra-based EBAE (flipped) and LB courses. The number of students in each group, N, is shown. For each group, a *p*-value obtained using a *t*-test shows that the difference between the pre-/posttest is statistically significant and the normalized gain (Norm g) from pre- to posttest shows how much students learned from what they did not already know based on the pretest.

Type of class	FCI	N	Mean	SD	p-value	Norm g
Calc LB	Pretest	461	51%	21%	<0.001	25%
	Posttest	350	63%	20%		
Alg flipped	Pretest	299	35%	18%	<0.001	30%
	Posttest	262	54%	20%		
Alg LB	Pretest	837	35%	17%	<0.001	23%
	Posttest	738	50%	19%		

Table 3.2 shows the intra-group (pooled data for the same type of courses) pre-/posttest data on the CSEM survey for algebra-based and calculus-based introductory physics II courses. We find that there are statistically significant differences between the pre-/posttest scores for each group but the normalized gain (Norm g) is largest (36%) for the EBAE courses.

Table 3.2. Intra-group CSEM pre-/posttest averages (Mean) and standard deviations (SD) for second-semester introductory physics in calculus-based LB and EBAE courses (here, EBAE flipped and interactive non-flipped courses are combined) and algebra-based LB courses. The total number of students in each group, N, is shown. For each group, a *p*-value obtained using a *t*-test shows that the difference between the pre-/posttest s is statistically significant and the normalized gain (Norm g) from pretest to posttest shows how much students learned from what they did not already know based on the pretest.

Type of Class	CSEM	N	Mean	SD	p-value	Norm g
Calc LB	Pretest	410	38%	14%	<0.001	21%
	Posttest	346	51%	17%		
Calc EBAE	Pretest	346	37%	16%	<0.001	36%
	Posttest	300	60%	19%		
Alg LB	Pretest	514	24%	11%	<0.001	25%
	Posttest	449	43%	17%		

Table 3.3 shows the inter-group FCI pre-/posttest score comparison between algebra-based LB and EBAE courses, first holding the instructor fixed (same instructor taught both the LB and EBAE courses, used the same homework and final exams) and second, combining all instructors who used similar methods in the same group (only one instructor used EBAE methods, but several who taught LB courses were combined). Table 3.3 shows that there is no statistically significant difference between the pretest scores of students in the LB and EBAE courses in introductory physics I on the FCI. Table 3.3 also shows that the effect sizes for comparing FCI posttest performance of students in EBAE courses with students in LB courses are 0.314 (same instructor

teaching both courses in the same semester) and 0.233 when different courses using similar methods are combined (which are considered small effect sizes).

Table 3.3. Inter-group comparison of the average FCI pre-/posttest scores of algebra-based students in LB courses with EBAE courses when (i) both courses are taught by the same instructor and (ii) different instructors using similar instructional methods are combined. The p-values and effect sizes are obtained when comparing the LB and EBAE courses in terms of students' FCI scores.

Comparison of LB and EBAE Groups		FCI-Pre	FCI-Post
(i) FCI Alg: LB vs. EBAE (same instructor)	LB	N: 466 Mean: 35% SD: 17%	N: 433 Mean: 48% SD: 20%
	Comparison LB and EBAE	p-value: 0.831 effect size: 0.017	p-value: <0.001 effect size: 0.314
	EBAE	N: 299 Mean: 35% SD: 18%	N: 262 Mean: 54% SD: 20%
(ii) FCI Alg: LB vs. EBAE (different instructors combined)	LB	N: 837 Mean: 35% SD: 17%	N: 738 Mean: 50% SD: 19%
	Comparison LB and EBAE	p-value: 0.901 effect size: 0.009	p-value: 0.001 effect size: 0.233
	EBAE	N: 299 Mean: 35% SD: 18%	N: 262 Mean: 54% SD: 20%

Table 3.4 shows the inter-group CSEM pre-/posttest score comparison between calculus-based LB and EBAE courses, first holding the instructor fixed (same instructor taught both the LB and EBAE courses and used the same homework and final exams) and second, combining all instructors who taught using similar methods in the same group. Table 3.4 shows that there is no statistically significant difference between the pretest scores of students in the LB and EBAE

courses in introductory physics II on the CSEM. Table 3.4 also shows that the effect sizes for comparing CSEM posttest performance of students in EBAE courses with students in LB courses are 0.357 (same instructor teaching both courses) and 0.494 when different courses using similar methods are combined (which are considered medium effect sizes).

Table 3.4. Inter-group comparison of the average CSEM pre-/posttest scores of calculus-based students in LB courses with EBAE courses when (i) both courses are taught by the same instructor and (ii) different instructors using similar instructional methods are combined. The p-values and effect sizes are obtained when comparing the LB and EBAE courses in terms of students' CSEM scores.

Comparison of LB and EBAE Groups		CSEM Pre	CSEM Post
(i) CSEM Calc: LB vs. EBAE (same instructor)	LB	N: 178 Mean: 40% SD: 13%	N: 154 Mean: 48% SD: 15%
	Comparison LB and EBAE	p-value: 0.895 effect size: 0.013	p-value: 0.001 effect size: 0.357
	EBAE	N: 208 Mean: 40% SD: 15%	N: 181 Mean: 54% SD: 19%
(ii) CSEM Calc LB vs. EBAE (different instructors pooled)	LB	N: 410 Mean: 38% SD: 14%	N: 346 Mean: 51% SD: 17%
	Comparison LB and EBAE	p-value: 0.886 effect size: 0.011	p-value: <0.001 effect size: 0.494
	EBAE	N: 346 Mean: 37% SD: 16%	N: 300 Mean: 60% SD: 19%

Table 3.5 shows the average FCI pre-/posttest scores for algebra-based and CSEM pre-/posttest scores for calculus-based courses (Av-Pre/Post), gain (Post – Pre), normalized gain (Norm g), and final exam scores (Av-Fin) for students in the flipped and LB courses taught by the

same instructor (with the same homework and final exam) with students divided into three groups based on their pretest scores. A closer look at the gains and normalized gains for the courses taught by the same instructor shows that students in all of the three pretest score categories in the flipped courses had higher gains and normalized gains compared to those in the LB courses taught by the same instructor. Moreover, for algebra-based physics I, the average final exam scores of the students in the flipped course taught by the same instructor in all the three pretest categories are somewhat higher than the LB course.

Table 3.5. Average FCI pre-/posttest scores for algebra-based and CSEM pre-/posttest scores for calculus-based courses (Av-Pre/Post), Gain (Post – Pre), normalized gain (Norm g) and final exam scores (Av-Fin) for students in the flipped and LB courses taught by the same instructor (with same homework and final exam) with students divided into three groups based on their pretest scores as shown. Students in the LB or flipped courses in the shaded region can be compared with each other and those in the unshaded region can be compared with each other.

	Pretest Split	Av-Pre	Av-Post	Gain	Norm g	Av-Fin
FCI Alg LB (Instructor 1)	bottom 1/3	18	36	18	22	48
	middle 1/3	32	45	13	20	54
	top 1/3	54	66	12	27	65
FCI Alg Flipped (Instructor 1)	bottom 1/3	17	41	24	29	54
	middle 1/3	32	49	17	25	54
	top 1/3	56	74	18	40	65
CSEM Calc LB (Instructor 2)	bottom 1/3	26	35	9	12	43
	middle 1/3	39	46	8	12	53
	top 1/3	53	60	6	14	59
CSEM Calc Flipped (Instructor 2)	bottom 1/3	25	42	18	24	51
	middle 1/3	39	49	11	18	56
	top 1/3	58	70	12	29	69

Table 3.6 shows the average FCI pre-/posttest scores for algebra-based and calculus-based courses (Av-Pre/Post), gain (Post-Pre), and normalized gain (Norm g) for students in the flipped

and LB courses with students divided into three groups based on their pretest scores. All equivalent (algebra-based or calculus-based physics I) courses which used the same instructional strategy (flipped or LB) were combined and students were divided into three groups based upon their pretest scores. A closer look at the gains and normalized gains for the algebra-based courses (for which there are both flipped and LB groups) shows that students in all of the three pretest score categories in the flipped courses had higher gains and normalized gains compared to those in the traditional courses. In the calculus-based LB courses, the highest one-third of the students had 83% and 82% as their FCI pretest and posttest scores, respectively. In Table 3.6, we do not list the average final exam performance since instructors used different exams which varied in difficulty.

Table 3.6. Average FCI pre-/posttest scores (Av-Pre/Post), Gain (Post – Pre), and normalized gain (Norm g) for students in the flipped and LB algebra-based and calculus-based courses. All courses in the same group were combined with students divided into three groups based upon their pretest scores as shown. Students in the LB or flipped courses in the shaded region can be compared with each other.

	Pretest Split	Av-Pre	Av-Post	Gain	Norm g
FCI Calc LB	bottom 1/3	31	46	15	22
	middle 1/3	55	68	13	28
	top 1/3	83	82	-1	-7
FCI Alg Flipped	bottom 1/3	17	41	24	29
	middle 1/3	32	49	17	25
	top 1/3	56	74	18	40
FCI Alg LB	bottom 1/3	19	35	16	19
	middle 1/3	33	46	14	20
	top 1/3	55	68	14	30

Table 3.7 shows the average CSEM pre-/posttest scores for algebra-based and calculus-based courses (Av-Pre/Post), gain (Post-Pre), and normalized gain (Norm g) for students in the EBAE and LB courses with students divided into three groups based on their pretest scores. All

equivalent (algebra-based or calculus-based physics II) courses which used the same instructional strategy (EBAE or LB) were combined and students were divided into three groups based upon their pre-test scores. A closer look at the gains and normalized gains for the calculus-based courses (for which there are both EBAE and LB groups) shows that students in all of the three pretest score categories in the EBAE courses had higher gains and normalized gains than those in the traditional courses.

Table 3.7. Average CSEM pre-/posttest scores (Av-Pre/Post), Gain (Post – Pre), and normalized gain (Norm g) for calculus-based students in the EBAE and LB courses and algebra-based students in LB courses. All courses in the same group were combined with students divided into three groups based upon their pretest scores as shown. Students in the LB or flipped courses in the shaded region can be compared with each other.

	Pretest Split	Av-Pre	Av-Post	Gain	Norm g
Calc EBAE CSEM	bottom 1/3	22	51	29	37
	middle 1/3	35	57	22	34
	top 1/3	56	70	15	33
Calc LB CSEM	bottom 1/3	23	39	16	20
	middle 1/3	35	47	12	19
	top 1/3	51	59	8	16
Alg LB CSEM	bottom 1/3	15	36	22	25
	middle 1/3	23	44	21	27
	top 1/3	35	51	16	25

We should note that differences in normalized gain should be interpreted carefully because we do not have a measure of the variability of normalized gain, and thus, differences of 5% may or may not be significant. We stress that we are not making statements about significant differences between EBAE courses and LB courses based on normalized gain, and any statements we have made about significant differences are supported by statistical analyses (e.g., Cohen's d, or comparison of post-test results).

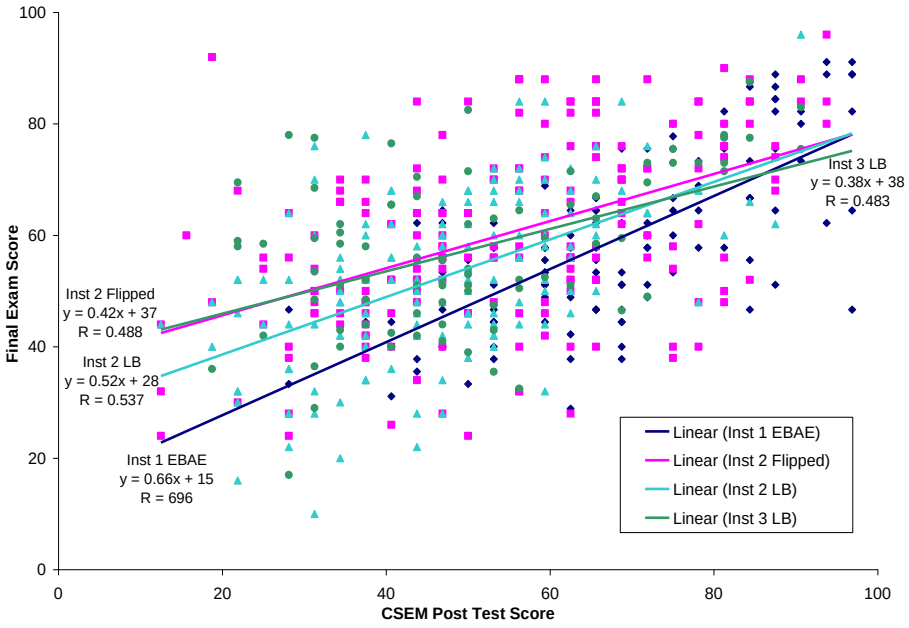

Figure 3.1 Linear regression of the CSEM posttest scores (conceptual) and final exam scores (heavy focus on quantitative problems) for four calculus-based introductory physics courses shows the correlation coefficients between 0.438-0.598. There were no clear trends in the correlation coefficients based upon whether the instructor (Inst) used EBAE strategies or whether the class was LB.

Figure 3.1 shows the CSEM posttest performance along with the final exam performance for three different instructors in flipped and LB calculus-based courses (one instructor taught an EBAE and an LB course, two instructors taught LB courses). Figure 3.1 shows that the linear regressions [83] for the flipped and LB courses are fairly similar and that there is a moderate correlation between CSEM posttest scores and final exam scores. We also plotted linear regressions for the algebra-based courses, but the data look similar to Figure 3.1 and so are not included here. Instead, we include all the correlation coefficients (CSEM posttest vs. final exam) for all the courses for which we managed to obtain posttest data. Table 3.8 summarizes the correlation coefficients between post-CSEM/FCI and final exam scores for each instructor who provided final exam data.

Table 3.8. Correlation coefficients (R) between post-CSEM/FCI and final exam scores for each instructor (Inst) who provided final exam data. The final exam data were not provided by physics II instructors in algebra-based courses.

Physics I (Calc)		**Physics I (Alg)**		**Physics II (Calc)**	
Instructor and course type	R	Instructor and course type	R	Instructor and course type	R
Inst 1: LB	0.495	Inst 1: Flipped	0.559	Inst 1: EBAE	0.696
Inst 2: LB	0.589	Inst 1: LB	0.516	Inst 2: Flipped	0.488
Inst 3: LB	0.787	Inst 2: LB	0.693	Inst 2: LB	0.537
				Inst 3: LB	0.483

3.4 DISCUSSION AND SUMMARY

In all cases investigated, we find that on average, introductory physics students in the courses which made significant use of evidence-based active engagement (EBAE) methods outperformed students in courses primarily taught using lecture-based (LB) instruction on standardized conceptual surveys (FCI or CSEM) in the posttest even though there was no statistically significant difference on the pretest. This was true both in the algebra-based and calculus-based physics I (primarily mechanics) and II (primarily E&M) courses. Also, the differences between EBAE and LB courses were observed both among students who performed well on the FCI/CSEM pretest (given in the first week of classes) and also those who performed poorly, thus indicating that EBAE instructional strategies help students at all levels.

On the other hand, the typical effect size for the differences between equivalent EBAE and traditional courses is between 0.23-0.49, which is considered small to medium. Thus, the benefits of these EBAE approaches were not as large as one may expect to observe. Why might that be the

case and how can instructors enhance student learning more than that observed in this investigation using EBAE instructional strategies?

There are many potential challenges to using EBAE instructional strategies. Below, we list some of the possible challenges and some strategies that may reduce those challenges. Many of these have been described elsewhere [68, 84-96] so we provide only a short summary:

- Lack of student engagement even with well-designed learning tools, which may occur for many different reasons: lack of student motivation, poor self-efficacy or poor time-management skills on the part of the students, lack of effective incentives for students to engage with the self-paced learning tools, etc. Strategies to address some of these difficulties have been described, e.g., providing students with effective strategies to learn [97, 98], using certain communication activities to foster student motivation [84, 85]. Other strategies to address these potential issues have also been described [87-89, 91-93].
- Lack of student engagement with in-class active learning activities (e.g., group problem solving). Many strategies to help address this issue have been described, e.g., designing in-class activities which foster both individual accountability and positive inter-dependence [11-12]. One example of fostering individual accountability is to include a short quiz or clicker questions related to content students were supposed to learn when working in groups, and positive interdependence means that the success of each student in a group is dependent on the success of others. For other strategies to help foster student engagement see Refs. [84, 94, 95] and references therein.
- Student misconceptions about learning, or resistance to EBAE instructional strategies which could be addressed at the beginning of the term by framing the instructional design

of the class [68, 96] and providing data on the effectiveness of the evidence-based strategies being used (and conversely the ineffectiveness of e.g., instructor explanations [99]).

- Large class sizes can be an impediment, and one approach faculty have used in flipped courses is to split the class in two, thus forming smaller class sizes as well as more room for students to form groups and move around the classroom. Undergraduate or graduate teaching assistants can also help in facilitating in-class activities. In group activities, students often work at different rates, and students who finish early can help others.
- Content coverage. There is often a lot of content covered in introductory physics courses and it may be challenging to cover the same amount of content while also including frequent active learning activities during class. Moving some of the content delivery outside of class (e.g., some pre-lecture reading or videos on certain 'easier' concepts, or moving the entire content delivery outside of class like in a flipped course) can help provide additional time for in-class activities.

We note that the instructors who taught the EBAE courses has control of designing the courses themselves and the researchers only provided them with guidance before and during they worked on designing the courses. The instructors may not have addressed some of the potential issues mentioned above sufficiently, e.g., by framing the courses at the beginning of the term, and providing incentives for students to engage both with in-class and out-of-class activities. However, these issues are challenging to fully address, especially in large classes such as those involved in this investigation (as suggested by the data), and iterative refinement of a course is needed in order to address them. Lastly, while we provided the instructors with information about active learning materials developed by physics education researchers, discussions indicated that they adapted or

created some of their own materials to fit the way the preferred to teach, and the extent to which the materials they adapted or created are conducive to effective learning is unclear.

In addition, Henderson and Dancy [100] found that many instructors try certain EBAE instructional strategies, but some discontinue use after one or two semesters. The faculty members who persist are usually the ones who get support from their peers (e.g., developing faculty learning communities, working with instructional designers at local teaching and learning centers) because there may be many implementation difficulties specific to a particular university even if a particular EBAE approach has been found to be effective elsewhere. Interacting with others, even from different departments, who have been engaged in evidence-based teaching (e.g., visiting their classes, getting feedback from them about one's own classes, etc.) can be extremely valuable. Often, teaching and learning centers are happy to send someone to observe a class and provide feedback as well as suggestions for future active learning activities.

Furthermore, we note that in this study, we found that student performance in a non-flipped EBAE course (which used active learning interspersed with short lectures in class) was comparable to student performance in a flipped course. It is important to point out that the instructor in this EBAE course ensured that the recitations are effectively used to promote active learning and that the activities used in the recitation were closely tied to the course learning goals. Flipping a course can be a time consuming process especially if the instructor is developing his/her own lecture videos for the first time and he/she has not implemented the EBAE strategies in his/her class earlier. Therefore, it is encouraging to observe that one does not need to flip his/her course completely, but can introduce EBAE activities in regular class and also in recitation. These active learning activities and materials can be modified and improved after each use by getting feedback from the students (and also getting feedback from the TAs teaching recitation).

As discussed earlier, learning gains in EBAE courses were not as high as one might expect. This should not be taken as discouragement, but rather as an indication that effective teaching is an iterative pursuit and one should learn from each course implementation and try to improve. The expectation that introducing a lot of EBAE instructional strategies which have been found to be effective elsewhere will result in large gains without refining the material and implementation can deter instructors from continuing the use of EBAE instructional strategies when the results are less than expected, especially given the time commitment reformed teaching can take initially. Instead, one should continue to make refinement and remember that any improvement in student learning is worth the effort!

In summary, in order to enhance student learning in EBAE classes it important not only to develop effective EBAE learning tools and pedagogies commensurate with students' prior knowledge but also to investigate how to implement them appropriately and how to motivate and incentivize their usage to get buy-in from students in order for them to engage with them as intended. Furthermore, for flipped classes, it is especially important to investigate strategies for having a diverse group of students engage with self-paced learning tools effectively. Investigation of various factors that can deter or incentivize their use is essential in order to develop a holistic learning environment to help students with diverse backgrounds benefit from the self-paced learning tools. Additionally, it will be valuable to examine and compare the effectiveness of self-paced learning tools, e.g., videos and concept questions provided to students in flipped classes, when implemented in a controlled environment in which students must effectively engage with the tool one-on-one in front of a researcher vs. an environment in which students are free to use the tool in whatever manner they choose. A framework for understanding and optimizing the factors that can support or hinder effective use of self-paced learning tools, e.g., those students are asked

to engage with in flipped courses, would be helpful in developing and implementing self-paced tools conducive to learning.

3.5 CHAPTER REFERENCES

1. A. Noack, T. Antimirova, and M. Milner-Bolotin, Student diversity and the persistence of gender effects on conceptual physics learning. Can. J. Phys. **87**(12), 1269 (2009).

2. H. Eshach, The use of intuitive rules in interpreting students' difficulties in reading and creating kinematic graphs, Can. J. Phys. **92**(1), 1 (2014).

3. E. M. Kennedy, and J.R. de Bruyn, Understanding of mechanical waves among second-year physics majors, Can. J. Phys. **89**(11), 1155 (2011).

4. A. C. K. Leung, A. Terrana, and S. Jerzak, Students' opinions on the educational value of physics laboratories: a cross-sectional survey, Can. J. Phys. **94**(9), 913 (2016).

5. A. Terrana, A. C. K. Leung, and S. Jerzak, Use of online resources by physics students in Canada: A cross sectional study, Can. J. Phys. **95**(2), 201 (2017).

6. H. Eshach, and I. Kukliansky, Developing of an instrument for assessing students' data analysis skills in the undergraduate physics laboratory, Can. J. Phys. **94**(11), 1205 (2016).

7. L. C. McDermott, Millikan Lecture 1990: What we teach and what is learned-Closing the gap, Am. J. Phys. **59**, 301 (1991).

8. E. Kim and S. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, Am. J. Phys. **70** (7), 759 (2002).

9. J. Fraser, A. Timan, K. Miller, J. Dowd, L. Tucker, and E. Mazur, Teaching and physics education research: bridging the gap, Reports on Progress in Physics **77** (3), 032401 (2014).

10. J. Docktor and J. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST PER **10**, 020119 (2014).

11. P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. 1. Group vs individual problem solving, Am. J. Phys. **60**, 627 (1992).

12. P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. 2. Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).

13. L. DesLauriers, E. Schelew and C. Wieman, Improved learning in a large-enrollment physics class, Science **332**, 862 (2011).

14. D. Hammer, Student resources for learning introductory physics, Am. J. Phys., Physics Education Research Supplement **68** (S1), S52 (2000).

15. E. Redish, R. Steinberg and J. Saul, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

16. R. Hake, Interactive engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66,** 64 (1998).

17. C. Singh, What can we learn from PER: Physics Education Research? The Phys. Teach. **52**, 568 (2014).

18. C. Singh, When physical intuition fails, Am. J. Phys. **70**(11), 1103 (2002).

19. F. Reif, Millikan Lecture 1994: Understanding and teaching important scientific thought processes, Am. J. Phys. **63**, 17 (1995).

20. B. Eylon and F. Reif, Effects of knowledge organization on task performance Cognition Instruct. **1** (1), 5 (1984).

21. S. Y. Lin and C. Singh, Effect of scaffolding on helping introductory physics students solve quantitative problems involving strong alternative conceptions, Phys. Rev. ST PER **11,** 020105 (2015).

22. S.Y. Lin and C. Singh, Using isomorphic problem pair to learn introductory physics: Transferring from a two-step problem to a three-step problem, Phys. Rev. ST PER **9,** 020114 (2013).

23. S.Y. Lin and C. Singh, Using isomorphic problems to learn introductory physics, Phys. Rev. ST PER **7**, 020104 (2011).

24. A. J. Mason and C. Singh, Assessing expertise in introductory physics using categorization task, Phys. Rev. ST PER **7**, 020110 (2011).

25. C. Singh, Categorization of problems to assess and improve proficiency as teacher and learner, Am. J. Phys. **77**, 73 (2009).

    S. Y. Lin and C. Singh, Categorization of quantum mechanics problems by professors and students, Euro. J. Phys. 31, 57 (2010).

26. C. Singh, Assessing student expertise in introductory physics with isomorphic problems. I. Performance on a non-intuitive problem pair from introductory physics, Phys. Rev. ST PER **4**, 010104 (2008).

27. C. Singh, Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer, Phys. Rev. ST PER **4**, 010105 (2008).

28. A. Collins, J. S. Brown, and S. E. Newman, Cognitive Apprenticeship: Teaching the crafts of reading, writing and mathematics, in *Knowing, learning, and instruction: Essays in honor of Robert Glaser*, edited by L. B. Resnick (Lawrence Erlbaum, Hillsdale, NJ, 1989), p. 453.

29. C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69,** 970 (2001).

30. N. Lasry, E. Mazur and J. Watkins, Peer instruction: From Harvard to the two-year college, Am. J. Phys. **76** (11) 1066 (2008).

31. A. J. Mason and C. Singh, Helping students learn effective problem solving strategies by reflecting with peers, Am. J. Phys. **78**, 748 (2010).

32. A. J. Mason and C. Singh, Impact of guided reflection with peers on the development of effective problem solving strategies and physics learning, The Phys. Teach. **54**, 295 (2016).

33. L. McDermott, P. Shaffer and the Physics Education Group at the University of Washington, Tutorials in Introductory Physics, Pearson Publishing, Inc. (2003).

34. C. Singh, Interactive learning tutorials on quantum mechanics, Am. J. Phys. **76** (4), 400 (2008).

35. E. Marshman and C. Singh, Interactive tutorial to improve student understanding of single photon experiments involving a Mach-Zehnder Interferometer, Euro. J. Phys. **37**, 024001 (2016).

36. C. Kalman, M. Milner-Bolotin, and T. Antimirova, Comparison of the effectiveness of collaborative groups and peer instruction in a large introductory physics course for science majors. Can. J. Phys. **88**(5), 325 (2010).

37. C. Singh, Impact of peer interaction on conceptual test performance, Am. J. Phys. **73**(5), 446 (2005).

38. C. Singh, Effectiveness of group interaction on conceptual standardized test performance, Proceedings of the 2002 Phys. Ed. Res. Conference, Boise (Eds. S. Franklin, K. Cummings and J. Marx), p. 67 (2002). http://dx.doi.org/10.1119/perc.2002.pr.017

39. R. Sayer, E. Marshman and C. Singh, The impact of peer interaction on the responses to clicker questions in an upper-level quantum mechanics course, Proc. 2016 Phys. Educ. Res. Conf., Sacramento, CA, p. 304, (2016). http://dx.doi.org/10.1119/perc.2016.pr.071

40. P. Black and D. Wiliam, Assessment and classroom learning, Assessment in Education **5** (1), 7 (1998).

41. R. Moll and M. Milner-Bolotin, The effect of interactive lecture experiments on student academic achievement and attitudes towards physics. Can. J. Phys. **87**(8), 917 (2009).

42. G. Novak, E.T. Patterson, A. Gavrin, and W. Christian, Just-in-Time Teaching: Blending Active Learning with Web Technology Upper Saddle River, NJ: Prentice Hall (1999).

43. R. Sayer, E. Marshman and C. Singh, A case study evaluating Just-in-Time Teaching and Peer Instruction using clickers in a quantum mechanics course, Phys Rev PER **12**, 020133 (2016).

44. https://cft.vanderbilt.edu/guides-sub-pages/flipping-the-classroom/

45. https://perusall.com/

46. D. Schwartz and J. Bransford, A time for telling, Cognition and Instruction **16**, 475 (1998).

47. R. Mayer, *Multimedia Learning* (Cambridge Press, 2001).

48. Z. Chen, T. Stelzer, and G. Gladding, Using multi-media modules to better prepare students for introductory physics lecture, Phys. Rev. ST PER **6**, 010108 (2010).

49. Z. Chen, and G. Gladding, How to make a good animation: A grounded cognition model of how visual representation design affects the construction of abstract physics knowledge, Phys. Rev. ST PER **10**, 010111 (2014).

50. M. C. Kim, and M. J. Hannafin, Scaffolding problem solving in technology-enhanced learning environments (TELEs): Bridging research and theory with practice, Comput. Educ. **56**(2), 403 (2011).

51. F. Reif and L. Scott, Teaching scientific thinking skills: Students and computers coaching each other, Am. J. Phys. **67** (9), 819 (1999).

52. N. Schroader, G. Gladding, B. Guttman, and T. Stelzer, Narrated animated solution videos in a mastery setting, Phys. Rev. ST PER **11**, 010103 (2015).

53. R. Azevedo, Computer environments as metacognitive tools for enhancing learning, Educ. Psychol., Special Issue on Computers as Metacognitive Tools for Enhancing Student Learning **40** (4), 193 (2005).

54. G. Gladding, B. Gutmann, N. Schroeder, and T. Stelzer, Clinical study of student learning using mastery style versus immediate feedback online activities, Phys. Rev. ST PER **11**, 010114 (2015).

55. M. Bower, B. Dalgarno, G. Kennedy, M. Lee, and J. Kenney, Design and implementation factors in blended synchronous learning environments: Outcomes from a cross-case analysis, Comput. Educ. **86**, 1 (2015).

56. V. Chandra and J. Watters, Re-thinking physics teaching with web-based learning, Comput. Educ. **58 (**1), 631 (2012).

57. C. Kulik, J. Kulik, and R. Bangert-Drowns, Effectiveness of mastery learning programs: A meta-analysis, Rev. Educ. Res. **60**, 265 (1990).

58. C. Kulik and J. Kulik, Effectiveness of computer-based instruction: An updated analysis, Comput. Hum. Behav. **7** (12), 75 (1991).

59. J. Kulik, Meta-analytic studies of findings on computer-based instruction. In *Technology Assessment in Education and Training*, edited by E. Baker and H. O'Neil, Jr. (Routledge, New York, NY, 1994) p. 9.

60. R. Azevedo, J. Guthrie, and D. Seibert, The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia, J. Educ. Comput. Res. **30**, 87 (2004).

61. D. Moos and R. Azevedo, Exploring the fluctuation of motivation and use of self-regulatory processes during learning with hypermedia, Instr. Sci. **36**, 203 (2008).

62. J. Greene, I. Costa, J. Robertson, Y. Pan, and V. Deekens, Exploring relations among college students' prior knowledge, implicit theories of intelligence, and self-regulated learning in a hypermedia environment, Comput. Educ. **55** (3), 1027 (2010).

63. C. M. Chen and C. H. Wu, Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance, Comput. Educ. **80**, 108 (2015).

64. C. Singh, Interactive video tutorials for enhancing problem-solving, reasoning, and meta-cognitive skills of introductory physics students, in Proc. 2003 Phys. Ed. Res. Conf. Madison, WI, AIP Publishing Melville NY, 2004, p.177. http://dx.doi.org/10.1063/1.1807283

65. C. Singh and D. Haileselassie, Developing problem solving skills of students taking introductory physics via web-based tutorials, J. Coll. Sci. Teaching **39 (**4), 34 (2010).

66. D. Berrett How 'flipping' the classroom can improve the traditional lecture. The Chronicle of Higher Education, Feb. 19 (2012).

67. E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Engelwood Cliffs, 1997).

68. D. C. Haak, J. HilleRisLambers, E. Pitre, and S. Freeman, Increased structure and active learning reduce the achievement gap in introductory biology, Science **332**, 1213 (2011).

69. L. Breslow, D. Pritchard, J. DeBoer, G. Stump, A. Ho, and D. Seaton, Studying learning in the worldwide classroom research into edX's first MOOC, Res. Pract. Assess. **8**, 13 (2013).

70. P. Laws, M. Willis, D. Jackson, K. Koenig and R. Teese Using research-based interactive video vignettes to enhance out-of-class learning in introductory physics, The Phys. Teach. **53**, 114 (2015).

71. D. Hestenes, M. Wells and G. Swackhamer, Force Concept Inventory, The Phys. Teach. **30,** 141 (1992).

72. D. Maloney, T. O'Kuma, C. Hieggelke and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. Supplement **69** (7), s12 (2001).

73. C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys **71**(6), 607 (2003).

74. L. Rimoldini and C. Singh, Student understanding of rotational and rolling motion concepts, Phys. Rev. ST PER **1,** 010102 (2005).

75. L. Ding, R. Chabay, B. Sherwood and R. Beichner, Valuating an assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST PER **1,** 10105 (2006).

76. C. Singh and D. Rosengrant, Students' conceptual knowledge of energy and momentum, Proc. Phys. Educ. Res. Conf., Rochester, p. 123 (2001). http://dx.doi.org/10.1119/perc.2001.pr.018

77. J. Li and C. Singh, Developing and validating a conceptual survey to assess introductory physics students' understanding of magnetism, Euro. J. Phys. **38** (2), 025702 (2017).

78. C. Singh, Student understanding of symmetry and Gauss's law of electricity, Am. J. Phys. **74** (10), 923 (2006).

79. A. Schoenfeld, Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics, *Handbook for Research on Mathematics Teaching and Learning* (NY: McMillan) 1992.

80. J. Heller and F. Reif, Prescribing effective human problem-solving processes: problem description in physics, Cognition and Instruction **1**, 177 (1984).

81. H. Ginsberg and S. Opper, *Piaget's Theory of Intellectual Development* (Prentice Hall, Englewood Cliffs, 1969).

82. G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 211 (1982).

83. G. Glass, K. Hopkins, *Statistical Methods in Education and Psychology*, 3rd Ed. Pearson 1996.

84. J. Bergmann, and A. Sams, Flip your classroom: Reach every student in every class every day (International Society for Technology in Education, Eugene, OR, 2012).

85. J. Kerssen-Griep, Teacher communication activities relevant to student motivation: Classroom facework and instructional communication competence, Communication Education and Instructional Processes **50**, 256 (2001).

86. S. B. Seidel, and K. D. Tanner, What if students revolt?—considering student resistance: origins, options, and opportunities for investigation, CBE-Life Sciences Education **12**, 586 (2013).

87. K. D. Tanner, Structure matters: Twenty-one teaching strategies to promote student engagement and cultivate classroom equity, CBE-Life Sciences Education **12**, 322 (2013).

88. R. Felder and R. Brent, Teaching and learning STEM: A practical guide (Jossey-Bass, San Francisco, CA, 2016).

89. M. Boekaerts, The crucial role of motivation and emotion in classroom learning, The Nature of Learning: Using Research to Inspire Practice, H. 91-111 (2010).

90. D. E. Ellis, Students' Responses to Innovative Instructional Methods: Exploring Learning-Centred Methods and Barriers to Change, UWSpace (2013).

91. P. Pintrich, A motivational science perspective on the role of student motivation in learning and teaching contexts. J. Educ. Psych. **95**, 667 (2003).

92. L. Ferlazzo, *Building a Community of Self-Motivated Learners: Strategies to Help Students Thrive in School and Beyond* (Routledge, 2015).

93. Motivating Learning, by the Carl Wieman Science Education Initiative. Retrieved online 7/7/2017 at http://www.cwsei.ubc.ca/resources/files/Motivating-Learning_CWSEI.pdf.

94. J. A. Fredericks, P. C. Blumenfeld, and A. H. Paris, School Engagement: Potential of the concept, state of the evidence, Rev. Educ. Res. 74, 59 (2004).

95. J. D. Klein and H. L. Schnackenberg, Effects of informal cooperative learning and the affiliation motive on achievement, attitude, and student interactions, Contemporary Educational Psychology **25**, 332 (2000).

96. G. A. Smith, First-day questions for the learner-centered classroom, National Teaching and Learning Forum, **17**, 1 (2008).

97. B. Oakley, *A Mind for Numbers: How to Excel in Math and Science (Even if You Flunked Algebra)*, Penguin, July, 2014.

98. https://www.youtube.com/watch?v=23Xqu0jXlfs

99. Z. Hrepic, D. Zollman, and S. Rebello, Students' understanding and perceptions of the content of a lecture, AIP Conf. Proc. **720**, 189 (2004).

100. C. Henderson and M. Dancy, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, Phys. Rev. ST PER **8**, 020104 (2012).

4.0 IMPACT OF EVIDENCE-BASED ACTIVE-ENGAGEMENT COURSES ON GENDER GAP IN INTRODUCTORY PHYSICS

4.1 INTRODUCTION

4.1.1 Physics Education Research-based Active Engagement Methods

In the past few decades, physics education research has identified challenges that students encounter in learning physics at all levels of instruction [1-7]. Building on these investigations, researchers are developing, implementing and evaluating evidence-based curricula and pedagogies to reduce these challenges to help students develop a coherent understanding of physics concepts and enhance their problem solving, reasoning and metacognitive skills [8-18]. In evidence-based curricula and pedagogies, the learning goals and objectives, instructional design, and assessment of learning are aligned with each other and there is focus on evaluating whether the pedagogical approaches employed have been successful in meeting the goals and enhancing student learning.

One highly successful model of learning is the field-tested cognitive apprenticeship model [19]. According to this model, students can learn effectively if the instructional design involves three essential components: "modeling", "coaching and scaffolding", and "weaning". In this approach, "modeling" means that the instructional approaches demonstrate and exemplify the criteria for good performance and the skills that students should learn (e.g., how to solve physics problems systematically). "Coaching and scaffolding" means that students receive guidance and support as they actively engage in learning the content and skills necessary for good performance. "Weaning" means gradually reducing the support and feedback to help students develop self-

reliance [19]. In traditional physics instruction, especially at the college level, there is often a lack of coaching and scaffolding: students come to class where the instructor lectures and does some example problems, then students are left on their own to work through homework with little or no feedback. This lack of prompt feedback and scaffolding can be detrimental to learning.

Some of the commonly used evidence-based active-engagement (EBAE) approaches implemented in physics include peer instruction with clickers popularized by Eric Mazur from Harvard University [20-22], tutorial-based instruction in introductory and advanced courses [23-25] and collaborative group problem solving [26-29], e.g., using context-rich problems [4-5]. In all of these evidence-based approaches, formative assessment plays a critical role in student learning [30]. Formative assessment tasks are frequent, low-stakes assessment activities which give feedback to students and instructors about what students have learned at a given point. Using frequent formative assessments helps make the learning goals of the course concrete to students, and provides them with a way to track their progress in the course with respect to these learning goals. When formative assessment tasks such as concept-tests, tutorials and collaborative group problem solving are interspersed throughout the course, learning is enhanced [30-31].

Moreover, technology is increasingly being exploited for pedagogical purposes to improve student learning. For example, Just-in-Time Teaching (JiTT) is an instructional approach in which instructors receive feedback from students before class and use that feedback to tailor in-class instruction [32-33]. Typically, students complete an electronic pre-lecture assignment in which they give feedback to the instructor regarding any difficulties they have had with the assigned reading material, lecture videos, and/or other self-paced instructional tools. The instructor then reviews student feedback before class and makes adjustments to the in-class activities. For example, during class, the instructor can focus on student difficulties found via electronic

feedback. Students may engage in discussions with the instructor and with their classmates, and the instructor may then adjust the next pre-lecture assignment based on the progress made during class. When JiTT was first conceived and implemented in the late 1990s in physics classes, the required internet technology for electronic feedback was still evolving; developments in digital technology since then have continued to make electronic feedback from students and the JiTT approach easier to implement in classes. For example, Eric Mazur's Perusall system [34] allows students to read the textbook and ask questions electronically and the system uses their questions to draft a "confusion report" which distills their questions to three most common difficulties, which can be addressed in class. It has been hypothesized that JiTT may help students learn better because out-of-class activities cause students to engage with and reflect on the parts of the instructional material they find challenging [32-33]. In particular, when the instructor focuses on student difficulties in lecture which were found via electronic feedback before class, it may create a "time for telling" [35] especially because students may be "primed to learn" better when they come to class if they have struggled with the material during pre-lecture activities. The JiTT approach is often used in combination with peer discussion and/or collaborative group problem solving inter-dispersed with lectures in the classroom.

In addition, in the last decade, the JiTT pedagogy has been extended a step further with the maturing of technology [36-41] and "flipped" [42, 43] classes with limited in-class lectures have become common with instructors asking students to engage with short lecture videos (or read certain section of the textbook) and concept questions associated with each video outside of the class and using most of the class-time for active-engagement. The effectiveness of flipped classes in enhancing student learning can depend on many factors including the degree to which evidence-based pedagogies that build on students' prior knowledge and actively engage them in the learning

process are used, whether there is sufficient buy-in from students, and the incentives that are used to get students engaged with the learning tools both inside and outside the classroom.

Moreover, research suggests that effective use of peer collaboration can enhance student learning in many instructional settings in physics classes, including in JiTT and flipped environments, and with various types and levels of student populations. Although the details of implementation vary, students can learn from each other in many different environments. For example, in Mazur's peer instruction approach [44], the instructor poses concrete conceptual problems in the form of conceptual multiple-choice clicker questions to students throughout the lecture and students discuss their responses with their peers. Heller et al. have shown that collaborative problem solving with peers in the context of quantitative "context-rich" problems [4-5] can be valuable both for learning physics and for developing effective problem solving strategies.

In evidence-based "active-engagement non-flipped" courses [45], lecture and interactive activities are combined during the prescribed class time to enhance student learning and students' out-of-class homework assignments are often similar to those assigned in traditionally taught classes. On the other hand, in flipped courses, there is very limited direct instruction (lecture) and the majority of in-class time is used to actively engage students in learning. The effectiveness of flipped classes depends on how the course is designed and incentivized and how out-of-class activities build on in-class activities. In addition, whether instructors create a low or high anxiety active-learning environment can play a critical role in student engagement. It can particularly impact learning for women and students from other underrepresented groups, whose sense of belonging and self-efficacy can either be enhanced or exacerbated depending upon the design of the active-learning environment. More about these types of issues is discussed in the next section.

Also, the lecture videos that students often watch outside of the class in a flipped class are self-paced, which has both advantages and disadvantages. While pedagogically developed, implemented and incentivized self-paced videos can provide a variety of students with an opportunity to learn at a pace that is commensurate with their prior knowledge, without appropriate pedagogy in the development, implementation, and incentives to learn from these tools, students may not engage with them as intended, especially if they do not have good time-management and self-regulation skills. For example, research on Massive Open Online Courses (MOOCs) [46] suggests that a majority of those who complete the entire online course already have a bachelor's degree. Moreover, a student who does not keep up with out-of-class activities such as watching videos and answering the concept questions associated with them before coming to class is unlikely to take full advantage of the interactive in-class activities in a flipped class. Thus, while a well-designed and implemented flipped course has the potential to help a variety of students learn to think like a physicist and can scaffold their learning of physics, many students may not engage and learn from the out-of-class videos if they are not intrinsically motivated and if the videos are not effective [36] or are not implemented and incentivized appropriately. Despite these caveats, well-designed and well-implemented interactive videos [47] and associated questions designed carefully can be beneficial as they can help a variety of students with different prior preparations and allow them to learn at their own pace. Moreover, if the videos are part of an adaptive video-suite for students with different prior knowledge and skills (for example, after a student views a video, he/she can be asked several questions, and if he/she struggles to answer those questions, he/she can be directed to another explanation video that other students who answer those questions correctly can skip). In particular, the videos can provide more scaffolding support

as needed to a student who is struggling. Then, after taking full advantage of these out-of-class activities, the EBAE activities can help all students.

4.1.2 Gender Gap in Introductory Physics Courses

Prior research has found that male students outperform female students on standardized conceptual assessments such as the Force Concept Inventory (FCI) [48] or the Conceptual Survey of Electricity and Magnetism (CSEM) [49]. The discrepancy between male and female students' performance is typically referred to as a "gender gap" [50-52]. While sometimes gender gap can be accounted for at least in part due to different prior preparation or coursework of male and female students, it has also been found even after controlling for these factors [50]. Prior research has also found that using evidence-based pedagogies can reduce the gender gap [53-54], but the extent to which this occurs varies. Others have found that the gender gap is not reduced despite significant use of evidence-based pedagogies [55]. Prior research has also found a gender gap on other assessments such as a conceptual assessment for introductory laboratories [56] and physics exams [50-51]. Yet others have found no differences in performance between male and female students on exams [52, 57-58].

The origins of gender gap on the FCI both at the beginning and end of a physics course have been a subject of debate with some researchers arguing that the test itself is gender-biased [59]. Some of the origins of the gender gap are related to societal gender stereotypes [60-63] that keep accumulating from an early age. For example, research suggests that even six year old boys and girls have gendered views about smartness in favor of boys [63]. Such stereotypes can impact female students' self-efficacy [64, 65], their beliefs about their ability to perform well, in disciplines such as physics in which they are underrepresented and which have been associated

with “brilliance”. They can also impact their intelligence mindset [66], which is related to beliefs about whether intelligence innate or whether it is something that can be developed and cultivated via focus and persistence in problem solving in a discipline such as physics. Thus, it may not be surprising that prior research has found that activation of a stereotype, i.e., stereotype threat (ST) about a particular group in a test-taking situation can alter the performance of that group in a way consistent with the stereotype [60-63]. In fact, some researchers have argued [60] that female students, when working on a physics assessment, undergo an implicit ST due to the prevalent societal stereotypes. In particular, Marchand and Taasoobshirazi [60] conducted a study in which high school students were randomly divided into three groups and all students received the following instructions before taking a physics test: “You will be given four physics problems to solve. These problems are based on physics material that you have already covered.” In the implicit ST condition, these were the only instructions, while in the explicit ST condition, students were also told: “This test has shown gender differences with males outperforming females on the problems” and in the nullified condition, students were told: “No gender differences in performance have been found on the test”. They found no statistically significant difference on the physics test between female students’ performance in the explicit ST condition and the implicit ST condition but female students in both these conditions performed significantly worse than male students. In contrast, the nullified condition in which female students were instead told that the test they are about to take is gender neutral erased the gender gap (no difference in performance between male and female students). The researchers hypothesized [60] that simply administering a physics test to female students creates an implicit stereotype threat (which is partly due to societal gender bias and related issue of anxiety and self-efficacy, which refers to the fact that many female students start doubting their own ability to perform well in a physics test).

4.1.3 Focus of the Research

In this study, we used the FCI [48] in the first semester introductory physics courses and the CSEM [49] in the second semester courses to assess student learning. We also investigated any possible gender gap at the beginning of the course as well as the extent to which evidence-based pedagogies can help reduce it. The FCI, CSEM and other standardized physics surveys [67-72] have been used to assess introductory students' understanding of physics concepts by a variety of educators and physics education researchers. One reason for their extensive use is that many of the items on the survey have strong distractor choices which correspond to students' common difficulties so students are unlikely to answer the survey questions correctly without having good conceptual understanding. Our research focuses on the following research questions for both algebra-based and calculus-based introductory physics courses:

RQ1. What is the gender gap on the FCI/CSEM pretest and posttest in LB and EBAE courses? By how much do both male and female students improve from pretest to posttest in LB and EBAE courses?

RQ2. How does the performance on the FCI/CSEM of male and female students in LB courses compare to EBAE courses in both the pretest and the posttest?

RQ3. To what extent do male and female students with high or low pretest scores perform differently in EBAE courses compared to LB courses when the comparison is made for one instructor who teaches both an EBAE and an LB course at the same time?

RQ4. To what extent do male and female students with high or low pretest scores perform differently in EBAE courses compared to LB courses when the comparison is made between EBAE and LB courses taught by different instructors?

RQ5. Is there any correlation between posttest and final exam scores for male and female students?

Thus, in our research, the performances of male and female students in EBAE courses in a particular type of course (algebra-based or calculus-based physics I or II) are compared with male and female students of LB courses in two situations: (I) the same instructor taught two courses, one of which was an EBAE course and the other an LB course with common homework and final exams, (II) student performances in all of the EBAE courses taught by different instructors were averaged and compared with LB courses of the same type, also averaged over different instructors.

Also, the students were divided into three subgroups based upon their pretest scores: top 1/3rd, middle 1/3rd and bottom 1/3rd. We calculated whether there was a statistically significant difference between male and female students' average scores on the pretest, posttest or final exam in two cases: (i) male students were divided into three subgroups according to the pretest scores of males only and female students were also divided into three subgroups according to the pretest scores of females only, and then the male and female students' average scores in each subgroup were compared and (ii) all students were divided into the three subgroups according to their pretest scores ***regardless*** of their gender and then male and female students in each of the three subgroups were separated and compared. This type of analysis based upon gender was carried out for the male and female students taught by the same instructor (teaching either LB or EBAE course) and also for different instructors teaching LB or EBAE courses of the same type combined. Whenever differences between these two groups were observed (e.g., with male or female students in the EBAE courses on average performing better than the corresponding students in the LB courses), we investigated which subgroup was benefiting most from the EBAE courses, e.g., those who performed well or poorly on the pretest given at the beginning of the course. Finally, we

investigated the typical correlation between the performance of male and female students' posttest performances on the validated conceptual surveys and their performance on the instructor-developed final exam (which typically places a heavy weight on quantitative physics problems).

4.2 METHODOLOGY

4.2.1 Courses and Participants

The participants in this study were students in 16 different algebra-based and calculus-based introductory physics courses. Out of all introductory physics courses (algebra-based or calculus-based physics I or II) included in this study, there were four EBAE courses: two completely flipped classes in algebra-based introductory physics I and one completely flipped and one interactive active engagement class in calculus-based introductory physics II. These courses include approximately 700 male and 750 female students in first semester courses and approximately 650 male and 500 female students in second semester courses at a typical large research university in the US (University of Pittsburgh). The details of the courses that fall into three categories are as follows:

1) A lecture-based (or LB) course is one in which the primary mode of instruction was via lecture. In addition to the three or four weekly hours for lectures, students attended an hour long recitation section taught by a graduate TA. In recitation, the TA typically answered student questions (mainly about their homework problems which were mostly textbook style quantitative problems), solved problems on the board and gave students a quiz in the last 10-20 minutes.

2) A flipped course is one in which the class was broken up into two almost equal size groups with each group meeting with the instructor for half the regular class time. For example, for a 200 student class scheduled to meet for four hours each week (on two different days), the instructor met with half the class (100 students) on the first day and the other half on the second day. This was possible in the flipped classes since the total contact hours for each instructor each week with the students was the same as in the corresponding LB courses. Students watched the lecture videos before coming to class and answered some conceptual questions which were based upon the lecture video content. They uploaded the answers to those conceptual questions before class onto the course website and were scored for a small percentage of their grade (typically 4-8%). Although students had to watch several videos outside of class in preparation for each class, each video was typically 5-10 minutes long, followed by concept questions. On average, students in a flipped class had to watch recorded videos which took a little less than half the allotted weekly time for class (e.g., for the courses scheduled for four hours each week, students watched on average 1.5 hours of videos each week, and in the courses scheduled for three hours each week, students watched around one hour of videos). These video times do not include the time that students would take to rewind the video, stop and think about the concepts and answer the concept questions placed after the videos that counted towards their course grade. In the spirit of JiTT, the instructors of the flipped courses adjusted the in-class activities based upon student responses to online concept questions which were supposed to be submitted the night before the class. About 90% of the students submitted their answers to the concept questions that followed the videos to the course website before coming to the class. The web-platforms used for managing, hosting and sharing these videos and for having online

discussions with students about them asynchronously (in which students and the instructor participated) were Classroom Salon or Panopto. In-class time was used for clicker questions involving peer discussion and then a whole class discussion of the clicker questions, collaborative group problem solving involving quantitative problems in which 2-3 students worked in a group (followed by a clicker question about the order of magnitude of the answer), and lecture-demonstrations with preceding clicker questions on the same concepts. In addition to the regular class times, students attended an hour long recitation section which was taught the same way as for students in the LB courses.

It is important to note that the instructors who taught the flipped courses also taught LB courses at the same time (usually teaching two courses in a particular semester: one flipped and one LB). Students in both flipped and LB courses completed the same homework and took the same final exam. For the calculus-based flipped courses, the students also took the same midterm exams. This was not possible for the algebra-based courses because the exams were scheduled at different times. However, in the algebra-based courses they took the same final exam and had the same homework.

3) In an EBAE interactive non-flipped course, the instructor combined lectures with research-based pedagogies including clicker questions involving peer discussion, conceptual tutorials, collaborative group problem solving, and lecture demonstrations with preceding clicker questions on the same concepts, similar to the flipped courses. In addition, students attended a reformed recitation which primarily used context-rich problems to get students to engage in group problem solving or worked on research-based tutorials while being guided by a TA. The instructor ensured that the problems students solved each week in the recitation activities were closely related to what happened in class. Students also worked

on some research-based tutorials during class in small groups, but if they did not finish them in the allotted time, they were asked to complete them at home and submit as homework.

From now on, we refer to the flipped and interactive non-flipped courses as EBAE courses except when relevant. We also note that the number of female students in algebra-based courses is larger than that of male students. Most of the algebra-based students have biological science or related majors like Biology, Psychology, Exercise Science, Neurology/Neuromedicine, Environmental Science, etc. In calculus-based courses, on the other hand, there are more male than female students. Most calculus-based students are in their first year in college, and have physical science related majors such as chemistry, mathematics, engineering (electrical, mechanical, chemical, civil etc.), and physics (typically only 5-10 physics majors out of several hundred students). The algebra-based or calculus-based physics courses are mandatory for these students. We do not have information about the background of the students, such as their prior experiences in physics or mathematics before college or whether they took any physics or math courses in high school (although a majority of these students have typically taken at least one high school physics course and the typical percentage of female students in calculus-based "advanced placement C" high school courses in the US is less than one third).

We also note that none of the instructors teaching the EBAE courses focused explicitly on whether the active-learning classroom environment helped foster a sense of belonging or focused on improving self-efficacy and instilling a growth mindset in all students. In particular, the instructors did not *explicitly* focus on whether the active-learning classroom was a low anxiety classroom for all students and whether women and other underrepresented students felt supported and had the same level of engagement with the active-learning activities.

4.2.2 Materials

The materials used in this study are the FCI and CSEM conceptual multiple-choice (five choices for each question) standardized surveys, which were administered in the first week of classes before instruction in relevant concepts (pretest) and after instruction in relevant concepts (posttest). The FCI was used in the first semester courses and the CSEM was used in the second semester courses. Apart from the data on these surveys that the researchers collected from all of these courses, each instructor administered his/her own final exam, which was mostly quantitative (60%-90% of the questions were quantitative, although some instructors had either the entire final exam or part of it in a multiple-choice format with five options for each question to make grading easier). Ten course instructors (who also provided the FCI or CSEM data from their classes) provided their students' final exam scores and most of them also provided a copy of their final exams.

4.2.3 Methods

Our main goals in this research were to compare the average performances of male and female students in introductory physics courses in different types of classes (e.g., Algebra-based or Calculus-based, EBAE or LB) and to compare male and female students' performances between courses that used EBAE pedagogies with the performances of students in LB courses by using standardized conceptual surveys, the FCI (for physics I) and CSEM (for physics II) as pre/posttests. We not only calculated the average gain (posttest - pretest scores) for each group for males and females but also calculated the average normalized gain, which is commonly used to determine how much students learned from pretest to posttest taking into account their initial scores on the pretest, to find out whether the gender gap increased, decreased or remained the

same. The normalized gain is defined as $\langle g \rangle = \frac{\%\langle S_f \rangle - \%\langle S_i \rangle}{100 - \%\langle S_i \rangle}$, in which $\langle S_f \rangle$ and $\langle S_i \rangle$ are the final (post) and initial (pre) class averages, respectively. Then, $\text{Norm}\, g = 100\langle g \rangle$ in percent [16]. This normalized gain provides valuable information about how much students have learned by taking into account what they already know based on the pretest. We wanted to investigate whether the normalized gain is higher in one course compared to another, and whether it is the same or different for males and females.

In order to compare EBAE courses with LB courses, we performed *t*-tests [73] on FCI or CSEM pre and posttest data for males and females. We also calculated the effect size in the form of Cohen's *d* defined defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma_{pooled}}$, where μ_1 and μ_2 are the averages of the two groups being compared (e.g., EBAE vs. LB or male vs. female) and $\sigma_{pooled} = \sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}$; here σ_1 and σ_2 are the standard deviations of the two groups being compared. We considered: $d < 0.5$ as small effect size, $0.5 \leq d < 0.8$ as medium effect size and $d \geq 0.8$ as large effect size, as described in [74].

Moreover, although we did not have control over the type of final exam each instructor used in his/her courses, we wanted to look for correlations between the FCI/CSEM posttest performance and the final exam performance for different instructors in the algebra-based and calculus-based EBAE or LB courses for male and female students separately. Including both the algebra-based and calculus-based courses, 10 instructors provided the final exam scores for their classes. We used these data to obtain linear regression plots between the posttest and the final exam performance for males, females and all students (combined) for each instructor and computed the correlation coefficient between the performance of students (male/female/all) on the

validated conceptual surveys and their performance on the final exam for different instructors. These correlation coefficients between the conceptual surveys and the final exam (with strong focus on quantitative problem solving) can provide an indication of the strength of the correlation between conceptual and quantitative problem solving of male and female students in these courses.

4.3 RESULTS

4.3.1 Comparison of the Gender Gap on the FCI/CSEM Pretest and Posttest in LB and EBAE Courses (RQ1)

4.3.1.1 Physics I

In Table 4.1, we present intra-group pre/posttest data (pooled data for the same type of courses) of male and female students on the FCI for the calculus-based and algebra-based introductory physics I courses. For the algebra-based courses, some were EBAE courses while others were LB courses, whereas all the calculus-based courses were LB. We found statistically significant improvements from the pretest to the posttest for each group (for both male and female students) in both LB and EBAE courses. However, both female and male students exhibited larger normalized gains in the EBAE courses. In the calculus-based LB course, the gender gap increased slightly from 13% to 17%, whereas in the algebra-based courses, the gender gap stayed roughly the same (varied between 11% and 13%) both in LB and EBAE courses. In both the pretest and the posttest in calculus-based and algebra-based courses, the difference in performance between male and female students was statistically significant and the effect sizes were typically in the medium range. Thus,

it appears that in algebra-based courses, using evidence-based pedagogies helped both female and male students learn more, but did not result in a reduction of the gender gap.

Table 4.1. Intra-group FCI pre/posttest averages (Mean) and standard deviations (SD) for first-semester introductory male and female students in calculus-based LB courses, and algebra-based EBAE and LB courses. The number of students in each group, N, is shown. For each group, a *p*-value obtained using a *t*-test shows that the difference between the pre/posttest is statistically significant and the difference between the male and female students is also statistically significant. The normalized gain (Norm g) from pretest to posttest and the effect size (Eff. size) shows how much male and female students learned from what they did not already know based on the pretest.

Type of Class	FCI	Female	Gender Comparison	Male
Calc LB	Pretest	N: 146 Mean: 43% SD: 20%	p-value: <0.001 ← **gender gap: 13%** → Eff. size: 0.635	N: 283 Mean: 55% SD: 20%
	Pretest vs. Posttest Comparison	↑ p-value: <0.001 Eff. size: 0.468 Norm g: 16% ↓		↑ p-value: <0.001 Eff. size: 0.686 Norm g: 30% ↓
	Posttest	N: 114 Mean: 52% SD: 20%	p-value: <0.001 ← **gender gap: 17%** → Eff. size: 0.868	N: 200 Mean: 68% SD: 19%
Alg EBAE	Pretest	N: 153 Mean: 30% SD: 15%	p-value: <0.001 ← **gender gap: 13%** → Eff. size: 0.749	N: 105 Mean: 43% SD: 20%
	Pretest vs. Posttest Comparison	↑ p-value: <0.001 Eff. size: 1.176 Norm g: 28% ↓		↑ p-value: <0.001 Eff. size: 0.955 Norm g: 32% ↓
	Posttest	N: 149 Mean: 49%	p-value: <0.001 ← **gender gap: 12%** →	N: 106 Mean: 61%

		SD: 18%	Eff. size: 0.653	SD: 19%
Alg LB	Pretest	N: 456 Mean: 30% SD: 14%	p-value: <0.001 ← **gender gap: 11%** → Eff. size: 0.691	N: 318 Mean: 41% SD: 18%
	Pretest vs. Posttest Comparison	↑ p-value: <0.001 Eff. size: 0.930 Norm g: 21% ↓		↑ p-value: <0.001 Eff. size: 0.863 Norm g: 28% ↓
	Posttest	N: 383 Mean: 44% SD: 18%	p-value: <0.001 ← **gender gap: 13%** → Eff. size: 0.686	N: 255 Mean: 57% SD: 20%

4.3.1.2 Physics II

In Table 4.2, we present intra-group (pooled data for the same type of courses) pre/posttest data for male and female students on the CSEM survey for algebra-based and calculus-based introductory physics II courses. Similar to the data shown in Table 4.1, we found statistically significant improvements on the CSEM for female and male students both in LB and EBAE courses, however, the learning gains for both female and male students were larger in EBAE courses. With regards to the gender gap, we found that in LB courses it stayed roughly the same (4% on pretest and 6% on posttest for calculus-based courses, 6% on the pretest and 8% on posttest for algebra-based courses). However, in the EBAE calculus-based course, the gender gap increased slightly from 4% to 10%, and it appears that male students may have benefited more from evidence-based pedagogies than female students (normalized gain for male students was 39% in EBAE courses compared to 29% for female students).

Table 4.2. Intra-group CSEM pre/posttest averages (Mean) and standard deviations (SD) for second-semester introductory male and female students in calculus-based LB and EBAE courses and algebra-based LB courses. The

total number of students in each group, N, is shown. For each group, a *p*-value obtained using a *t*-test shows that the difference between the pre/posttest is statistically significant and the difference between the male and female students is also statistically significant. The normalized gain (Norm g) from pretest to posttest and the effect size (Eff. size) shows how much male and female students learned from what they did not already know based on the pretest.

Type of Class	CSEM	Female	Gender Comparison	Male
Calc LB	Pretest	N: 84 Mean: 34% SD: 13%	p-value: 0.007 ← **gender gap: 4%** → Eff. size: 0.349	N: 234 Mean: 38% SD: 13%
	Pretest vs. Posttest Comparison	↑ p-value: <0.001 Eff. size: 0.821 Norm g: 18% ↓		↑ p-value: <0.001 Eff. size: 0.894 Norm g: 22% ↓
	Posttest	N: 78 Mean: 45% SD: 16%	p-value: 0.003 ← **gender gap: 6%** → Eff. size: 0.381	N: 248 Mean: 51% SD: 17%
Calc EBAE	Pretest	N: 112 Mean: 35% SD: 14%	p-value: 0.017 ← **gender gap: 4%** → Eff. size: 0.272	N: 220 Mean: 39% SD: 16%
	Pretest vs. Posttest Comparison	↑ p-value: <0.001 Eff. size: 1.143 Norm g: 28% ↓		↑ p-value: <0.001 Eff. size: 1.384 Norm g: 39% ↓
	Posttest	N: 98 Mean: 53% SD: 18%	p-value: <0.001 ← **gender gap: 10%** → Eff. size: 0.538	N: 193 Mean: 63% SD: 19%
Alg LB	Pretest	N: 301 Mean: 22% SD: 8%	p-value: <0.001 ← **gender gap: 6%** → Eff. size: 0.452	N: 201 Mean: 27% SD: 13%

	Pretest vs. Posttest Comparison	↑ p-value: <0.001 Eff. size: 1.450 Norm g: 23% ↓		↑ p-value: <0.001 Eff. size: 1.325 Norm g: 29% ↓
	Posttest	N: 266 Mean: 40% SD: 16%	p-value: <0.001 ← **gender gap: 8%** → Eff. size: 0.451	N: 172 Mean: 48% SD: 18%

4.3.2 Comparison of the Performance of Male and Female Students on the FCI/CSEM in LB and EBAE Courses in Pretest and Posttest (RQ2)

4.3.2.1 Physics I

Table 4.3 shows the between-course male and female student FCI pre/posttest score comparison between algebra-based LB and EBAE courses, first holding the instructor fixed (same instructor taught both the LB and EBAE courses, used the same homework and final exams) and second, combining all instructors who used similar methods in the same group (only one instructor used EBAE methods, but several who taught LB courses were combined). Table 4.3 shows that on the pretest, the performance of male and female students in the LB courses was similar to the EBAE courses. However, on the posttest both male (female) students in the EBAE courses outperformed male (female) students in the LB courses (effect sizes ranging from 0.196 to 0.324). Coupled with the gender gaps shown in Table 4.1, these data suggest that while both female and male students learned more in EBAE courses, the gender gap remained roughly the same in EBAE courses as in LB courses.

Table 4.3. Between-course comparison of the average FCI pre/posttest scores of algebra-based male and female students in LB courses with EBAE courses when (i) both courses are taught by the same instructor and (ii) different instructors using similar instructional methods are combined. The p-values and effect sizes are obtained for male and female students separately when comparing the LB and EBAE courses in terms of students' FCI scores.

	FCI	Female-Pretest	Female-Posttest	Male-Pretest	Male-Posttest
(i) FCI Alg: LB vs. EBAE (same instructor)	LB	N: 260 Mean: 30% SD: 13%	N: 246 Mean: 43% SD: 18%	N: 166 Mean: 42% SD: 19%	N: 154 Mean: 56% SD: 21%
	LB vs. EBAE Comparison	p-value: 0.846 Eff. size: 0.020	p-value: 0.002 Eff. size: 0.324	p-value: 0.786 Eff. size: 0.034	p-value: 0.041 Eff. size: 0.258
	EBAE	N: 153 Mean: 30% SD: 15%	N: 149 Mean: 49% SD: 18%	N: 105 Mean: 43% SD: 20%	N: 106 Mean: 61% SD: 19%
(ii) FCI Alg: LB vs. EBAE (different instructors combined)	LB	N: 456 Mean: 30% SD: 14%	N: 383 Mean: 44% SD: 18%	N: 318 Mean: 41% SD: 18%	N: 255 Mean: 57% SD: 20%
	LB vs. EBAE Comparison	p-value: 0.861 Eff. size: 0.017	p-value: 0.009 Eff. size: 0.255	p-value: 0.432 Eff. size: 0.091	p-value: 0.088 Eff. size: 0.196
	EBAE	N: 153 Mean: 30% SD: 15%	N: 149 Mean: 49% SD: 18%	N: 105 Mean: 43% SD: 20%	N: 106 Mean: 61% SD: 19%

4.3.2.2 Physics II

Table 4.4 shows the between-course CSEM pre/posttest score comparison between calculus-based LB and EBAE courses, first holding the instructor fixed (same instructor taught both the LB and EBAE courses and used the same homework and final exams) and second, combining all

instructors who taught using similar methods into the same group. Table 4.4 shows that on the pretest, the performance of male and female students in the LB courses was similar to the EBAE courses. However, on the posttest, both male (female) students in the EBAE courses outperformed male (female) students in the LB courses (effect sizes ranging from 0.299 to 0.623). Interestingly, the effect sizes for male students were slightly higher than the effect sizes for female students, suggesting that male students may have benefited more from evidence-based pedagogies. The gender gap data shown in Table 4.2 can be interpreted in a similar manner. Thus, our data suggest that in calculus-based physics II, while both female and male students learned more in EBAE courses, male students may have benefited more than female students, resulting in a slight increase in the gender gap from pretest to posttest in EBAE courses.

Table 4.4. Between-course comparison of the average CSEM pre/posttest scores of calculus-based male and female students in LB courses with EBAE courses when (i) both courses are taught by the same instructor and (ii) different instructors using similar instructional methods are combined. The p-values and effect sizes are obtained for male and female students separately when comparing the LB and EBAE courses in terms of students' CSEM scores.

	CSEM	Female-Pretest	Female-Posttest	Male-Pretest	Male-Posttest
(i) CSEM Calc: LB vs. EBAE (same instructor)	LB	N: 51 Mean: 35% SD: 12%	N: 44 Mean: 44% SD: 15%	N: 126 Mean: 42% SD: 12%	N: 110 Mean: 50% SD: 16%
	LB vs. EBAE Comparison	p-value: 0.590 Eff. size: 0.097	p-value: 0.119 Eff. size: 0.299	p-value: 0.845 Eff. size: 0.024	p-value: 0.001 Eff. size: 0.455
	EBAE	N: 75 Mean: 36% SD: 14%	N: 68 Mean: 48% SD: 17%	N: 133 Mean: 42% SD: 15%	N: 113 Mean: 58% SD: 20%
(ii) CSEM Calc LB vs. EBAE	LB	N: 84 Mean: 34%	N: 78 Mean: 45%	N: 234 Mean: 38%	N: 248 Mean: 51%

(different instructors combined)		SD: 13%	SD: 16%	SD: 13%	SD: 17%
	LB vs. EBAE Comparison	p-value: 0.595 Eff. size: 0.077	p-value: 0.003 Eff. size: 0.448	p-value: 0.679 Eff. size: 0.039	p-value: <0.001 Eff. size: 0.623
	EBAE	N: 112 Mean: 35% SD: 14%	N: 98 Mean: 53% SD: 18%	N: 220 Mean: 39% SD: 16%	N: 193 Mean: 63% SD: 19%

4.3.3 Comparison between EBAE and LB Courses Taught by the Same Instructor in terms of Male and Female students' Performances, Divided according to Pretest Scores (RQ3)

Tables 4.5 and 4.6 show the average algebra-based FCI and calculus-based CSEM pretest, posttest, gain, normalized gain and final exam scores for male and female students, along with the p-values between each subgroup of male and female students for pretest, posttest and the final exam in the EBAE and LB courses taught by the same instructor (with the same homework and final exam) with students divided into three subgroups based on their pretest scores. The male students were divided into three subgroups according to the pretest scores of male students only and female students were also divided into three subgroups according to the pretest scores of female students only. (Tables 4.10 and 4.11 in the chapter appendix show similar type of information as Tables 4.5 and 4.6 except that the total number of students was divided into three subgroups according to their pretest scores regardless of their gender and then male and female students were separated for comparison.)

4.3.3.1 Algebra-based Physics I

The data in Table 4.5 show that in both the LB and EBAE algebra-based courses, on the pretest, there was a gender gap on the FCI between male and female students in each group (bottom 1/3, middle 1/3, top 1/3) and since female and male students had comparable gains, the gender gap was maintained in most cases on the posttest. This is consistent with the data shown in Table 4.1 which indicate that in both LB and EBAE courses, when including all students, the gender gap on the FCI stayed roughly the same from pretest to posttest. When comparing the LB with the EBAE course, we see that both female and male students seemed to benefit equally (gains and normalized gains on the FCI were higher in the EBAE course compared to the LB course), with the exception of the top 1/3 of the students. The top 1/3 of the female students had similar FCI gains in the LB and EBAE course (13% and 14%, respectively), whereas the top 1/3 of the male students showed larger FCI gains in the EBAE course.

Table 4.5. Average FCI pretest scores (Pretest), posttest scores (Posttest), gain (Gain), normalized gain (Norm g) and final exam scores (Final) for male and female students in the EBAE and LB courses taught by the same instructor (with same homework and final exam). Male students were divided into three groups based upon their pretest scores and female students were also divided into three groups based upon their pretest scores separately. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest, posttest and final exam.

	Pretest Split		Pretest	Posttest	Gain	Norm g	Final
FCI Alg LB	bottom 1/3	Mean Female Score	17	35	18	21	47
		p-value	0.010	0.078			0.815
		Mean Male Score	23	42	19	25	48
	middle 1/3	Mean Female Score	28	42	14	19	53
		p-value	<0.001	0.012			0.078
		Mean Male Score	39	55	16	26	60

	top 1/3	Mean Female Score	44	57	13	23	60
		p-value	<0.001	0.002			0.422
		Mean Male Score	65	74	8	24	64
FCI Alg EBAE	bottom 1/3	Mean Female Score	15	39	23	27	54
		p-value	<0.001	0.164			0.318
		Mean Male Score	21	43	22	28	50
	middle 1/3	Mean Female Score	28	46	17	24	52
		p-value	<0.001	<0.001			0.008
		Mean Male Score	38	58	21	33	61
	top 1/3	Mean Female Score	49	63	14	28	60
		p-value	<0.001	<0.001			0.011
		Mean Male Score	63	78	15	41	68

On the final exam, the data in Table 4.5 suggest that in the EBAE class, the top and middle 1/3 of the male students performed better than the top and middle 1/3 of the female students. Similar, although not as strong, trends can be seen in the LB course. Since students in these two courses took the same final exams, these data suggest that the top and middle 1/3 of the male students benefited slightly more from EBAE pedagogies than top and middle 1/3 of the female students. On the other hand, the bottom 1/3 of the female students benefited more from EBAE pedagogies than the bottom 1/3 of the male students (performance of bottom 1/3 of female students is 54% in the EBAE course and only 47% in the LB course, whereas the performance of the bottom 1/3 of male students is 50% in the EBAE course and 48% in the LB course).

4.3.3.2 Calculus-based Physics II

The data in Table 4.6 indicate that in the LB course, there was a gender gap on the CSEM in the pretest, but on the posttest, this gender gap decreased and was no longer statistically significant. This result appears to be inconsistent with the data shown in Table 4.2 which indicate that in the

LB courses, the CSEM gender gap remained roughly the same (or slightly increased). However, the data in Table 4.2 includes all LB courses, whereas the data in Table 4.6 includes only one course which was taught by the same instructor who taught the EBAE course shown in Table 4.6. So in this particular calculus-based LB course, the gender gap on the CSEM decreased slightly, but if we include all calculus-based LB courses, the gender gap was roughly the same (or increased slightly).

Table 4.6. Average CSEM pretest scores (Pretest), posttest scores (Posttest), gain (Gain), normalized gain (Norm g) and final exam scores (Final) for male and female students in the EBAE and LB courses taught by the same instructor (with same homework and final exam). Male students were divided into three groups based upon their pretest scores and female students were also divided into three groups based upon their pretest scores separately. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest, posttest and final exam.

	Pretest Split		Pretest	Posttest	Gain	Norm g	Final
CSEM Calc LB	bottom 1/3	Mean Female Score	24	36	12	16	42
		p-value	0.006	0.738			0.209
		Mean Male Score	28	38	10	13	47
	middle 1/3	Mean Female Score	34	44	10	16	52
		p-value	<0.001	0.661			0.770
		Mean Male Score	40	46	6	10	50
	top 1/3	Mean Female Score	49	56	8	15	53
		p-value	0.060	0.513			0.018
		Mean Male Score	54	59	5	12	62
CSEM Calc EBAE	bottom 1/3	Mean Female Score	22	36	14	18	48
		p-value	0.002	0.050			0.196
		Mean Male Score	27	44	18	24	53
	middle 1/3	Mean Female Score	34	48	14	22	52
		p-value	<0.001	0.455			0.104
		Mean Male Score	41	51	10	18	58

		Mean Female Score	56	63	7	16	68
	top 1/3	p-value	0.382	0.003			0.621
		Mean Male Score	58	74	16	38	70

For the calculus-based EBAE course, the data in Table 4.6 suggest that the gender gap increased. The gender gaps for bottom 1/3, middle 1/3 and top 1/3 of the students are 5%, 7% and 2% in the pretest but 8%, 3% and 11% in the posttest, respectively. Thus, with the exception of the middle 1/3 of the students, the gender gap on the CSEM increased. This suggests that the bottom 1/3 and top 1/3 of the male students may have benefited more from EBAE pedagogies compared to the respective female students. It appears that this was indeed the case when we compare the normalized gains in the LB and EBAE courses: for the bottom 1/3 and top 1/3 of the male students, their CSEM normalized gains were 13% and 12% in the LB course, but 24% and 38% in the EBAE course, respectively. For the bottom 1/3 and top 1/3 of the female students, their CSEM normalized gains were 16% and 15% in the LB course, and 18% and 16% in the EBAE course. On the final exam, in both the LB and EBAE course, it appears that male students performed slightly better than the female students. However, only the 9% gender gap between the top 1/3 of the male and female students in the LB course is statistically significant.

4.3.4 Comparison between EBAE and LB Courses Taught by Different Instructors in terms of Male and Female students' Performances, Divided according to Pretest Scores (RQ4)

Tables 4.7 and 4.8 show the average algebra-based and calculus based FCI and CSEM pretest score, posttest score, gain and normalized gain for male and female students, along with the p-values between each subgroup of male and female students for the pretest and posttest in the EBAE

and LB courses. All equivalent (algebra-based or calculus-based physics I or II) courses which used the same instructional strategy (EBAE or LB) were combined and students were divided into three groups based upon their pretest scores. Male students were divided into three subgroups according to the pretest scores of male students only and female students were also divided into three subgroups according to the pretest scores of female students only, and their scores were compared (cases in which male and female scores are significantly different have been highlighted). We note that Tables 4.12 and 4.13 in the chapter appendix show the same data except that the students were divided into three subgroups according to their pretest scores regardless of their gender. Then, male and female students were separated for comparison and the cases in which male and female scores are significantly different from each other have been highlighted. In these tables (4.7, 4.8, 4.12 and 4.13), the average final exam performance is not listed because different instructors used different exams which varied in difficulty.

Table 4.7. Average FCI pretest scores (Pretest), posttest scores (Posttest), gain (Gain) and normalized gain (Norm g) for male and female students in the EBAE and LB algebra-based and calculus-based courses. All courses in the same group were combined. Male students were divided into three groups based upon their pretest scores and female students were also divided into three groups based upon their pretest scores separately. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest and posttest. ***Note that FCI data for calculus-based EBAE classes are not available.***

	Pretest Split		Pretest	Posttest	Gain	Norm g
FCI Calc LB	bottom 1/3	Mean Female Score	24	39	15	19
		p-value	<0.001	<0.001		
		Mean Male Score	35	52	17	26
	middle 1/3	Mean Female Score	43	52	9	16
		p-value	<0.001	<0.001		
		Mean Male Score	60	75	14	36
	top 1/3	Mean Female Score	67	72	6	17

		p-value	<0.001	0.023		
		Mean Male Score	85	82	-3	-18
FCI Alg EBAE	bottom 1/3	Mean Female Score	15	39	23	27
		p-value	<0.001	0.164		
		Mean Male Score	21	43	22	28
	middle 1/3	Mean Female Score	28	46	17	24
		p-value	<0.001	<0.001		
		Mean Male Score	38	58	21	33
	top 1/3	Mean Female Score	49	63	14	28
		p-value	<0.001	<0.001		
		Mean Male Score	63	78	15	41
FCI Alg LB	bottom 1/3	Mean Female Score	17	33	15	18
		p-value	<0.001	<0.001		
		Mean Male Score	23	41	18	23
	middle 1/3	Mean Female Score	29	44	15	22
		p-value	<0.001	<0.001		
		Mean Male Score	40	55	15	24
	top 1/3	Mean Female Score	44	57	13	23
		p-value	<0.001	<0.001		
		Mean Male Score	62	76	14	36

4.3.4.1 Physics I

For the calculus-based LB course, the data in Table 4.7 suggest that the gender gap on the FCI increased from pretest to posttest for each ability level. The gender gap for the bottom, middle, and top 1/3 of the students was 11%, 17%, 18% in the pretest, but in the posttest, it was 13%, 23%, 10%, respectively. This is consistent with the data in Table 4.1 which indicates that, including all students, the FCI gender gap increased slightly from the pretest to the posttest except the top 1/3 group.

Table 4.8. Average CSEM pretest scores (Pretest), posttest scores (Posttest), gain (Gain) and normalized gain (Norm g) for male and female students in the EBAE and LB algebra-based and calculus-based courses. All courses in the same group were combined. Male students were divided into three groups based upon their pretest scores and female students were also divided into three groups based upon their pretest scores separately. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest and posttest. ***Note that CSEM data for algebra-based EBAE classes are not available.***

	Pretest Split		Pretest	Posttest	Gain	Norm g
CSEM Calc EBAE	bottom 1/3	Mean Female Score	21	43	22	28
		p-value	0.044	<0.001		
		Mean Male Score	23	57	35	45
	middle 1/3	Mean Female Score	32	52	20	29
		p-value	<0.001	0.403		
		Mean Male Score	37	55	18	29
	top 1/3	Mean Female Score	51	67	15	31
		p-value	0.039	0.017		
		Mean Male Score	56	74	18	41
CSEM Calc LB	bottom 1/3	Mean Female Score	21	37	16	20
		p-value	0.001	0.429		
		Mean Male Score	25	40	15	20
	middle 1/3	Mean Female Score	32	43	12	17
		p-value	<0.001	0.030		
		Mean Male Score	37	49	12	20
	top 1/3	Mean Female Score	47	60	13	24
		p-value	0.025	0.744		
		Mean Male Score	52	57	6	12
CSEM Alg LB	bottom 1/3	Mean Female Score	15	34	20	23
		p-value	0.017	0.100		
		Mean Male Score	16	39	23	27
	middle 1/3	Mean Female Score	22	40	19	24

		p-value	<0.001	0.009		
		Mean Male Score	25	47	22	30
		Mean Female Score	31	47	16	23
	top 1/3	p-value	<0.001	<0.001		
		Mean Male Score	40	59	19	31

For the algebra-based LB and EBAE courses, the data in Table 4.7 suggest that the gender gap on the FCI was present at each ability level in the pretest and it remained roughly the same in the posttest (consistent with the data in Table 4.1). For the LB courses, the gender gap on the FCI for bottom, middle, top 1/3 of the students was 6%, 11%, 18% on the pretest and 7%, 11%, 19% on the posttest. For the EBAE courses, the gender gap for bottom, middle, top 1/3 of the students was 6%, 10%, 14% on the pretest and 4%, 12%, 15% on the posttest. Interestingly, in both type of courses, it appears that the gender gap on the FCI was more pronounced at higher ability levels (based on FCI pretest scores). This was especially true in the LB course where the performance of the top 1/3 of the male students was on the average 18% (19%) higher compared to the top 1/3 of the female students on the pretest (posttest). The data in Table 4.7 suggest that both female and male students learned more in the EBAE course compared to the LB, but the learning gains were not much larger in the EBAE course compared to the LB course.

4.3.4.2 Physics II

The data in Table 4.8 suggest that in the calculus-based EBAE courses, the gender gap on the CSEM increased, but only for the bottom 1/3 of the students. On the pretest, the gender gap between the bottom 1/3 of the male and female students was 2%, whereas in the posttest, the gender gap was 14%. This suggests that the bottom 1/3 male students benefited much more from EBAE pedagogies than the bottom 1/3 of the female students. For the middle and top 1/3 of the students,

the gender gap remained roughly the same. By comparison, in the calculus-based LB courses, the gender gap stayed roughly the same. The gender gap between bottom 1/3, middle 1/3, top 1/3 of the students was 4%, 5%, 5% in the pretest and 3%, 6%, -3% in the posttest. These findings are consistent with the data shown in Table 4.2, which indicate that the gender gap on the CSEM stayed roughly the same in the LB courses, but increased slightly in the EBAE courses.

Comparing the LB with the EBAE courses in terms of normalized gain, the data in Table 4.8 suggest that students at all levels benefit from EBAE pedagogies. However, it appears that the bottom 1/3 and top 1/3 of the male students benefited more from EBAE pedagogies compared to the corresponding female students. The normalized CSEM gains for the bottom 1/3 and top 1/3 of the male students were 20% and 12% in the LB course but the EBAE course they were much larger at 45% and 41%. For the bottom 1/3 and top 1/3 of the female students on the other hand, the normalized CSEM gains were 20% and 24% in the LB course but only slightly larger at 28% and 31% in the EBAE course. A very similar trend was observed in Table 4.6. Thus, it appears that for calculus-based physics II, the bottom 1/3 and top 1/3 of the male students may have benefited more from EBAE pedagogies compared to the bottom 1/3 and top 1/3 of the female students.

Similar to the calculus-based LB courses, for the algebra-based LB courses, the gender gap stayed roughly the same for students in the bottom, middle and top 1/3 of the class (based on CSEM pretest scores).

4.3.5 Correlation between CSEM Posttest and Final Exam Scores for Male and Female Students (RQ5)

Figure 4.1 plots the CSEM posttest performance along with the final exam performance for male, female and all students in calculus-based EBAE course. Figure 4.1 shows that the linear

regressions [73] and there are moderate to strong correlation between posttest and the final exam scores. We also plotted linear regressions for the other courses and the data look similar to Figure 4.1 but are not included here. Instead, we include all the correlation coefficients (CSEM posttest vs. final exam) for all the courses for which we were able to obtain both the posttest and final exam data. Table 4.9 summarizes the correlation coefficients between CSEM/FCI posttest and final exam scores for each instructor who provided final exam data.

Despite the fact that different instructors had different final exams and at least some of the content in the final exam does not match the FCI or CSEM tests (e.g., in physics I, many topics were covered which are not on the FCI, such as momentum and collisions, static equilibrium and rotations, fluid dynamics and others), the correlation coefficients including males and females range from 0.415 to 0.816, which are considered to be moderate to high correlations. Although there are some differences between the correlation coefficients for male and female students in a given course, there is no clear discernable trend.

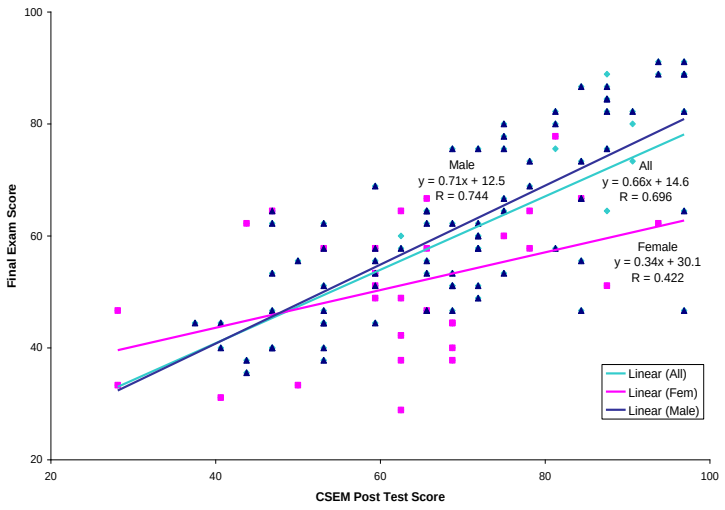

Figure 4.1. Linear regression and correlation coefficients of the CSEM posttest scores (conceptual) and final exam scores (heavy focus on quantitative problem solving) for male students, female students and all students (males and

females combined together) for Inst 1 (EBAE instructor) for calculus-based introductory physics courses. The correlation coefficients for other FCI/CSEM instructors have been summarized in Table 4.9.

Table 4.9. Correlation coefficients (R) between CSEM/FCI posttest and final exam scores of male and female students for each instructor (Inst) who provided final exam data. The final exam data were not provided by physics II instructors in algebra-based courses.

Physics I (Calc)			**Physics I (Alg)**			**Physics II (Calc)**		
Inst and course	Female	Male	Inst and course	Female	Male	Inst and course	Female	Male
Inst 1: LB	0.427	0.525	Inst 1: EBAE	0.568	0.628	Inst 1: EBAE	0.422	0.744
Inst 2: LB	0.674	0.573	Inst 1: LB	0.510	0.554	Inst 2: EBAE	0.571	0.415
Inst 3: LB	0.774	0.816	Inst 2: LB	0.626	0.728	Inst 2: LB	0.553	0.540
						Inst 3: LB	0.230	0.592

4.4 DISCUSSION AND SUMMARY

4.4.1 General Findings for EBAE and LB Courses Regardless of Gender

In all cases investigated, we find that on average, introductory students in the courses which made significant use of EBAE methods outperformed those in courses primarily taught using LB instruction on standardized conceptual surveys (FCI or CSEM) on the posttest even though there was no statistically significant difference on the pretest. This was true both in the algebra-based and calculus-based physics I (primarily mechanics) and II (primarily E&M) courses. Also, the differences between EBAE and LB courses were observed both among students who performed well on the FCI/CSEM pretest (given in the first week of classes) and also those who performed

poorly, thus indicating that EBAE instructional strategies helped students at all levels. However, the typical effect sizes for the differences between equivalent EBAE and LB courses was between 0.23-0.49, which are small. Thus, the benefits of these EBAE approaches were not as large as one might expect to observe. There are many potential challenges to using EBAE instructional strategies effectively, including but not limited to:

- Content coverage. There is often a lot of content covered in introductory physics courses and it is challenging to cover the same amount of content while also including frequent active learning activities during class and students are expected to take responsibility for learning some those things outside of classes.
- Lack of student buy-in of EBAE pedagogies, which may result in lack of appropriate engagement with self-paced learning tools outside of class. It is therefore important for instructors to frame the course for students and discuss the various instructional approaches that will be used in a course and why they are expected to be beneficial for student learning. Providing data to the students that support the use of evidence-based active learning strategies [47] can be helpful, and when possible, including explicit discussions connecting students' and instructors' goals for taking the course can also be beneficial [75].
- Lack of student engagement with in-class active learning activities (e.g., clicker questions and group problem solving). Students may not recognize on their own that they will learn best if they engage with the in-class activities to the best of their ability. Therefore, ensuring that in-class activities help all students learn is important. Furthermore, since peer collaboration is exploited in many EBAE classes to enhance student learning, ensuring that these activities are designed and incentivized in a manner that not only fosters positive inter-dependence (success of one student is contingent on the success of the group) but also

individual accountability (students are expected to show that they learned from working in a group) is essential [4-5].

- Large class sizes may be an impediment. One approach faculty used in flipped courses was to split the class in two (the instructor met with each group for only half of the time as compared to an LB course, his/her total contact hours with students remained the same), thus forming smaller class sizes. But even if the class size goes from 200 to 100 students by this process of breaking the class into two halves, it may still be challenging to manage the in-class activities effectively. Undergraduate or graduate teaching assistants need to be trained to effectively help in facilitating in-class activities. In group activities, students often work at different rates, so effective approaches need to be adopted to ensure that those who finish early can help others.

4.4.2 Impact on Gender Gap

We found that the EBAE courses did not result in reducing the gender gap. For algebra-based courses, students at all levels learned more in the EBAE courses; however, it appears that both female and male students benefited from evidence-based pedagogies equally and the gender gap present in the pretest was also found on the posttest. For calculus-based courses, our data suggest that male students actually benefited more from evidence-based pedagogies, which resulted in an increase in the gender gap from the pretest to the posttest. One hypothesis for why the gender gap was steady in algebra-based courses but grew in the calculus-based courses is that in the calculus-based courses there are significantly fewer women which can impact their sense of belonging and self-efficacy. These issues were not investigated in this study.

Previous research has also found that sometimes evidence-based pedagogies result in a reduction of the gender gap [53-54], while in other cases they do not [55]. The reasons for the gender gap even in the pretest are complex and some have attributed the persistence of gender gap to issues such as societal gender stereotypes, stereotype threat, high anxiety classes, lack of social belonging for women in physics classes, the culture promoting fixed intelligence mindset (with men having the innate ability to excel in subjects such as physics), and low self-efficacy [59-66].

Some have suggested that the gender gap found on conceptual assessments may at least in part be due to stereotype threat [57, 60-63], and the extent to which the classroom environment is perceived as threatening by female students which in turn can depend on the instructor and the instructional design. For example, as discussed earlier, research suggests that, for high school female students, taking a physics test can create an 'implicit stereotype threat' and can degrade their performance [60]. Such threat may be present even when taking the FCI or CSEM test at the beginning of the semester as a pretest and lead to a gender gap in performance. Also, female students may have a lower sense of social belonging and low self-efficacy in a physics class due to societal stereotypes about who belongs in physics and who is capable of doing physics.

Some have suggested that classes which are not only collaborative but also emphasize collaboration and reduce competition (e.g., by not grading on a curve) are likely to be perceived more positively by female students and may partly be responsible for the reduced gender gap in Ref. [53]. Even EBAE classrooms can be characterized as high or low anxiety classes depending on the extent to which the instruction was designed to be inclusive and whether it explicitly focused on promoting a sense of belongingness, self-efficacy and growth mindset for all students. The extent to which the instructor plays an encouraging role to promote these positive motivational factors, and emphasizes that he/she is there as a guide to help all students succeed and also

emphasizes that struggling is a stepping stone to success and should be viewed positively may also play a role in dispelling the negative impacts of societal gender stereotypes about physics that accumulate over a female student's lifetime. Since fixed mindset about innate intelligence can be a factor for the poor performance of female students, instructors should take advantage of research finding about the importance of promoting growth mindset [66] in their physics classes. In particular, research shows that students who believe that the brain is like a muscle, and intelligence is malleable and can increase with effort are more likely to persevere and perform better than those who think that intelligence is fixed [66]. Moreover, research suggests that mindset can be changed with a very short intervention [66]. Since according to the national data [76], fewer female students are likely to have taken challenging high school physics courses (e.g., Advanced Placement) before taking the college-level course, college EBAE courses which do not explicitly take into account these motivational factors may unknowingly create a high anxiety classroom environment for students who have less prior knowledge (who are more likely to be female students). For example, if students work in small groups in an EBAE course and some students in the group "show off" their knowledge and the instructional design does not promote a growth mindset, or the importance of hard work and persistence in learning physics, students who have taken less challenging physics course may have their self-efficacy issues exacerbated as opposed to reduced. Therefore, these motivational issues should be addressed in all physics classes as part of the instructional design to create inclusive classroom environment, as suggested in Ref. [77].

In summary, in order to enhance student learning in EBAE courses, it important not only to develop effective EBAE learning tools and pedagogies commensurate with students' prior knowledge but also to investigate how to implement them appropriately and how to motivate and incentivize their usage to get buy-in from students in order for them to engage with them as

intended. Furthermore, reducing the gender gap on conceptual assessments is a challenging endeavor and evidence-based pedagogies may not be sufficient. In order to reduce gender gap, it may be useful to pay attention to other factors, e.g., improving the sense of belonging and self-efficacy of female students, improving their intelligence mindset (so that they do not think of male students as having an innate ability to excel in physics that they do not have and view intelligence as something that is malleable and can be cultivated by focus and effort), and reducing competition and emphasizing collaboration.

4.5 CHAPTER REFERENCES

1. L.C. McDermott, Millikan Lecture 1990: What we teach and what is learned-Closing the gap, Am. J. Phys. **59**, 301 (1991).

2. J. Fraser, A. Timan, K. Miller, J. Dowd, L. Tucker, and E. Mazur, Teaching and physics education research: Bridging the gap, Reports on Progress in Physics **77**(3), 032401 (2014).

3. J. Docktor and J. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST PER **10**, 020119 (2014).

4. P. Heller, R. Keith and S. Anderson, Teaching problem solving through cooperative grouping. 1. Group vs individual problem solving, Am. J. Phys. **60**, 627 (1992).

5. P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. 2. Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).

6. D. Hammer, Student resources for learning introductory physics, Am. J. Phys. Physics Education Research Supplement **68** (S1), S52 (2000).

7. E. Redish, R. Steinberg and J. Saul, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

8. R. Hake, Interactive engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

9. C. Singh, What can we learn from PER: Physics Education Research?, The Phys. Teach. **52**, 568 (2014).

10. C. Singh, When physical intuition fails, Am. J. Phys. **70**(11), 1103 (2002).

11. F. Reif, Millikan Lecture 1994: Understanding and teaching important scientific thought processes, Am. J. Phys. **63**, 17 (1995).

12. S.Y. Lin and C. Singh, Effect of scaffolding on helping introductory physics students solve quantitative problems involving strong alternative conceptions, Phys. Rev. ST PER **11**, 020105 (2015).

13. S.Y. Lin and C. Singh, Using isomorphic problem pair to learn introductory physics: Transferring from a two-step problem to a three-step problem, Phys. Rev. ST PER **9**, 020114 (2013).

14. S.Y. Lin and C. Singh, Using isomorphic problems to learn introductory physics, Phys. Rev. ST PER **7**, 020104 (2011).

15. A. Mason and C. Singh, Assessing expertise in introductory physics using categorization task, Phys. Rev. ST PER **7**, 020110 (2011).

    A. Mason and C. Singh, Using categorization of problems as an instructional tool to help introductory students learn physics, Physics Education **51**, 025009 (2016).

16. C. Singh, Categorization of problems to assess and improve proficiency as teacher and learner, Am. J. Phys. **77**(1), 73 (2009).

17. C. Singh, Assessing student expertise in introductory physics with isomorphic problems. I. Performance on a nonintuitive problem pair from introductory physics, Phys. Rev. ST PER **4**, 010104 (2008).

18. C. Singh, Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer, Phys. Rev. ST PER **4**, 010105 (2008).

19. A. Collins, J. S. Brown, and S. E. Newman, Cognitive Apprenticeship: Teaching the crafts of reading, writing and mathematics, in *Knowing, learning, and instruction: Essays in honor of Robert Glaser*, edited by L. B. Resnick (Lawrence Erlbaum, Hillsdale, NJ, 1989), p. 453.

20. C.H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

21. A. Mason and C. Singh, Helping students learn effective problem solving strategies by reflecting with peers. Am. J. Phys. **78** (7), 748 (2010).

22. A. Mason and C. Singh, Impact of guided reflection with peers on the development of effective problem solving strategies and physics learning, The Phys. Teach. **54**, 295 (2016).

23. L. McDermott, P. Shaffer and the Physics Education Group, University of Washington, *Tutorials in Introductory Physics*, Prentice Hall, NJ, 2002.

24. C. Singh, Interactive learning tutorials on quantum mechanics, Am. J. Phys. **76**, 400 (2008).

25. E. Marshman and C. Singh, Interactive tutorial to improve student understanding of single photon experiments involving a Mach-Zehnder Interferometer, Euro. J. Phys. **37**, 024001 (2016).

26. C. Kalman, M. Milner-Bolotin and T. Antimirova, Comparison of the effectiveness of collaborative groups and peer instruction in a large introductory physics course for science majors, Can. J. Phys. **88**(5) 325 (2010).

27. C. Singh, Impact of peer interaction on conceptual test performance, Am. J. Phys. **73** (5) 446 (2005).

28. C. Singh, Effectiveness of group interaction on conceptual standardized test performance, Proceedings of the 2002 Phys. Ed. Res. Conference, Boise (Eds. S. Franklin, K. Cummings and J. Marx), p. 67 (2002). http://dx.doi.org/10.1119/perc.2002.pr.017

29. R. Sayer, E. Marshman and C. Singh, The impact of peer interaction on the responses to clicker questions in an upper-level quantum mechanics course, Proc. 2016 Phys. Educ. Res. Conf., Sacramento, CA, p. 304 (2016). http://dx.doi.org/10.1119/perc.2016.pr.071

30. P. Black and D. Wiliam, Assessment and classroom learning, Assessment in Education **5**(1), 7 (1998).

31. R. Moll and M. Milner-Bolotin, The effect of interactive lecture experiments on student academic achievement and attitudes towards physics, Can. J. Phys. **87**(8), 917 (2009).

32. G. Novak, E.T. Patterson, A. Gavrin and W. Christian, Just-in-Time Teaching: *Blending Active Learning with Web Technology*, Upper Saddle River, NJ: Prentice Hall 1999.

33. R. Sayer, E. Marshman and C. Singh, A case study evaluating Just-in-Time Teaching and Peer Instruction using clickers in a quantum mechanics course, Phys. Rev. ST PER **12**, 020133 (2016).

34. https://perusall.com/

35. D. Schwartz and J. Bransford, A time for telling, Cognition and Instruction **16**, 475 (1998).

36. R. Mayer, *Multimedia Learning* (Cambridge Press, 2001).

37. Z. Chen, T. Stelzer and G. Gladding, Using multi-media modules to better prepare students for introductory physics lecture, Phys. Rev. ST PER **6**, 010108 (2010).

38. F. Reif and L. Scott, Teaching scientific thinking skills: Students and computers coaching each other, Am. J. Phys. **67** (9), 819 (1999).

39. N. Schroader, G. Gladding, B. Guttman, and T. Stelzer, Narrated animated solution videos in a mastery setting, Phys. Rev. ST PER **11**, 010103 (2015).

40. C. Singh, Interactive video tutorials for enhancing problem-solving, reasoning, and meta-cognitive skills of introductory physics students, Proc. 2003 Phys. Ed. Res. Conf., Madison, WI, AIP Publishing Melville NY, **720**, p.177 (2003). http://dx.doi.org/10.1063/1.1807283

41. C. Singh and D. Haileselassie, Developing problem solving skills of students taking introductory physics via web-based tutorials, J. Coll. Sci. Teaching **39 (**4), 42 (2010).

42. J.L. Bishop and M.A. Verleger, The Flipped Classroom: A Survey of the Research, 2013 ASEE Annual Conference & Exposition (2013).

43. D. Berrett, How 'flipping' the classroom can improve the traditional lecture, The Chronicle of Higher Education, Feb. 19 (2012).

44. E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Engelwood Cliffs) 1997.

45. D. Haak, J. HilleRisLambers, E. Pitre, and S. Freeman, Increased structure and active learning reduce the achievement gap in introductory biology, Science **332**, 1213 (2011).

46. L. Breslow, D. Pritchard, J. DeBoer, G. Stump, A. Ho and V. Seaton, Studying learning in the worldwide classroom: research into edX's first MOOC, Res. Prac. Assess. **8**, 13 (2013).

47. P. Laws, M. Willis, D. Jackson, K. Koenig and R. Teese, Using research-based interactive video vignettes to enhance out-of-class learning in introductory physics, The Phys. Teach. **53**, 114 (2015).

48. D. Hestenes, M. Wells and G. Swackhamer, Force Concept Inventory, The Phys. Teach. **30**, 141 (1992).

49. D. Maloney, T. O'Kuma, C. Hieggelke and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism. Am. J. Phys. Supplement **69** (7) s12 (2001).

50. A. Madsen, S.B. McKagan, and E.C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST PER **9**, 020121 (2013).

    A.L. Traxler, X.C. Cid, J. Blue and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).

    B.A. Adegoke, Impact of interactive engagement on reducing the gender gap in quantum physics learning outcomes among senior secondary school students, Physics Education **47** (4) 462 (2012).

51. J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and introductory physics performance, AIP Conf. Proc. **1064**, p.15 (2008).

52. S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch and N.R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison, Euro. J. Phys. **34**(2), 421 (2013).

53. M. Lorenzo, C. Crouch and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**(2), 118 (2006).

54. R. Beichner, J. Saul, D. Abbott, J. Morse, D. Deardorff, R. Allain, S. Bonham, M. Dancy and J. Risley, The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project, in *Research-Based Reform of University Physics*, edited by E. Redish and P. Cooney, 2007 (published online).

55. M. Cahill, K. Hynes, R. Trousil, L. Brooks, M. McDaniel, M. Repice, J. Zhao, R. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST PER 10:2 (2014).

56. J. Day, J. Stang, N. Holmes, D. Kumar and D. Bonn, Gender gaps and gendered action in a first-year physics laboratory, Phys. Rev. ST PER 12:2 (2016).

57. A. Maries and C. Singh, Stereotype threat? : Effects of inquiring about test takers' gender on conceptual test performance in physics, Proceedings of the 5th International Conference on Women in Physics, AIP Conf. Proc., Melville, NY **1697** (2015). http://dx.doi.org/10.1063/1.4937713

58. V. P. Coletta and J.A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73,** 1172 (2005).

    V. P. Coletta, J.A. Phillips, and J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413,** p. 23 (2012).

59. T. Majors and P. Engelhardt, Gender & LEAP Pedagogy: What does the Gender Force Concept Inventory have to say?, PERC Proceedings p. 167 (2014) (available online).

60. G. C. Marchand and G. Taasoobshirazi, Stereotype threat and women's performance in physics, International Journal of Science Education **35** (18), 3050 (2013).

61. M. Appel and N. Kronberger, Stereotypes and the achievement gap: Stereotype threat prior to test taking, Educational Psychology Review **24** (4), 609 (2012).

62. C. McKown and R.S. Weinstein, The development and consequences of stereotype consciousness in middle childhood, Child Development **74**, 498 (2003).

63. L. Bian, S.-J. Leslie and A. Cimpian, Gender stereotypes about intellectual ability emerge early and influence children's interests, Science **355**, 6323 (2017).

64. A. Bandura, Self-efficacy: Toward a Unifying Theory of Behavioral Change, Psychological Review **84**(2), 19 (1977).

65. K. Miller, J. Schell, A. Ho, B. Lukoff, and E. Mazur, Response switching and self-efficacy in Peer Instruction classrooms, Phys. Rev. ST PER **11**, 010104 (2015).

66. C. Dweck, *Mindset: The new psychology of success*. Random House Incorporated, 2006.

67. C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys **71** (6), 607 (2003).

68. L. Rimoldini and C. Singh, Student understanding of rotational and rolling motion concepts, Phys. Rev. ST PER **1**, 010102 (2005).

69. L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Valuating an assessment tool: Brief electricity and magnetism asssessment, Phys. Rev. ST PER **1**, 10105 (2006).

70. C. Singh and D. Rosengrant, Students' conceptual knowledge of energy and momentum, Proc. Phys. Educ. Res. Conf., Rochester, p. 123 (2001). http://dx.doi.org/10.1119/perc.2001.pr.018

71. J. Li and C. Singh, Developing a magnetism conceptual survey and assessing gender differences in student understanding of magnetism, Proceedings of the Phys. Ed. Res. Conference, AIP Conf. Proc., Melville, New York, **1413,** 43 (2012). http://dx.doi.org/10.1063/1.3679989

    J. Li and C. Singh, Developing and validating a conceptual survey to assess introductory physics students' understanding of magnetism, Euro. J. Phys. **38** (2), 025702 (2017).

72. C. Singh, Student understanding of symmetry and Gauss's law of electricity, Am. J. Phys. **74** (10), 923 (2006).

    C. Singh, Student understanding of symmetry and Gauss's law, Proceedings of the Phys. Ed. Res. Conference, AIP Conf. Proc., Melville, NY, **790,** 65 (2005). http://dx.doi.org/10.1063/1.2084702

73. G. Glass and K. Hopkins, *Statistical Methods in Education and Psychology*, 3rd Ed. Pearson, 1996.

74. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Routledge, 1988.

75. G.A. Smith, First-day questions for the learner-centered classroom, National Teaching and Learning Forum **17**, 1 (2008).

76. https://www.aip.org/statistics/women

77. L. Aguilar, G. Walton and C. Wieman, Psychological insights for improved physics teaching, Phys. Today **67**, 43 (2014).

4.6 CHAPTER APPENDIX

4-10. Average FCI pretest scores (Pretest), posttest scores (Posttest), gain (Gain), normalized gain (Norm g) and final exam scores (Final) for male and female students in the flipped and LB courses taught by the same instructor

(with same homework and final exam) with students divided into three groups regardless of their gender based on their pretest scores. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest, posttest or final exam.

	Pretest Split		Pretest	Posttest	Gain	Norm g	Final
FCI Alg LB	bottom 1/3	Mean Female Score	19	35	17	20	48
		p-value	0.192	0.171			0.082
		Mean Male Score	15	40	25	30	40
	middle 1/3	Mean Female Score	32	46	13	20	53
		p-value	0.967	0.955			0.772
		Mean Male Score	32	45	13	19	54
	top 1/3	Mean Female Score	49	62	13	26	64
		p-value	0.046	0.121			0.934
		Mean Male Score	56	70	13	31	64
FCI Alg EBAE	bottom 1/3	Mean Female Score	16	40	24	28	54
		p-value	0.149	0.610			0.438
		Mean Male Score	18	42	24	29	51
	middle 1/3	Mean Female Score	31	45	14	21	52
		p-value	0.038	0.013			0.159
		Mean Male Score	33	54	21	32	57
	top 1/3	Mean Female Score	53	71	18	38	63
		p-value	0.091	0.209			0.212
		Mean Male Score	58	75	17	40	67

4-11. Average CSEM pretest scores (Pretest), posttest scores (Posttest), gain (Gain), normalized gain (Norm g) and final exam scores (Final) for male and female students in the flipped and LB courses taught by the same instructor (with same homework and final exam) with students divided into three groups regardless of their gender based on

their pretest scores. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest, posttest and final exam.

	Pretest Split		Pretest	Posttest	Gain	Norm g	Final
CSEM Calc LB	bottom 1/3	Mean Female Score	27	37	11	15	44
		p-value	0.945	0.505			0.921
		Mean Male Score	26	35	8	11	44
	middle 1/3	Mean Female Score	39	49	10	16	52
		p-value	0.701	0.298			0.688
		Mean Male Score	39	45	7	11	53
	top 1/3	Mean Female Score	55	62	7	15	57
		p-value	0.507	0.613			0.578
		Mean Male Score	53	59	6	13	59
CSEM Calc EBAE	bottom 1/3	Mean Female Score	25	41	16	21	50
		p-value	0.542	0.480			0.436
		Mean Male Score	24	44	20	26	53
	middle 1/3	Mean Female Score	38	49	11	18	52
		p-value	0.433	0.909			0.288
		Mean Male Score	39	49	11	17	57
	top 1/3	Mean Female Score	57	63	6	14	69
		p-value	0.949	0.015			0.869
		Mean Male Score	57	73	16	36	69

4-12. Average FCI pretest scores (Pretest), posttest scores (Posttest), gain (Gain) and normalized gain (Norm g) for male and female students in the flipped and LB algebra-based and calculus-based courses. All courses in the same group were combined with students divided into three groups regardless of their gender based upon their pretest scores. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest and posttest. ***Note that FCI data for Calculus-based EBAE classes are not available.***

	Pretest Split	Pretest	Posttest	Gain	Norm g

FCI Calc LB	bottom 1/3	Mean Female Score	30	43	13	19
		p-value	0.083	0.016		
		Mean Male Score	32	49	17	25
	middle 1/3	Mean Female Score	53	60	7	14
		p-value	0.049	<0.001		
		Mean Male Score	55	71	16	36
	top 1/3	Mean Female Score	83	84	1	7
		p-value	0.918	0.707		
		Mean Male Score	82	82	-1	-3
FCI Alg EBAE	bottom 1/3	Mean Female Score	16	40	24	28
		p-value	0.149	0.610		
		Mean Male Score	18	42	24	29
	middle 1/3	Mean Female Score	31	45	14	21
		p-value	0.038	0.013		
		Mean Male Score	33	54	21	32
	top 1/3	Mean Female Score	53	71	18	38
		p-value	0.091	0.209		
		Mean Male Score	58	75	17	40
FCI Alg LB	bottom 1/3	Mean Female Score	19	34	15	19
		p-value	0.467	0.427		
		Mean Male Score	20	36	16	20
	middle 1/3	Mean Female Score	32	46	13	20
		p-value	0.617	0.501		
		Mean Male Score	33	47	15	22
	top 1/3	Mean Female Score	50	64	14	27
		p-value	0.002	0.005		
		Mean Male Score	56	71	15	34

4-13. Average CSEM pretest scores (Pretest), posttest scores (Posttest), gain (Gain) and normalized gain (Norm g) for male and female students in the flipped and LB algebra-based and calculus-based courses. All courses in the same group were combined with students divided into three groups regardless of their gender based upon their

pretest scores. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pretest and posttest. ***Note that CSEM data for algebra-based EBAE classes are not available.***

	Pretest Split		Pretest	Posttest	Gain	Norm g
CSEM Calc EBAE	bottom 1/3	Mean Female Score	23	44	22	28
		p-value	0.571	0.001		
		Mean Male Score	22	57	35	45
	middle 1/3	Mean Female Score	35	56	22	33
		p-value	0.771	0.891		
		Mean Male Score	35	56	21	32
	top 1/3	Mean Female Score	54	65	10	23
		p-value	0.718	0.033		
		Mean Male Score	55	72	17	38
CSEM Calc LB	bottom 1/3	Mean Female Score	23	39	16	20
		p-value	0.367	0.827		
		Mean Male Score	24	39	15	20
	middle 1/3	Mean Female Score	35	46	11	17
		p-value	0.833	0.535		
		Mean Male Score	35	48	13	20
	top 1/3	Mean Female Score	51	61	10	21
		p-value	0.966	0.460		
		Mean Male Score	51	58	7	14
CSEM Alg LB	bottom 1/3	Mean Female Score	15	36	21	25
		p-value	0.402	0.576		
		Mean Male Score	15	38	23	27
	middle 1/3	Mean Female Score	23	42	20	25
		p-value	0.981	0.192		
		Mean Male Score	23	46	23	30
	top 1/3	Mean Female Score	33	45	13	19
		p-value	0.002	<0.001		

	Mean Male Score	37	56	19	30

5.0 IMPACT OF STEREOTYPE THREAT ON STUDENT PERFORMANCE AND GENDER GAP IN INTRODUCTORY PHYSICS

5.1 INTRODUCTION

Prior research has found that in the introductory physics courses, male students often outperform female students on conceptual assessments such as the Force Concept Inventory or FCI [1, 2] and the Conceptual Survey of Electricity and Magnetism or CSEM [3], a phenomenon sometimes referred to as the "gender gap". Furthermore, prior research has also found that activation of a stereotype about a particular group in a test-taking situation, i.e., stereotype threat (ST), can alter the performance of that group in a way consistent with the stereotype. For example, Spence et al. [4] conducted a study in which a group of students was told immediately before taking a mathematics test that in prior administrations of the test, a gender gap has been found (with female students performing worse than male students), while another group was not provided with this information. Female students who were informed about the stereotype right before the test performed significantly worse than those who were not exposed to this stereotype, but the performance of male students was unaffected. The researchers concluded that informing female students about the stereotype acts as a stereotype threat and leads to deteriorated performance [4]. Spence et al. [4] also describe another study in which, when students were told that the mathematics test they are about to take is gender neutral, no gender gap was observed, but in the control condition, when students were not given any such information about the gender neutrality of the mathematics test, a gender gap was observed. The researchers hypothesized that a stereotype

threat may be present for female students in a mathematics test-taking situation unless they are explicitly told that the mathematics test is previously found to be gender neutral [4].

Other researchers have found more subtle stimuli that can activate stereotype threat and result in deteriorated performance [5], e.g., asking students to indicate their ethnicity before taking a test [6]. In particular, prior research suggests that asking African American students to indicate their ethnicity before taking a difficult test on verbal ability resulted in decreased performance compared to students of the same race who were not asked for this information [6]. Yet others have found that asking for gender or ethnicity before taking a test did not impact students' performance on standardized tests [7, 8].

Even in the context of physics, some researchers have argued [9] that a stereotype threat is automatically triggered in a physics test-taking situation due to prevalent societal stereotypes. In a somewhat similar study to Spence et al. [4], Marchand and Taasoobshirazi [9] used three different conditions immediately before students took a four question quantitative physics test: an explicit stereotype threat condition (students were told that female students had performed worse than male students on this test), an implicit stereotype threat condition (no information regarding past performance of male and female students was given), and a nullified condition (students were told that no gender differences were found on previous administrations of the test). Under all three conditions [9], students received the following instructions: "You will be given four physics problems to solve. These problems are based on physics material that you have already covered." In the implicit stereotype threat condition, these were the only instructions, while in the explicit stereotype threat condition, students were also told: "This test has shown gender differences with males outperforming females on the problems" and in the nullified condition, students were told: "No gender differences in performance have been found on the test". While male students

performed similarly in all three conditions, females in the explicit and implicit stereotype threat conditions had comparable performances but performed statistically significantly worse than female students in the nullified condition.

5.2 GOALS OF THE INVESTIGATIONS

Since the stereotype threat has the potential to exacerbate the gender gap typically found in conceptual physics assessments, in the Study 1 described here, our goal was to investigate whether asking introductory physics students to indicate their gender before taking the CSEM impacted their performance, both when it was administered as a pretest (before instruction) and as a posttest (after traditional lecture-based instruction in relevant concepts). In the Study 2 described here, our goal was to investigate the prevalence of the belief that men generally perform better in physics than women (a gender stereotype) among introductory physics students and the extent to which agreeing with this gender stereotype is correlated with the performance of female and male students in algebra-based and calculus-based introductory physics I and II on the commonly used conceptual standardized physics assessments, the FCI and the CSEM. We also investigated whether there was a difference between the condition in which the gender stereotype question was asked immediately before or immediately after the introductory physics students took the FCI or the CSEM to understand whether there was a qualitative difference between the performance of female students in these before and after conditions.

As noted, Marchand and Taasoobshirazi [9] have posited that their research suggests that many female students automatically experience a certain level of stereotype threat while taking a physics test due to the societal stereotypes about physics being a discipline for intelligent men. We

hypothesized that while a certain level of stereotype threat may be implicitly present for many female students in the introductory physics courses as Marchand and Taasoobshirazi [9] have argued based upon their research, the stereotype threat may be worse, on average, for female students taking introductory physics if they agree with the gender stereotype that men generally perform better in physics than women. Moreover, without explicit intervention to improve women's sense of belonging, self-efficacy and growth mindset (as opposed to fixed mindset, e.g., agreeing with the statement that according to my own belief, I expect men to generally perform better in physics than women), being in a physics course in which they are severely underrepresented can have worse negative impact on the performance of the female students who believe in the gender stereotype than those who do not believe in the stereotype. In particular, it is possible that for those female students who agree with the gender stereotype, the ecosystem of the physics classrooms in which they are underrepresented may act as an additional level of stereotype threat (over and above what Marchand and Taasoobshirazi [9] argue many female students experience automatically in physics test-taking situations due to common societal biases), and they may perform worse than female students who do not agree with the stereotype. Our goal was to especially investigate this issue.

5.3 METHODOLOGY

The participants in this study were students in various algebra-based and calculus-based introductory physics courses. Also, introductory physics courses (algebra-based physics I and II or calculus-based physics I and II) included in this study were large introductory physics courses at a typical large research university (University of Pittsburgh or Pitt) except in one study, as

described below, calculus-based introductory physics students from another large research university (University of Cincinnati) participated. The two semester calculus-based course sequence at Pitt and Cincinnati is mainly taken by college freshman who are engineering, chemistry, mathematics or physics majors. Approximately, 30% of the students in these calculus-based courses are females (somewhat higher percentage at Pitt than at Cincinnati). The first semester course covers mainly mechanics and waves and the second semester course covers mainly electricity and magnetism and some wave-optics. The algebra-based introductory physics course sequence at Pitt is taken mainly by the biological science and neuroscience majors and those who are pre-medical students. It is taken in the junior or senior year. Introductory mechanics and waves are covered in the first semester algebra-based course and electricity and magnetism are covered in the second semester algebra-based course, although other topics are also included in the course in order to cover the topics in the medical entrance examination. Approximately, 60% of the students in these algebra-based courses are females.

The students in the calculus-based course had four hours of lecture time and the algebra-based course had three hours of lecture time per week. Both calculus-based and algebra-based courses had one hour of recitation time. In the recitations, the graduate teaching assistants typically fielded questions about the homework from the students and solved example problems on the board. Each week, after students submitted the textbook style mostly quantitative homework on a particular topic, they were typically given a recitation quiz in the last 15-20 minutes of the recitation class.

In order to compare the performances of students under different conditions, we performed *t*-tests [10] on FCI or CSEM pre and posttest data for males and females. We also calculated the effect size in the form of Cohen's *d* defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma_{pooled}}$, where μ_1 and μ_2 are the averages

of the two groups being compared and $\sigma_{pooled} = \sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}$; here σ_1 and σ_2 are the standard deviations of the two groups being compared. We considered: $d < 0.5$ as small effect size, $0.5 \leq d < 0.8$ as medium effect size and $d \geq 0.8$ as large effect size, as described in Ref. [10].

5.3.1 Study 1

In this study, 170 students in an introductory algebra-based physics II course (as noted, mostly biological and neuroscience majors and pre-medical students) took the CSEM as a pretest (in the recitation class in the first week of classes before instruction in relevant concepts) and as a posttest (in the recitation class during the last week of classes after instruction in relevant concepts). Students were assigned to two conditions, one which asked them to indicate their gender (checkbox format with options male, female, and prefer not to specify) before then took the CSEM and one in which they were asked for such information after taking the CSEM. We then compared the performance of students under the two conditions.

5.3.2 Study 2

In this study, we investigated the following: 1) the prevalence of the belief in the gender stereotype among introductory physics students in the algebra-based and calculus-based courses and 2) the extent to which believing the stereotype is correlated with female and male students' performance on the FCI and CSEM. This study involved over 1800 calculus-based students (mainly engineering, mathematics and physical science majors) and over 1600 algebra-based students (mainly pre-medical and biological and neuroscience majors) enrolled in first and second semester

introductory physics courses. The majority of these students, after taking the FCI or CSEM, were asked to indicate the extent to which they agree with the following statement: "According to my own personal beliefs, I expect men to generally perform better in physics than women" on a five-point Likert scale (strongly disagree, disagree, neutral, agree, and strongly agree). Then, students were grouped according to their beliefs (agree/strongly agree, neutral which was explained to students as neither agree nor disagree, disagree/strongly disagree) and we investigated performance differences (e.g., we compared the performance of female students who agree with the stereotype with that of female students in the same class who disagree with the stereotype) on both the pretest (before instruction) and the posttest (after instruction in relevant concepts).

We note that if students are asked to indicate the extent to which they agree with the gender stereotype (according to my own belief, I expect men to generally perform better in physics than women) before taking the FCI or CSEM, this may act as an additional stereotype threat (over and above the stereotype threat that Marchand and Taasoobshirazi [9] posit many female students automatically experience in a physics classes in a test-taking situation), especially for the female students who agree with this gender stereotype. Thus, to avoid any additional stereotype threat to female students (and potential consequences on performance on the standardized test), all students at one large state-related university (University of Pittsburgh) were given the gender stereotype question right after they had completed answering the FCI or CSEM questions. However, we wanted to test whether asking the gender stereotype question before students take the conceptual survey qualitatively impacts female and male students' performance. Since it was agreed that at Pitt, the gender stereotype question would be asked at the end (after students had taken the standardized test in the recitation) so that female students do not experience additional stereotype threat, another group of calculus-based introductory physics students at the University of

Cincinnati was asked the gender stereotype question right before taking the FCI, and the qualitative trends amongst male and female students who agreed or disagreed with the stereotype were compared with the corresponding calculus-based cohort at Pitt for whom the stereotype question was asked right after taking the FCI.

Finally, we note that it is possible that introductory physics students who answered the stereotype question as neutral (which was explained to them as neither agreeing nor disagreeing with the gender stereotype) may instead actually agree with the stereotype but avoided stating that they agree with it since they were aware it is not politically correct. Therefore, we carried out a separate analysis in which we compared the performance of introductory physics students in a course who disagree with the stereotype with the performance of all the rest of the students (i.e., those who were either neutral or agreed with the stereotype) of a particular gender in that course. That type of analysis of data shows very similar trends in those data as the ones presented here by comparing the male and female students who only agreed or disagreed with the stereotype (and discarding the neutral responses provided by students).

5.4 RESULTS

Before discussing the results, we note that whether we use matched pretest and posttest data or consider all students who took the pretest or posttest (unmatched), the qualitative trends are unchanged, so we report data from all students who took the pretest or posttest.

5.4.1 Study 1

Table 5.1 shows the pretest and posttest performance of introductory algebra-based female (N=99) and male (N=71) students on the CSEM in the two conditions: students were/were not asked to provide gender information before taking the CSEM (gender salient/not salient condition, respectively). Table 5.1 shows that there were no statistically significant differences between the performance of male or female students in the two conditions (e.g., female students who wrote their gender before taking the CSEM did not perform worse than female students who wrote their gender after taking the CSEM) in the pretest or the posttest.

Table 5.1. Female (F) and male (M) students' pretest and posttest performance on the CSEM depending on the testing condition. The standard deviations are abbreviated as SD. The *p* values are obtained using a t-test and *d* refers to the effect size (Cohen's *d* [10]).

Algebra	PHYSICS II (CSEM)			
	PRETEST		POSTTEST	
	Female	Male	Female	Male
NOT SALIENT	N: 46 Mean: 23 SD: 10	N: 27 Mean: 27 SD: 10	N: 12 Mean: 38 SD: 17	N: 7 Mean: 46 SD: 19
Comparison	↑ *p*: 0.692 *d*: 0.080 ↓	↑ *p*: 0.603 *d*: 0.124 ↓	↑ *p*: 0.332 *d*: 0.330 ↓	↑ *p*: 0.866 *d*: 0.074 ↓
SALIENT	N: 53 Mean: 24 SD: 10	N: 44 Mean: 28 SD: 13	N: 80 Mean: 43 SD: 14	N: 58 Mean: 48 SD: 16

5.4.2 Study 2

Table 5.2 shows the percentage of male and female introductory students in algebra-based and calculus-based physics I and II courses who agreed/were neutral/disagreed with the stereotype (According to my own belief, I expect men to generally perform better in physics than women). Only 7-13% of algebra-based and calculus-based students (regardless of their gender) agreed with this gender stereotype. Thus, it appears that this stereotype was not very common amongst college introductory physics students. However, if the neutral responses are combined with those who agree with the stereotype, approximately 25% of the introductory physics students of both genders either agree with the stereotype or are neutral (neither agree nor disagree).

Table 5.2. Percentage of female (F) and male (M) students who agreed/were neutral/disagreed with the stereotype that men generally perform better in physics than women in algebra-based (Alg.) and calculus-based (Calc.) introductory physics. The total number of female/male students is indicated at the bottom (N).

	Alg. Physics I (FCI)				Alg. Physics II (CSEM)				Calc. Physics I (FCI)				Calc. Physics II (CSEM)			
	Pretest		Posttest		Pretest		Posttest		Pretest		Posttest		Pretest		Posttest	
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Disagree	77	73	74	73	80	78	76	79	83	72	83	74	83	74	77	73
Neutral	14	21	13	21	9	15	12	13	10	21	7	18	9	19	10	20
Agree	9	7	13	7	11	7	12	7	8	7	10	8	9	7	13	7
N	668	365	450	251	553	330	348	219	253	453	217	354	231	527	181	396

Before presenting data from Study 2, we note that in all classes regardless of whether they were algebra-based or calculus-based, whether they were introductory physics I or II, large gender differences were found in our investigation between the performance of male and female students both on the pretest and posttest. The gender gap on the FCI is typically 10-20% depending upon

whether it is the pretest or posttest and whether it is the algebra-based or calculus-based introductory course. The gender gap on the CSEM is typically smaller, especially on the pretest since the overall scores of each group on the pretest (particularly for the algebra-based course) are not significantly better than random guessing (20% average). As noted earlier, some have attributed gender differences in student performance in a particular class on the standardized conceptual assessments (the FCI or the CSEM) on these tests being biased against female students [11]. Others have argued that many female students in a physics class automatically experience a stereotype threat due to societal biases especially when taking a test [9].

Tables 5.3 and 5.4 show the pretest and posttest performances on the FCI and CSEM of female and male students in the algebra-based (Table 5.3) and calculus-based physics (Table 5.4) courses who agreed/disagreed with the gender stereotype. The tables also list *p* values and effect sizes (Cohen's d) for the comparison of the performance of female/male students who agree with the stereotype with that of female/male students who disagree with the stereotype.

Table 5.3 shows that for the algebra-based introductory physics students, neither on the FCI nor on the CSEM are there major differences between the female (male) students who agree and female (male) students who disagree with the gender stereotype, in the pretest or in the posttest.

Table 5.3. Numbers of algebra-based students (N), averages (Mean) and standard deviations (SD) for the performance on the FCI (for Physics I) or CSEM (for Physics II) in pre-/post- tests of female and male students who agree/disagree with the stereotype that men generally perform better in physics than women. The *p* values (*p*) and effect sizes (*d*) shown with the performance of female/male students for each class type were obtained when comparing the average performance of female/male students who agree with that of female/male students who disagree with the stereotype. These students answered the stereotype question after taking the FCI/CSEM.

Algebra	PHYSICS I (FCI)				PHYSICS II (CSEM)			
	PRETEST		POSTTEST		PRETEST		POSTTEST	
	Female	Male	Female	Male	Female	Male	Female	Male

Disagree	N: 512 Mean: 32 SD: 15	N: 265 Mean: 44 SD: 19	N: 333 Mean: 48 SD: 17	N: 182 Mean: 61 SD: 19	N: 441 Mean: 22 SD: 8	N: 257 Mean: 26 SD: 10	N: 263 Mean: 37 SD: 13	N: 174 Mean: 44 SD: 16
Comparison	↑ *p*: 0.718 *d*: 0.047 ↓	↑ *p*: 0.111 *d*: 0.316 ↓	↑ *p*: 0.849 *d*: 0.028 ↓	↑ *p*: 0.588 *d*: 0.152 ↓	↑ *p*: 0.043 *d*: 0.256 ↓	↑ *p*: 0.957 *d*: 0.012 ↓	↑ *p*: 0.680 *d*: 0.067 ↓	↑ *p*: 0.632 *d*: 0.128 ↓
Agree	N: 60 Mean: 31 SD: 14	N: 25 Mean: 50 SD: 16	N: 57 Mean: 47 SD: 17	N: 17 Mean: 57 SD: 24	N: 61 Mean: 20 SD: 6	N: 24 Mean: 26 SD: 11	N: 43 Mean: 37 SD: 12	N: 16 Mean: 46 SD: 16

Table 5.4. Numbers of calculus-based students (N), averages (Mean) and standard deviations (SD) for the performance on the FCI (for Physics I) or CSEM (for Physics II) in pre-/post- tests of female and male students who agree/disagree with the stereotype that men generally perform better in physics than women. The *p* values (*p*) and effect sizes (*d*) shown with the performance of female/male students for each class type were obtained when comparing the average performance of female/male students who agree with that of female/male students who disagree with the stereotype. These students answered the stereotype question after taking the FCI/CSEM.

Calculus	PHYSICS I (FCI)				PHYSICS II (CSEM)			
	PRETEST		POSTTEST		PRETEST		POSTTEST	
	Female	Male	Female	Male	Female	Male	Female	Male
Disagree	N: 209 Mean: 43 SD: 19	N: 326 Mean: 60 SD: 21	N: 180 Mean: 56 SD: 19	N: 262 Mean: 70 SD: 19	N: 191 Mean: 35 SD: 14	N: 388 Mean: 40 SD: 14	N: 140 Mean: 50 SD: 18	N: 288 Mean: 58 SD: 19
Comparison	↑ *p*: 0.801 *d*: 0.056 ↓	↑ *p*: 0.592 *d*: 0.102 ↓	↑ *p*: 0.169 *d*: 0.313 ↓	↑ *p*: 0.249 *d*: 0.231 ↓	↑ *p*: 0.233 *d*: 0.240 ↓	↑ *p*: 0.954 *d*: 0.011 ↓	↑ *p*: 0.020 *d*: 0.505 ↓	↑ *p*: 0.185 *d*: 0.283 ↓
Agree	N: 19	N: 30	N: 22	N: 30	N: 20	N: 37	N: 23	N: 29

	Mean: 42 SD: 16	Mean: 58 SD: 20	Mean: 51 SD: 18	Mean: 75 SD: 20	Mean: 33 SD: 10	Mean: 40 SD: 18	Mean: 42 SD: 14	Mean: 64 SD: 23

Table 5.4 shows that for the calculus-based introductory physics students, on the FCI, there are no statistically significant differences between the female (male) students who agree and the female (male) students who disagree with the stereotype, in the pretest or in the posttest (although the trends for the average scores suggest that the female students who agree with the stereotype perform worse than the female students who disagree with it and male students who agree with the stereotype perform better than the male students who disagree with it and for a larger N these results may become statistically significant). Also, Table 5.4 shows that for the calculus-based students, on the CSEM, the trends are similar to the trends on the FCI and the differences between female (or male) students who agree or disagree with the gender stereotype are not statistically significant on the pretest. However, on the CSEM posttest, there is a statistically significant difference (a difference of 8%) between the calculus-based female students who agree and the female students in the same course who disagree with the stereotype.

As mentioned earlier, we also analyzed the performance of another group of calculus-based students from University of Cincinnati who answered the gender stereotype question before taking the FCI because we wanted to investigate whether asking students the gender stereotype question before taking the FCI may act as another source of stereotype threat, especially for female students who agree with the stereotype. The results are shown in Table 5.5 and are qualitatively similar to the FCI data shown in Table 5.4. One hypothesis for this similarity is that female students who believe in the stereotype that men generally perform better in physics than women may experience similar stereotype threat regardless of whether they are asked the gender stereotype question before

or after taking the standardized test (recall that the authors of Ref. [9] posit that implicit stereotype threat is there for female students in a physics exam situation even if they are not reminded of a gender stereotype before taking a test). In other words, our finding is consistent with the findings of Marchand and Taasoobshirazi in a somewhat different context [9].

Table 5.5. Numbers of calculus-based students (N), averages (Mean) and standard deviations (SD) for the performance on the FCI (for Physics I) or CSEM (for Physics II) in pre-/post- tests of female and male students who agree/disagree with the stereotype (that men generally perform better in physics than women). The *p* values (*p*) and effect sizes (*d*) shown with the performance of female/male students for each class type were obtained when comparing the average performance of female/male students who agree with that of female/male students who disagree with the stereotype. These students answered the stereotype question before taking the FCI.

	PHYSICS I (FCI)			
	PRETEST		POSTTEST	
	Female	Male	Female	Male
Disagree	N: 151 Mean: 40 SD: 17	N: 381 Mean: 54 SD: 20	N: 46 Mean: 55 SD: 18	N: 289 Mean: 68 SD: 20
Comparison	↑ *p*: 0.416 *d*: 0.16 ↓	↑ *p*: 0.412 *d*: 0.09 ↓	↑ *p*: 0.357 *d*: 0.27 ↓	↑ *p*: 0.072 *d*: 0.08 ↓
Agree	N: 35 Mean: 37 SD: 19	N: 103 Mean: 52 SD: 20	N: 14 Mean: 50 SD: 19	N: 47 Mean: 70 SD: 19

Finally, the analysis in which students who disagreed with the stereotype were put in one group and all the other students were put in another is shown in Tables 5.6 and 5.7. The results are qualitatively the same as those shown in Tables 5.2 and 5.3.

Table 5.6. Numbers of algebra-based students (N), averages (Mean) and standard deviations (SD) for the performance on the FCI (for Physics I) or CSEM (for Physics II) in pre-/posttests of female and male students who agree/disagree with the stereotype that men generally perform better in physics than women. Students who selected 'neutral' were considered to have agreed with the stereotype question. The *p* values (*p*) and effect sizes (*d*) shown with the performance of female/male students for each class type were obtained when comparing the average performance of female/male students who agree with that of female/male students who disagree with the stereotype.

Algebra	PHYSICS I (FCI)				PHYSICS II (CSEM)			
	PRETEST		POSTTEST		PRETEST		POSTTEST	
	Female	Male	Female	Male	Female	Male	Female	Male
Disagree	N: 512 Mean: 32 SD: 15	N: 265 Mean: 44 SD: 19	N: 333 Mean: 48 SD: 17	N: 182 Mean: 61 SD: 19	N: 441 Mean: 22 SD: 8	N: 257 Mean: 26 SD: 10	N: 263 Mean: 37 SD: 13	N: 174 Mean: 44 SD: 16
Comparison	↑ *p*: 0.106 *d*: 0.147 ↓	↑ *p*: 0.288 *d*: 0.125 ↓	↑ *p*: 0.610 *d*: 0.056 ↓	↑ *p*: 0.784 *d*: 0.040 ↓	↑ *p*: 0.041 *d*: 0.210 ↓	↑ *p*: 0.596 *d*: 0.073 ↓	↑ *p*: 0.060 *d*: 0.238 ↓	↑ *p*: 0.704 *d*: 0.065 ↓
Agree	N: 156 Mean: 30 SD: 15	N: 100 Mean: 42 SD: 20	N: 117 Mean: 47 SD: 18	N: 69 Mean: 60 SD: 22	N: 112 Mean: 20 SD: 7	N: 73 Mean: 27 SD: 12	N: 85 Mean: 34 SD: 13	N: 45 Mean: 45 SD: 17

Table 5.7. Numbers of calculus-based students (N), averages (Mean) and standard deviations (SD) for the performance on the FCI (for Physics I) or CSEM (for Physics II) in pre-/post- tests of female and male students who agree/disagree with the stereotype that men generally perform better in physics than women. Students who selected 'neutral' were considered to have agreed with the stereotype question. The *p* values (*p*) and effect sizes (*d*) shown with the performance of female/male students for each class type were obtained when comparing the average performance of female/male students who agree with that of female/male students who disagree with the stereotype.

Calculus	PHYSICS I (FCI)		PHYSICS II (CSEM)	
	PRETEST	POSTTEST	PRETEST	POSTTEST

	Female	Male	Female	Male	Female	Male	Female	Male
Disagree	N: 209 Mean: 43 SD: 19	N: 326 Mean: 60 SD: 21	N: 180 Mean: 56 SD: 19	N: 262 Mean: 70 SD: 19	N: 191 Mean: 35 SD: 14	N: 388 Mean: 40 SD: 14	N: 140 Mean: 50 SD: 18	N: 288 Mean: 58 SD: 19
Comparison	↑ *p*: 0.700 *d*: 0.063 ↓	↑ *p*: 0.095 *d*: 0.178 ↓	↑ *p*: 0.295 *d*: 0.187 ↓	↑ *p*: 0.959 *d*: 0.006 ↓	↑ *p*: 0.042 *d*: 0.317 ↓	↑ *p*: 0.178 *d*: 0.136 ↓	↑ *p*: 0.002 *d*: 0.549 ↓	↑ *p*: 0.534 *d*: 0.072 ↓
Agree	N: 44 Mean: 42 SD: 18	N: 127 Mean: 56 SD: 22	N: 37 Mean: 53 SD: 18	N: 92 Mean: 70 SD: 20	N: 40 Mean: 32 SD: 10	N: 139 Mean: 38 SD: 15	N: 41 Mean: 40 SD: 16	N: 108 Mean: 59 SD: 21

5.5 DISCUSSION AND SUMMARY

The research in Study 1 suggests that asking algebra-based introductory physics students to indicate their gender before taking the CSEM did not impact their performance, consistent with a previous study conducted with the AP calculus exam and the Computerized Placement test [7]. One possible explanation for this finding supported by previous research [9] is that stereotype threat for female students occurs implicitly regardless of whether or not students are asked to indicate their gender before taking the CSEM test because the stereotype is automatically activated for female students in the test-taking situation in physics and math. In other words, one possible explanation is that the threat may be present for this group regardless of being explicitly asked about such personal information explicitly [9]. Other high-stakes tests (e.g., MCAT, SAT) commonly require students to indicate their gender before taking the tests. If the results of Study 1 were to hold for these tests as well, then the common practice of asking for personal information

such as gender may not make a difference in the performance of the stereotypically underperforming group.

The data from Study 2 suggest that both in the algebra-based and calculus-based physics I and II, male students perform significantly better than female students both on the pretest and posttest. In prior studies, this type of discrepancy between male and female students' performance has been found even after controlling for factors such as different prior preparation or coursework of male and female students [2]. Students' intelligence mindset and self-efficacy can impact their performance [12-14]. The origins of gender gap on the FCI both at the beginning and end of a physics course have been a subject of debate with some researchers arguing that the test itself is gender-biased [11, 15-16]. Some of the origins of the gender gap can be attributed to societal gender stereotypes [9, 17-19] that keep accumulating from an early age. For example, research suggests that even six year old boys and girls have gendered views about smartness in favor of boys [19]. Such stereotypes can impact female students' self-efficacy [12-14], their beliefs about their ability to perform well, in disciplines such as physics in which they are underrepresented and which have been associated with brilliance. As noted, some researchers have argued [9] that female students, when working on a physics assessment, undergo an implicit stereotype threat due to the prevalent societal stereotypes. Prior research has also found that using evidence-based pedagogies can reduce the gender gap [20], but the extent to which this occurs varies. Others have found that the gender gap is not reduced despite significant use of evidence-based pedagogies [21]. Prior research has also found a gender gap on other assessments such as a conceptual assessment for introductory laboratories [22]. Yet others have found no differences in performance between male and female students on exams [23-24]. There are other studies that shed light on different aspects of gender gap [25-39] which can also be used to interpret the findings of Study 2.

In Study 2, we investigated the prevalence of the belief that men generally perform better in physics than women among introductory physics students and found that this type of belief is not very common (around 7-13% of algebra-based and calculus-based students agreed with this stereotype). We also investigated the extent to which agreeing with the stereotype was correlated with students' performance on the FCI and CSEM. The analysis of data from Study 2 suggests that algebra-based female students who agreed with the gender stereotype (men generally perform better in physics than women) and female students who disagreed with the stereotype showed similar performance (within 2%) in both the pretest and the posttest. In other words, for algebra-based students, there were no major differences between female students who agreed with the stereotype and female students who disagreed with it. For calculus-based students, there were no differences on the FCI (although there was a discernable trend emerging and larger number of students may make it statistically significant), but for the CSEM, in the posttest, female students who agreed with the gender stereotype performed worse than female students who disagreed with it. In other words, at the end of the full year of a calculus-based introductory physics sequence, a statistically significant difference in the CSEM posttest for the calculus-based students emerged in that the female students who agreed with the stereotype performed significantly worse than female students who disagreed with the stereotype (this result is not only statistically significant but also has practical implications since there is 8% difference in female student performance between those who agree and disagree with the stereotype).

We note that in algebra-based courses, approximately 60% of the students were female (compared to approximately 30% in the calculus-based courses). Thus, in a calculus-based course, female students who agreed with the stereotype are likely to be impacted more by the associated stereotype threat since they see fewer female students compared to male students in their physics

class. In other words, the observation that there are fewer female students in a physics class compared to male students can reinforce the stereotype and hence has the potential to cause a larger stereotype threat. This could lead to increased anxiety for female students in a test-taking situation in a calculus-based course compared to an algebra-based course. In an algebra-based course, the observation that there is a larger percentage of female students in class may mitigate the impact of the gender stereotype. In other words, in the algebra-based courses, the larger number of women has the potential to reduce the impact of the additional stereotype threat even for the female students who agree with the gender stereotype.

As Marchand and Taasoobshirazi [9] have argued that, based upon their research, it is possible that a certain level of stereotype threat may be implicitly present for many female students in an introductory physics course. However, we hypothesize that the stereotype threat may be worse, on average, for female students taking introductory physics if they agree with the gender stereotype (that men generally perform better in physics than women). Moreover, without explicit intervention to improve women's sense of belonging, self-efficacy and growth mindset, being in a calculus-based physics course in which they are severely underrepresented may have had worse negative impact on the performance of the female students who believe in the stereotype than those who do not believe in this stereotype. Thus, one reason for the emergence of the statistically significantly different performance between the female students who disagree and agree with the gender stereotype on the CSEM posttest for calculus-based students may be the cumulative impact of increased stereotype threat. In particular, for women who agree with the gender stereotype, there may be additional stereotype threat over and above what Marchand and Taasoobshirazi [9] posit many female students experience in a physics test taking situation implicitly. Such an additional threat can create added level of anxiety that can impact female students' performance from several

angles. For example, due to added level of anxiety, female students who agree with the stereotype may, on average, be less excited about learning physics and this decreased level of excitement can potentially lead to task avoidance, i.e., less time learning physics. Moreover, when learning physics, some of the limited number of chunks in their working memory [40] may be taken up by the added anxiety instead of the physics involved in the problems they may be working on. Thus, the anxiety can reduce the level of focus and effectiveness of the study session. Moreover, during an exam, these female students who experience the added level of anxiety due to the additional stereotype threat may not be able to use all of their limited cognitive resources effectively to solve the problems and their working memory [40] may again be used up partly by the anxiety of taking the physics test. Since physics is a hierarchical discipline in which different concepts build on each other, it is possible that these negative effects have compounding impact over time and may at least partly be responsible for the statistically significantly different performance of the female students in the calculus-based courses on the CSEM who agreed or disagreed with the stereotype at the end of the entire academic year physics sequence.

Finally, we note that the results of this investigation can be useful for designing professional development for instructors and TAs to help them make their classes more inclusive [10, 11]. Our data indicate that agreeing with the gender stereotype that men generally perform better in physics than women is correlated with decreased performance for female students on the CSEM at the end of the yearlong calculus-based course. Since one possible explanation of this finding is that female students who agree with the stereotype may experience increased stereotype threat compared to the female students who do not agree with the stereotype, TAs and instructors need to be careful to not propagate these types of stereotypes, both in their actions and statements. In particular, instructors and TAs should try to send the message to their students (both explicitly

and implicitly) that success in physics is primarily determined by effort and engaging in appropriate learning strategies rather than by something innate, e.g., gender (i.e., they should send the message that all students regardless of their gender can excel by effort and deliberate practice). In a book chapter titled "Is Math a Gift? Beliefs that Put Females at Risk" [12], Dweck argues that a fixed mindset (belief that intelligence is fixed or innate) is more detrimental to female students than male students. She describes a study in which two groups of adolescents were taught the same math lesson (which included historical information about the mathematicians who originated the ideas discussed in the lesson) in two different ways. For one group, the mathematicians were portrayed as geniuses and their "innate ability" and "natural talent" were highlighted, whereas for the other group, the mathematicians' commitment and hard work were highlighted. After the lesson, students were given a difficult math test and were told that the test would measure their mathematical ability. Female students who received the lesson which portrayed the mathematicians as geniuses performed worse than their male counterparts. On the other hand, for students who received the lesson which highlighted the mathematicians' hard work, there were no gender differences in performance. Dweck argues that when female students receive messages that mathematical ability is a gift, some of them may interpret that this gift is something they do not possess [12-14]. It is possible that accumulated societal stereotypes influence how female students interpret these messages and they may assume that if mathematical ability is a gift, male students are likely to have this gift, whereas they are not likely to have it. Therefore, it is important that professional development workshops for physics instructors and TAs focus on the findings of this research vis-à-vis other studies on stereotype threat [9, 17-19], and help instructors and TAs reflect upon the importance of encouraging their students to develop a growth mindset, namely that

intelligence is malleable and it can be cultivated with hard work and productive learning strategies regardless of gender or other characteristics (e.g., race/ethnicity) of an individual.

5.6 CHAPTER REFERENCES

1. D. Hestenes, G. Wells, and M. Swackhammer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

2. A. Traxler, X. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016);

   A. Madsen S. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

3. D. Maloney T. O'Kuma, C. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, s12 (2001).

4. S. Spence C. M. Steele, and D. M. Quinn, Stereotype threat and women's math performance, J. Exp. Soc. Psychol. **35**, 4 (1999).

5. S. Wheeler and R. Petty, The effects of stereotype activation on behavior: A review of possible mechanisms, Psychol. Bull. **127**, 797 (2001).

6. C. Steele and J. Aronson, Stereotype threat and the intellectual test performance of African Americans, J. Pers. Soc. Psychol. **69**, 797 (1995).

7. L. Strickler and C. Ward, Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance, J. Appl. Soc. Psychol. **34**, 665 (2004).

8. A. Maries and C. Singh, Stereotype threat? Effects of inquiring about test takers' gender on conceptual test performance in physics, AIP Conf. Proc. **1697**, 120008-1 (2015).

9. G. Marchand and G. Taasoobshirazi, Stereotype threat and women's performance in physics, Int. J. Sci. Educ. **35**, 3050 (2013).

10. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Routledge (1988).

11. A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. PER **14**, 010103 (2018).

12. C. Dweck, in *Why aren't more women in science? Top researchers debate the evidence, Washington DC, 2006*, edited by S. J. Ceci and W. Williams (American Psychological Association, Washington DC, 2006).

13. C. Dweck, Mindset: *The New Psychology of Success*, Random House Inc. (2006).

14. A. Bandura, Self-efficacy: Toward a unifying theory of behavioral change, Psychological Review **84** (2), 19 (1977).

15. T. Majors and P. Engelhardt, Gender & LEAP pedagogy: What does the Gender Force Concept Inventory have to say? PERC Proceedings p. 167 (available online) (2014).

16. R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).

17. M. Appel and N. Kronberger, Stereotypes and the achievement gap: Stereotype threat prior to test taking, Educational Psychology Review **24** (4), 609 (2012).

18. C. McKown and R. Weinstein, The development and consequences of stereotype consciousness in middle childhood, Child Development **74**, 498 (2003).

19. L. Bian, S. Leslie, and A. Cimpian, Gender stereotypes about intellectual ability emerge early and influence children's interests, Science **355**, 6323 (2017).

20. M. Lorenzo, C. Crouch and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74** (2), 118 (2006).

21. M. Cahill, K. Hynes, R. Trousil, L. Brooks, M. McDaniel, M. Repice, J. Zhao, and R. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST PER **10**, 020101 (2014).

22. J. Day, J. Stang, N. Holmes, D. Kumar, and D. Bonn, Gender gaps and gendered action in a first-year physics laboratory, Phys. Rev. ST PER **12** (2), 020104 (2016).

23. V. Coletta and J. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

24. V. Coletta, J. Phillips, and J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413,** 23 (2012).

25. S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch and N. Walet, Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison, Euro. J. Phys. **34** (2), 421 (2013).

26. Z. Y. Kalender, E. Marshman, T. Nokes-Malach, C. Schunn, and C. Singh, Motivational characteristics of underrepresented ethnic and racial minority students in introductory physics courses. Proceedings of the 2017 Physics Education Research Conference, 204-207 (2018). https://doi.org/10.1119/perc.2017.pr.046

27. T. Nokes-Malach, E. Marshman, Z. Y. Kalender, C. Schunn, and C. Singh, Investigation of male and female students' motivational characteristics throughout an introductory physics course sequence. *Proceedings of the 2017 Physics Education Research Conference*, 276-279 (2018). https://doi.org/10.1119/perc.2017.pr.064

28. A. Traxler and E. Brewe, Equity investigation of attitudinal shifts in introductory physics, Phys. Rev. ST Phys. Educ. Res. **11**, 020132 (2015).

29. Z. Hazari, G. Potvin, R. Lock, F. Lung, G. Sonnert, and P. Sadler, Factors that affect the physical science career interest of female students: Testing five common hypotheses, Phys. Rev. ST Phys. Educ. Res. **9**, 020115 (2013).

30. Z. Hazari, G. Sonnert, P. Sadler, and M. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**(8), 978 (2010).

31. Z. Hazari, P. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**(5), 82 (2013).

32. M. Besterfield-Sacre, M. Moreno, L. Shuman, and C. Atman, Gender and ethnicity differences in freshmen engineering student attitudes: A cross-institutional study, J. Eng. Educ. **90** (4), 477 (2001).

33. R. Felder, G. Felder, M. Mauney, C. Hamrin, and E. Dietz, A longitudinal study of engineering student performance and retention. III. Gender differences in student performance and attitudes, J. Engineering Education **84** (2), 151-163 (1995).

34. C. Moss-Racusin, J. Dovidio, V. Brescoll, M. Graham, and J. Handelsman, Science faculty's subtle gender biases favor male students, Proc. of the National Academy of Sciences of the United States of America **41**, 16474 (2012). https://doi.org/10.1073/pnas.1211286109

35. E. Marshman, Z. Y. Kalender, C. Schunn, T. Nokes-Malach, and C. Singh, A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences, Can. J. Phys. **96** (4) 391-405 (2017). https://doi.org/10.1139/cjp-2017-0185

36. N. I. Karim, A. Maries, and C. Singh, Do evidence-based active-engagement courses reduce the gender gap in introductory physics? Eur. J. Phys. **39**(2), 025701 (2018).

37. N. Abramzon, P. Benson, E. Bertschinger, S. Blessing, G. Cochran, A. Cox, B. Cunningham, J. Galbraith-Frew, J. Johnson, L. Kerby, E. Lalanne, C. O'Donnell, S. Petty, S. Sampath, S.

Seestrom, C. Singh, C. Spencer, K. Sparks Woodle, and S. Yennello, Women in physics in the United States: Recruitment and retention, AIP Conf. Proc. **1697**, 060045 (2015).

38. C. Seron, S. Silbey, S. Cech and B. Rubineau, Persistence is cultural: Professional socialization and the reproduction of sex segregation, Work and Occupations **43**(2), 178-214. (2016).

39. E. Seymour and N. M. Hewitt, *Talking about Leaving: Why Undergraduates Leave the Sciences*, Boulder, CO: Westview Press (1997).

40. H. Simon, *Models of Thought*, Vols. 1 and 2. Yale University Press (1979).

6.0 SUMMARY AND FUTURE DIRECTIONS

The studies discussed in this thesis can be extended in several ways. In Chapter 2, we investigated the pedagogical content knowledge (PCK) of teaching assistants (TAs) at identifying introductory physics students' difficulties in electricity and magnetism using the Conceptual Survey of Electricity and Magnetism (CSEM). In the past, the pedagogical content knowledge of teaching assistants has been investigated using the Force Concept Inventory (FCI) and the Test of Understanding Graphs in Kinematics (TUG-K). In the future, the PCK of TAs and/or instructors can be investigated using other standardized physics tests.

The main goal for investigating the PCK of TAs pertaining to a particular physics concept is to find out how well they know about students' difficulties in this area. This knowledge is very important for TAs because they need to know "where the students are" in order to determine "where to take them" using evidence-based curricula and pedagogies. In other words, PCK is necessary in order to create Piagetian 'Optimal Mismatch' [1] or to ensure that students are in Vygotsky's Zone of Proximal Development (ZPD) [2] using appropriate curricula and pedagogies. In the future, knowledge about PCK of TAs can be used for TAs' professional development to improve their PCK. It will also be useful to investigate how TAs (and instructors) take advantage of the PCK in helping students learn physics. Also, it will be useful to investigate how a TA's level of confidence and teaching experience is correlated with their overall PCK.

PCK is very important for implementing effective teaching strategies in an evidence-based active-engagement (EBAE) classroom. The impact of EBAE classes on student performance is discussed in Chapter 3. In that chapter, we have discussed the impact of EBAE classes by comparing students' performances in EBAE classes with those of traditional lecture-based (LB)

classes. We did observe improvement in students' performances in EBAE classes compared to those in LB classes on two standardized physics posttests (FCI and CSEM) but we did not observe as much improvement as one might expect from EBAE classes. In other words, we found that there is still room for improvement in implementation of EBAE pedagogies which can be studied in future. In particular, in Chapter 3, we briefly discussed possible reasons of not observing large improvements in the EBAE classes. The two big steps that can be taken for the improvement of EBAE pedagogies in the future are student buy-in (framing the purpose of such pedagogies appropriately to students so that they engage with it deeply) and instructor/TA training. Since EBAE pedagogies are relatively new and not used by many science instructors, students are not familiar with the EBAE teaching techniques which require them to be actively engaged in the learning process. Since learning requires effort, students may resist the fact that they have to do the thinking during the class. Therefore, they may expect TAs/instructors to lecture instead of expecting them to be involved in the learning process with the TA/instructors as their guide. More, we need instructors and TAs to properly implement the EBAE pedagogies (commensurate with their students' prior knowledge and skills) in order to obtain its full potential advantages. Therefore, instructor/TA training and student buy-in are two research areas that can be explored in future.

Chapter 4 also focuses on evidence-based active-engagement (EBAE) classes but here we explored its impact on gender gap. Whether (and what types of) EBAE classes help reduce the gender gap or not is still a subject not well-understood. Researchers like Lorenzo [3] and Beichner [4] have reported that active-engagement classes reduce the gender gap while other researchers including the study in Chapter 4 didn't. In our study, we used the Force Concept Inventory (FCI) and the Conceptual Survey of Electricity and Magnetism (CSEM) as pretest and posttest

instruments in the algebra- and calculus-based introductory physics classes, both in the EBAE classes and traditional lecture-based (LB) classes, to investigate whether gender gap is reduced. We found that the gender gap is maintained almost at the same level in the algebra-based classes but it increased a little in the calculus-based EBAE classes. This result may be due to the fact that some EBAE classes may have unintended effects for some students (e.g., those who are underrepresented or marginalized in some ways) and that there is a large room for future improvements in the EBAE classes. For example, the instructors and TAs in the EBAE classes can use interventions to reduce stereotype threat, increase the sense of social belongingness and self-efficacy, and help improve growth mindset [5] of the female (and other underrepresented) students. This may improve their performance and reduce the gender gap similar to the study of Lorenzo et al. [1]. Although such interventions are not common in physics classrooms, they can be integrated as a part of EBAE pedagogies in order to reduce the gender gap. However, focus on such interventions aiming at improving social belongingness, self-efficacy [6] and growth mindset is more important in EBAE classes since students often working in small groups. Therefore, EBAE classes investigated by future researchers can focus on the impact of such interventions within the EBAE pedagogies to reduce the gender gap.

In Chapter 5, we presented the effect of stereotype threat on gender gap including the fact that asking students to indicate their gender did not affect female students' performances on FCI or CSEM. We also found some evidence that female students who disagree with a statement about gender stereotypes in physics perform significantly better than female students who agree with the statement. Since interventions are extremely important for reducing stereotype threat and hence reducing gender gap, future investigations can focus on such interventions. Research suggests that such interventions focusing on improving students' social belongingness, self-efficacy and growth

mindset should always be subtle so that the students do not perceive it as an intervention [7] (otherwise its impact can be mitigated). There are many possibilities for future research along these directions in physics classrooms. For example, a short essay, a paragraph or a letter from several senior students (from all types of students – males, females, racial/ethnic minorities and majorities) describing how they had struggled but successfully completed the course with a good grade at the end by working hard and studying smart in physics might mean a lot to a student under threat and can encourage her to be less anxious about their physics class. There can be interventions to increase female (and underrepresented minority) students' sense of belongingness in the physics classroom, e.g., by having them realize that many students in their class (regardless of their demographical standing) face similar struggle with belonging and they are not alone. Moreover, future studies can also focus on the impact of interventions involving the growth mindset, e.g., that excelling in physics and math are not something that only male students can do because of their innate ability and that the brain is malleable like a muscle and everyone can become intelligence through systematic practice and hard work [5]. These types of interventions are widely used in social sciences but are still rare in the physics domain. However, the effectiveness of the interventions focusing on these issues and their effective implementation needs to be explored in the context of physics classrooms to increase all students' self-efficacy and confidence in learning physics so that they are not anxious about spending the time to engage deeply in learning physics and can use their cognitive resources optimally while learning physics.

6.1 CHAPTER REFERENCES

1. H. Ginsberg and S. Opper, Piaget's theory of intellectual development. Englewood Cliffs, NJ, Prentice Hall (1969).

2. L.S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes, Cambridge, MA, Harvard University Press (1978).

3. M. Lorenzo, C. Crouch and E. Mazur, Reducing the gender gap in the physics classroom, Am J Phys **74** (2), 118 (2006).

4. R. Beichner, J. Saul, D. Abbott, J. Morse, D. Deardorff, R. Allain, S. Bonham, M. Dancy and J. Risley, The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project, in *Research-Based Reform of University Physics*, edited by E. Redish and P Cooney (published online) (2007).

5. C. Dweck, Mindset: The New Psychology of Success. Random House Incorporated (2006).

6. L. Aguilar, G. Walton and C. Wieman, Psychological insights for improved physics teaching Phys. Today **67**, 43 (2014).

7. G. L. Cohen and J. Garcia, Educational theory, practice, and policy and the wisdom of social psychology, Policy Insights from the Behavioral and Brain Sciences **1**(1), 13 (2014).