Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario

Engaging Students to Think Critically in a Large History Class

Mairi Cowan, Tyler Evans-Tokaryk,
Elaine Goettler, Jeffrey Graham,
Christopher Landon, Simone Laughton,
Sharon Marjadsingh, Caspian Sawczak
and Alison Weir
University of Toronto Mississauga

Published by

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada, M5E 1E5

Phone:	(416) 212-3893
Fax:	(416) 212-3899
Web:	www.heqco.ca
E-mail:	info@heqco.ca

Cite this publication in the following format:

Cowan, M., Evans-Tokaryk, T., Goettler, E., Graham, J., Landon, C., Laughton, S., Marjadsingh, S., Sawczak, C., & Weir, A. (2014). *Engaging Students to Think Critically in a Large History Class.* Toronto: Higher Education Quality Council of Ontario.

Higher Education
Quality Council
of Ontario

The opinions expressed in this research document are those of the authors and do not necessarily represent the views or official policies of the Higher Education Quality Council of Ontario or other agencies or organizations that may have provided support, financial or otherwise, for this project.

Acknowledgments

The authors of this report would like to express their gratitude for the help received from University of Toronto Mississauga Department of Historical Studies research assistants Crystal Rosbrook, now a BEd student at the Ontario Institute for Studies in Education; Peter N. Stavros, now pursuing a JD at Harvard Law School; and Shameran Zaya, now an MT student at the Ontario Institute for Studies in Education. These three students assisted ably and cheerfully with literature reviews, preliminary analyses of qualitative data, and administration of the pre- and post-intervention tests. We would also like to thank Asal Aslemand, who recently completed a MSc in statistics at the University of Toronto and who will soon be starting a PhD at the Ontario Institute for Studies in Education, for her assistance with clarifying the research design at the pilot stage, as well as HEQCO research consultants Susan Elgie and Ruth A. Childs, Associate Professor at the Ontario Institute for Studies in Education, for their guidance in improving the focus of this exploratory study.

Table of Contents

List of Tables

List of Figures

Executive Summary

Engagement Strategies Yield Mixed Results in a Large History Course

In their efforts to foster active engagement in the classroom, instructors are increasingly looking to integrate instructional technologies such as online quizzes and clickers into their large courses. While studies of STEM (science, technology, engineering, and mathematics) education have demonstrated that such approaches have the potential not only to enhance the quality of students' learning experiences generally, but also to help improve their critical thinking skills specifically, much less is known about the effectiveness of instructional technologies in humanities education. This exploratory study seeks to add to our understanding of pedagogical best practices in the humanities by testing the efficacy of engagement strategies in a history course. One main finding of this study is that the adoption of a cluster of engagement strategies similar to those used in physics education did help develop the critical thinking skills of some students in a large first-year history course, but not always to a greater extent than more conventional approaches to instruction.

Project Description

The present study considered two different groups of students in a first-year history course at the University of Toronto Mississauga (UTM). It measured students' ability to think critically in the discipline-specific task of selecting appropriate sources for historical research, one of the stated learning objectives for the course. Students in the intervention group received online quizzes each week as an incentive to complete their assigned readings before class, as well as clicker questions during lecture to help them to recall their reading, activate prior knowledge, and practise newly acquired skills. The control group was taught the same course material for the same amount of time but without the online quizzes and clicker questions. Both groups were tested on the degree to which they had mastered the critical thinking skill of identifying and selecting sources for historical research.

Methods

Students enrolled in the Fall 2012 section of HIS101 Introduction to Historical Studies were the intervention group, and students enrolled in the Winter 2013 section of the same course were the control group. The same instructor taught identical material in both sections, but employed a suite of engagement strategies for the intervention group. The primary instrument for assessing students' improvement in critical thinking skills was a pre-intervention test administered in week 2 and a post-intervention test in week 6. Other means for assessing student improvement included a writing assignment and a set of questions on the final examination.

The differences between pre-intervention and post-intervention test scores were partitioned into five ordered categories, labeled -2 (None), -1 (Slight), 0 (Some), 1 (Good), and 2 (Very Good), where 0 is the middle group, defined by the quintiles of students' levels of improvement. Statistical analysis of the data collected through these instruments was conducted using the statistical software package R. Ordinal logistic regression was used to model the probabilities that students would fall into one of five Improvement categories, given that they either had or did not have the intervention of engagement strategies and whether they fit into a combination of other variables such as academic performance in the course (according to final grade, partitioned into quintiles) and year of study. Log-linear models were used to determine whether there is a relationship between students' attitudes to the technological interventions and their success in the course.

Findings

The engagement strategies that can be effective at improving students' critical thinking skills in other disciplines had mixed results in a large history course. For most students in the different improvement categories, there was no significant difference between those who had the intervention and those who did not. An exception was the 1 or "Good" improvement category in terms of academic performance. Students in this improvement category improved the most when using the engagement strategies. A similar result was found for the direct-entry first-year students. The students in the 2 or "Very Good" improvement category, and those who already had some experience in university, by contrast, actually improved more in a conventional classroom (without engagement strategies) than in one with clickers and online quizzes. This finding suggests that some types of students had a higher probability of greater improvement *with* the engagement strategies, while some other types of students had a higher probability of greater improvement *without* them.

To measure the longer-term impact of the intervention, students' improvement from the pre-intervention test to the post-intervention test was compared to their improvement from the pre-intervention test to the final examination. The differences between the control group and intervention group were not, however, statistically significant.

A supplementary analysis was conducted to determine whether the intervention had an impact on students' ability to use their critical thinking skills in a more authentic context. Researchers assessed students' ability to select an appropriate source while conducting research for a writing assignment, and determined whether their performance could be predicted by their improvement from the pre-intervention test to the post-intervention test. Again, there was no statistically significant relationship between the pre- and post-intervention tests and the writing assignment in either the control group or the intervention group.

An interesting and unanticipated finding of this exploratory study was a dramatic decline in attendance for students in the control group. Average attendance in lecture throughout the term in the intervention group was 78% of enrolled students, but it dropped to an average of just under 50% for the control group when the clickers and online quizzes were excluded from the course. In previous years (when these engagement strategies were a part of the course), the attendance rate had been between 70% and 80% of enrolled students.

The study also found that there was no relationship between students' evaluations of the engagement strategies and individual student learning outcomes. Students who liked the strategies were neither more nor less likely to improve than those who did not like them. This finding has implications for the development of education quality indicators, as it suggests that student experience or satisfaction measures do not necessarily relate to academic success or to student learning, and thus should be considered separately from, and should not be mixed in with nor substituted for, assessment of learning outcomes.

Data on an array of different variables for each student (e.g., age, gender, declared major, number of credits, number of transfer credits) were collected and analyzed, but none of these had a statistically significant effect on student improvement.

Recommendations/Further Research

An intriguing but preliminary finding of this exploratory study was an interactive effect between the pre-intervention test and tutorial marks for the control group: students with higher tutorial marks in the control group improved their critical thinking skills the most. It is not clear whether this relationship is correlational or causal, but the result suggests that further research on the impact of small tutorials on the development of students' critical thinking skills would be helpful. Also helpful would be more research on why students found certain kinds of questions on the pre- and post-intervention test more difficult than others; on the longer-term impacts of these engagement strategies; and on the efficacy of these engagement strategies in teaching other skills. Perhaps most importantly, research should be done comparing students using variations of this research design (e.g., in the fall semester of one year against the fall semester of the next, or with the same subset of students in a first-year course, or a cohort of students across multiple years) rather than fall and winter semesters of the same academic year. This could help determine if the results of this study were affected by more subtle aspects of a first-semester effect and whether there may be longer-term impacts of technology-integrated pedagogies on student learning.

Introduction

This exploratory study investigates how students in a large introductory history course at the University of Toronto Mississauga (UTM) develop their critical thinking skills through active engagement with course material. Inspired by the close connection between practice and pedagogy in STEM (science, technology, engineering, and mathematics) fields, we draw upon the literature about physics education in particular and science education more generally to provide humanities instructors with information about how we can best help students in large classes to think critically in a discipline-specific context. In order to find and interpret this information, we have been forging cross-disciplinary alliances in our research team among specialists in the fields of history, writing studies, psychology, statistics, information studies and academic counselling.

We considered students in the two different sections of the Introduction to Historical Studies course (HIS101) during the 2012-2013 academic year at UTM. In the fall semester, for the intervention group, we employed a cluster of instructional strategies aimed at helping students become engaged in class and use critical thinking skills to solve authentic problems in history research. These strategies included online quizzes each week as an incentive for students to complete their readings before coming to class, as well as “clicker questions” in lecture (questions posed during lecture to which students responded by using their student response system, or “clickers”) to get students to recall what they had read, activate prior knowledge either from within or from outside the course, and practise the application of newer skills to already-learned material. In the winter semester, for the control group, we covered the same course material but we took a more conventional approach in its delivery (no online quizzes, no clickers). In both semesters, we measured the level to which students had mastered the critical thinking skill of identifying and selecting sources for historical research. Our instruments for measuring this skill were pre- and post-intervention tests, a writing assignment that asked students to select an appropriate source for a specific area of historical research, and a series of questions on the final examination.

Research Questions

Our main research questions have been developed with a practical application in mind, namely, to provide the instructors of large humanities classes with data and guidelines that will help them make informed decisions about what to include in lectures and how to deliver the lecture material.

Motivated particularly by a desire to provide instructors with information about whether and to what extent instructional strategies aimed at increasing student engagement can help to develop their students’ critical thinking skills, researchers in this study began by seeking answers to the following question:

1. Can the cluster of strategies for active engagement that is increasingly used in physics education improve the critical thinking skills of humanities students?

As we began to observe our results, we developed the following additional questions:

2. If these strategies do help improve students’ critical thinking skills, is their effectiveness constant across all levels of student ability, year of study, or previous exposure to history courses?
3. Is there a positive relationship between students’ level of satisfaction with the strategies and the extent to which their learning is enhanced by these strategies?

Literature Review

Considerable research has been conducted into the consequences of large class size on student learning outcomes in postsecondary education, the effectiveness of specific instructional technologies to offset some of the disadvantages or perceived disadvantages of large classes, and methods for developing critical thinking skills among students. With a shortage of research into the intersection of these areas in the context of history education specifically – on whether instructional technologies can improve the development of critical thinking skills among undergraduate students in large history classes – our research team has drawn heavily from studies in fields other than history education.

A primary area of inspiration for our exploratory study has been the field of physics education research, where work by practising physicists has led to the development of some radical departures from traditional pedagogy (Adams et al., 2006; Deslauriers, Schelew & Wieman, 2011; Koenig, 2010; Redish, 2003). A similar intervention to that of our project was recently undertaken in a large-enrolment first-year physics class at the University of British Columbia and reported in the journal *Science* (Deslauriers, Schelew & Wieman, 2011). The researchers in this study compared the learning achieved during a unit about electromagnetic waves from a traditional lecture format (control group) to that achieved through an approach that included pre-class reading assignments, pre-class reading quizzes, in-class clicker questions, small-group active learning tasks, and targeted in-class instructor feedback (experimental group). The academic performance of students in the experimental group at UBC was considerably better than that of students in the control group, strongly suggesting greater improvement with the intervention. The authors conclude that “use of deliberate practice teaching strategies can improve both learning and engagement in a large introductory physics course as compared with what was obtained with the lecture method”, and add that “this result is likely to generalize to a variety of postsecondary courses” (p. 864). Another study similar to ours was performed at the University of Colorado Boulder (Knight & Wood, 2005), in which the authors attempted to determine whether student learning gains in a large lecture course in developmental biology could be increased by partially changing a traditional lecture format to include greater interaction. Their experimental section was designed to include more student participation and cooperative problem-solving, as well as frequent in-class assessment of understanding. Knight and Wood find that “even a partial shift toward a more interactive and collaborative course format can lead to significant increases in student learning gains”, including in students’ development of skills necessary for solving conceptual problems (p. 304).

Nothing quite like these studies in physics and biology education has been undertaken in history education. History educators have considered the current state of history education among students and the general public (Barton, 2004; Friesen, Muise & Northrup, 2009), and they have also made suggestions for how to improve historical thinking through history instruction at the elementary and secondary levels (*Historical Thinking Project*, 2011; Osborne, 2003; Peck & Seixas, 2008; Seixas & Morton, 2012), but the field of history education research has tended to be the preserve of scholars in faculties of education rather than departments of history. Professional historians ought to be involved in developing the scholarship of teaching and learning of history both because their disciplinary expertise is essential for determining what students require and what faculty expect, and also because their input in the creation of pedagogical knowledge is necessary if this knowledge is ever to be widely applied in university history classes (Huber & Morreale, 2002; Reichard, 2006). Yet, in spite of calls for more participation by historians in the field of history education research, practising academic historians remain underrepresented in the scholarship of teaching and learning (Cowan & Landon, 2011; Pace, 2004).

Large Classes

Awareness has been growing among instructors and policy-makers of the important place that large classes occupy in many undergraduate curricula, as has concern for the impact that large class sizes might have on the learning outcomes of students (Keirle & Morgan, 2011; Kerr, 2011). The threshold beyond which a class is defined as “large” varies greatly in the literature, however, from as low as 40 students (Tolley et al., 2012), to just over 50 (Chapman & Ludlow, 2010), to 80 (Moulding, 2010), to greater than 100 (Johnson & Robson, 2008), to 200 (Mandel & Süssmuth, 2011), to more than 550 (Exeter et al., 2010), to 800 (Snowball & Boughey, 2012). Some studies avoid choosing any threshold at all, preferring instead to look at increased enrolment in a relative sense or at levels of student involvement (Hunt, 2012; Johnson, 2010; Keirle & Morgan, 2011). Overall, the definition of a “large” class is subjective, and can be expected to differ based on discipline, institution, and the expectations and aims of individual instructors.

Instructional Technologies

Anecdotal instructor experience confirmed by research in the field indicates that there are many undergraduate students who do not complete the course readings assigned to them and that this contributes to their falling behind in their course work (Berry et al., 2010; Burchfield & Sappington, 2000; Hatteberg & Steffy, 2013; Hobson, 2004; Jolliffe & Harl, 2008). One possible way of encouraging students to do more of their course readings is to have them write short critical reviews (Saltmarsh & Saltmarsh, 2008), but assessing these reviews often requires resources beyond those available in a large course with only one instructor and hundreds of students. Bean (2011) recommends a number of student-focused strategies, such as reading guides and modeling the reading process in lecture or tutorial, but these strategies seem insufficient when confronted with the fact that over 70% of first-year students across the disciplines are not completing their assigned reading (Hobson, 2004). Perhaps a more promising approach is to incentivize reading by embedding random reading comprehension quizzes into lectures (McDougall & Cordeiro, 1993) or by requiring students to complete low-stakes online quizzes before class.

The results of several studies suggest that online quizzes have the potential to help scaffold student learning by providing immediate feedback, encouraging preparation outside of class, promoting understanding, affording opportunities for self-assessment and reflection, and fostering proactive engagement with course material (Anthis & Adams, 2012; Gikandi, Morrow & Davis, 2011; Metz, 2008). Some studies have also shown that online quizzes are an effective strategy for reinforcing pre-lecture reading assignments, though it has been suggested that they are most valuable when used in the early weeks of a semester as a means of establishing good study habits (Balter, Enstrom & Klingenberg, 2013; Stull et al., 2011).

Another engagement strategy that has been found to be effective at the tertiary level is the use of clickers (sometimes also called “student response systems”). In higher education classrooms, clickers are used with the aims of facilitating learning strategies, supporting formative assessment, and enhancing student-faculty interactions (Brewer, 2004; James, 2006). They may have the potential to improve student learning, especially because they increase the immediacy of feedback for the learner and allow for more opportunities for student participation (Bloemhof & Christensen Hughes, 2013; Bartsch & Murphy, 2011; Britten, 2011; Denker, 2013; Fortner-Wood et al., 2012; Mayer et al., 2009; Poole, 2012; Stowell et al., 2010; Trees & Jackson, 2007; Zurmehly & Leadingham, 2008) and perhaps, more basically, because they increase student attendance (Britten, 2011; Poole, 2012). Several instructors additionally note the positive aspect of receiving information in real time about student performance, which lets them adjust their instructional pace accordingly (Bartsch & Murphy, 2011; Britten, 2011; Zurmehly & Leadingham, 2008). Whatever the advantages of clickers, they also have limitations, either because no improvement or only slight improvements have been found in student learning among those groups using the clickers (Britten, 2011; Fortner-Wood et al., 2012;

Matus et al., 2011), or because the adoption of clickers into an already-running class requires a considerable time investment on the part of the instructor (Zurmehly & Leadingham, 2008), even if the adoption is being “patched in” to traditional lecture courses (Brewer, 2004; Johnson & Robson, 2008). Recent meta-analyses of published research indicate that the most common challenges instructors face when adopting clickers are related to the technology (e.g., the system or students’ remotes do not work properly) or the time commitment required for writing questions and tracking response data (Caldwell, 2007; Kay & Lesage, 2009). For a comprehensive summary of recent research documenting challenges instructors face when adopting clickers in their courses, see Table 2 in Gok, 2011.

Leger et al. (2013) have studied efforts to promote better engagement and deeper learning in a large first-year geography class by using both online quizzes and clicker questions, and their research yields some interesting results. The authors found that clickers had a positive effect on student engagement and student attendance, but also that students frequently cheated on their attendance by having friends borrow their clickers. Students themselves praised the opportunities afforded by clickers to practise their skills in class, but they also offered somewhat contradictory comments on the value of this practice, with some saying that it did motivate them to complete readings before class, some saying that it did not motivate them to do their work, and some calling it rote learning. Student opinion on the value of clickers and weekly online quizzes in promoting engagement was almost evenly split (Leger et al., 2013).

Only a few studies have looked at clickers in large history classes in particular, and these tend to be impressionistic and focused on the personal insight of the instructor (Britten, 2011; Cole & Kosc, 2010). This research suggests that clickers might augment student participation and thereby also improve student learning, but additional studies could help to provide more information regarding optimal use of this type of technology to support teaching and learning activities in the context of history education specifically. Collaborative and active-learning activities in the classroom have been shown to have a positive impact on the development of students’ critical thinking skills in disciplines other than history (Gokhale, 1995; Johnson, Johnson & Smith, 2007; Jones, 2006; Nelson, 1994; Olivares, 2005; Schamber & Mahoney, 2006; Springer et al., 1999, all cited in Lo, 2010).

Critical Thinking Skills

The development of critical thinking skills is commonly recognized as an important goal of postsecondary education. At the University of Toronto Mississauga specifically, one of the undergraduate degree level expectations is that students demonstrate the competency of “critical thinking and analytical skills inside and outside the discipline” (University of Toronto Mississauga, n.d.). Various models have been proposed for defining and assessing these skills both in general postsecondary education and also in discipline-specific contexts (Behar-Horenstein & Niu, 2011; Golding, 2011), but the very concept of “critical thinking” remains, at least for some researchers, frustratingly ill-defined. The central yet insufficiently examined problem of defining “critical thinking” is articulated succinctly by Jennifer Wilson Mulnix, who observes a “widespread disagreement as to what critical thinking actually is or amounts to”, in spite of its recurrent inclusion within current pedagogical efforts:

> In a climate where colleges and universities are increasingly demanding that their faculties instill critical thinking skills in undergraduate students, it is imperative that we begin to think critically about this concept. Yet…what counts as ‘critical thinking’ seems to vary widely. (Mulnix, 2010, p. 464)

It seems, then, that more research needs to be done in order to develop pedagogical theories and approaches that support the development of learners’ critical thinking skills in discipline-specific contexts.

While this report does not attempt to wade very far into the debate concerning definitions of “Critical Thinking”, it does recognize the importance of identifying and fostering specific skills that the authors think are necessary for students to think critically about history. In particular, this report considers the skills associated with selecting sources for historical research by focusing on how well students understand what is being asked in a research question and on how clearly students understand that different types of historical sources are appropriate for different kinds of historical questions. (Additional information on the skills measured in this study can be found below in the section on “The Instruments” and in Appendix 3 “Sample Questions from the Pre- and Post-Intervention Test”.)

The importance of fostering critical thinking in history courses specifically has been recognized, even if a concrete and shared definition of “critical thinking” remains elusive (Crenshaw, Hale & Harper, 2011; Peace, 2010; Warren, Memory & Bolinger, 2004). Regardless of how critical thinking is defined, the challenge of promoting the range of “higher order thinking skills” involved in the discipline of history has been flagged as a particular challenge in large undergraduate classes (Keirle & Morgan, 2011). Yet observers of pedagogical developments in the field of history education have noted that very little work has been done on how best to foster critical thinking skills among history students or on how to measure instructional effectiveness in history classes (Chapnick, 2010; Lévesque, 2008; Pace, 2004; Seixas, 2006; Seixas et al., 2005). One possible reason for this gap is that the discourse of history education has become too highly infused with the assumed duality of “knowledge” (often referred to as “course content”) and “skill”, when in fact these domains are closely connected and perhaps even interdependent (Barton, 2004; Counsell, 2002). Some researchers have been challenging the supposed content/skill duality by looking at which mental operations are implicitly expected of students in history courses and pointing to differences between novice and expert strategies in various aspects of studying the past (McKeown & Beck, 1994; Perfetti et al., 1994; Wineburg, 2001). Our study links content and skills instruction by measuring a skill that is required for students to understand the more sophisticated, often ambiguous, “content” of history: identifying and selecting appropriate sources for historical investigation.

Methods

The Large Class

Our test course was HIS101 Introduction to Historical Studies at the University of Toronto Mississauga (UTM). This course is taught each year in both the fall (September-December) and winter (January-April) semesters. With an enrolment cap of 350 students each semester, it is the largest history course at UTM. HIS101 is a program requirement for students in history, history of religions, and classical civilizations, meaning that all students seeking a specialist, major, or minor in these programs must complete the course before graduating; because this course is neither a formal prerequisite nor even recommended preparation for most historical studies upper-year courses, however, students may, and sometimes do, take it in their second, third, or even final year of university study.

The official course description for HIS101 on the syllabus is as follows:

> This writing-intensive course introduces Historical Studies through a variety of exercises that will allow students to read models of good writing and to practise the integration of successful strategies into their own work. In both lectures and tutorials, students will have the opportunity to try different tools and approaches for developing the skills useful at every stage of the creative process from pre-writing and preliminary research through to editing and undergraduate publication.

The official learning objectives for HIS101 on the syllabus are that by the end of the course, students should be able to:

1. Select and read primary and secondary sources critically to support historical investigation;
2. Apply the stylistic conventions of a variety of historical writing genres; and,
3. Write clearly and persuasively on historical topics.

Pedagogical Goals

The two main pedagogical goals of HIS101 are to help students (1) learn some basic content in world history prior to the year 1800; and (2) develop research and writing skills specific to the discipline of history. These pedagogical goals are designed to get students thinking like historians, rather than merely thinking about history. Therefore, students are encouraged to consider the connections between research, writing, and basic historical content throughout the course. For their weekly readings, students are assigned chapters from a world history textbook to cover basic content, a section from a writing guide to cover a theoretical approach to the development of a particular skill, and an academic article that illustrates how that skill is actually used by a professional historian. In their lectures, students are presented with instruction in both historical content and historical skills. For example, in the week described in the syllabus as “Mesoamerica and South America; Selecting Sources”, students hear about the Spanish conquest of the Aztec Empire and also about which primary and secondary sources are the best to choose for that topic. In their tutorials, students are guided through the practical application of the week’s skill using that week’s content. For example, in the week on “The Mediterranean; The First Draft”, students practise the skill of crafting a good thesis statement using a series of laws from ancient Rome.

The interconnection of content with skill in HIS101 is further reinforced in the course’s assignments, which include a primary source study and revised primary source study. Students learn from these assignments about the history of the society that produced the primary source, which is historical content, along with the importance of using both primary and secondary sources in historical research and the value of careful revision in light of formative feedback from teaching assistants, which are basic skills.

Overview of General Methods

Our study ran during the 2012-2013 academic year with the same instructor teaching the same material in both semesters but using different instructional techniques in each. Course evaluations following both semesters indicated that students found the course environment conducive to their learning and that they considered the course instructor to be knowledgeable in her field, approachable, and enthusiastic in her delivery of lecture material.

In the fall semester, for the intervention group, we included a cluster of engagement strategies in the form of online quizzes each week before lecture and clicker questions in lecture. Each online quiz was available during the entire week leading up to its associated lecture, and students could take the quiz as often as they liked during the week, with only their highest score being counted toward their grade. Collectively, quizzes were worth 10% of the final grade in the course, and the lowest quiz grade was dropped. The online quizzes contained questions in various forms, including multiple choice, multiple answer, fill in the blanks, and matching of columns. All questions drew on weekly readings, and the online system provided students with feedback on their answers. Some questions were focussed on “content”, others on “skills.” See Appendix 1 for examples of online quiz questions.

During the fall semester lectures, we also used the clickers in every lecture. As in the most prominent examples of physics education research, our instructional design for these lectures was based on the concept of "deliberate practice" for the development of expertise. Deliberate practice involves the invention of specific tasks to overcome weakness, and the careful monitoring of performance to provide cues for further improvement. As Ericsson et al. note, deliberate practice "is a highly structured activity, the explicit goal of which is to improve performance" (1993, p. 368). In our project, this deliberate practice involved presenting on an overhead screen authentic historical questions and tasks that required the students to use their clickers and employ historian-like critical thinking during class time while being provided with frequent feedback. Collectively, clicker questions were worth 5% of the final grade for the course, and we dropped the lowest two weeks' worth of responses. Clicker questions served several purposes, including simple polling of student opinion, assessment of how well the students understood the assigned readings or earlier lecture material, and consolidation of material recently delivered in lecture. For this study, perhaps the most important purposes of clicker questions were to gauge how well students understood a concept at a specific point in time so that the instructor could adjust her pace accordingly, and to allow students the opportunity to solve problems for themselves or with the assistance of peer instruction in order to practise their historical skills on relevant content. See Appendix 2 for examples of clicker questions.

In the winter semester of HIS101, for the control group, we presented the same material using a more conventional approach. Instead of online quizzes, we simply informed students that they were expected to complete all readings before coming to class. Instead of clickers, the instructor paused in her lectures to pose the same questions as those that had been used in the fall semester, but sought volunteers to answer them. For example, in the fall semester, the instructor introduced the distinction between "primary sources" and "secondary sources" through a slide asking, "Which of the following is a definition for 'primary source' according to *Writing History*" followed by five options; students used their clickers to answer and the instructor confirmed that 95% of the students were correct. In the winter semester, the instructor presented a slide that simply stated "A definition for 'primary source', according to *Writing History*" but was otherwise blank; she asked for a student to volunteer the correct answer. The first student to raise his hand offered a generally correct response that the instructor clarified before presenting an additional slide with the formal definition. In both cases, less than a minute was spent answering the question. In order to help ensure that the same material was being covered for the same amount of time in both semesters, the instructor had a colleague sit in on the lectures to time the question sessions and ensure that they were of comparable duration and composition. During the fall semester, this colleague made note of how long the instructor spent on each section of the lecture. During the winter semester, the instructor had these notes in front of her so that she could time each section as precisely as possible to match the fall.

On average, the instructor used the same amount of time to teach material in the fall semester with clicker technology as she did in the winter semester with standard question-and-answer techniques. The pacing was somewhat different with some clicker questions in the fall semester requiring either more or less time to pose and answer than the same questions asked and answered verbally in the winter semester, but ultimately the lectures in both semesters covered precisely the same material and were completed within the same two-hour time period.

The Instruments

We measured students' mastery of one fundamental critical thinking skill for historical studies, namely, how to identify and select appropriate sources for historical research. Ethics approval was obtained from the University of Toronto Office of Research Ethics. We used several different instruments to assess this mastery, including a pair of pre- and post-intervention tests, a writing assignment and a series of questions on the final examination.

Pre- and Post-Intervention Tests

The pair of pre- and post-intervention tests was our main instrument for measuring critical thinking. We piloted a draft of this test with students in HIS101 during 2011-2012 to ensure that our foils were well chosen, and then we adjusted our final version of the test in response to student performance on the earlier version. The final version of the test consisted of 27 skill-related questions and additional user confidence questions. We created two sets of questions (identical in form but using different specific sources) and randomly allocated half the questions from each set to the pre-intervention test and half to the post-intervention test. We had graduate students in history and professors of history take the final version of our test to verify that all questions were answerable by experts.

The pre-intervention test was administered at the start of lecture in week 2, before the skill was taught in the course, and the post-intervention test was administered at the start of lecture in week 6, after the skill had been taught. Instruction of the skill was conducted exclusively through course readings and lectures; tutorial assistants did not teach this skill in tutorials during these weeks, focusing instead on other aspects of the course. We deliberately excluded tutorials from this study because we wanted to measure the effectiveness of teaching strategies in large lectures, even when resources were not available for additional small-group instruction. An announcement was made at the start of lecture in week 2 and then again in week 6 to tell students that we had prepared for them an exercise that would help us know what they currently understood about using sources in historical research. Students who were present for the pre-intervention test were told that if they took the test they would receive 0.5% added to their final grade at the end of the semester, and students who were present for the post-intervention test were told that if they improved upon their result in the pre-intervention test they would receive 0.5% added to their final grade at the end of the semester. The response rate (those who wrote both the pre-intervention test and the post-intervention test) across both semesters was 61.9%.

For both the pre-intervention test and the post-intervention test, students were given 40 minutes to write the test, a time limit determined by piloting the test with graduate students and history professors and then doubling the amount of time for the undergraduate students. The tests were written under standard test conditions, which were enforced by a faculty member from the academic skills centre and two research assistants. The course instructor left the room while the tests were being written.

The pre-intervention test and post-intervention test were made up of multiple choice questions about source usage, mostly in the form of case studies. The questions were organized into eight categories asking students to:

- Choose a primary source for a historical situation that is real, but not covered in HIS101 by week 6;
- Choose a primary source for a hypothetical (unreal, but realistic-sounding) historical situation;
- Choose a secondary source for a historical situation that is real, but not covered in HIS101 by week 6;
- Choose a secondary source for a hypothetical (unreal but realistic-sounding) historical situation;
- Choose a primary source for a historiographical essay about how people in a specific historical period viewed an earlier historical period (i.e., an investigation into the history of history);
- Identify what kind of source is listed in a real reference;
- Identify what kind of source is listed in a hypothetical (unreal but realistic-sounding) reference; and,
- Indicate the confidence level and metacognitive understanding of their performance.

We used real examples to test the students’ ability to apply critical thinking skills in authentic historical situations, and we used hypothetical examples to ensure that students would not be able to bring prior knowledge of the historical context to bear on the question. We predicted that the questions about historiography would be the most challenging, because they require students to apply a more complex set of critical thinking skills that need to interact on two different levels: students must be able to understand that they are looking for a primary source about the study of a specific area of history rather than a primary source about the specific area of history itself. See Appendix 3 for sample questions from the pre- and post-intervention tests.

The pre- and post-intervention tests were not identical, but two versions of each of the 27 skills-related questions were created with the same form for their respective categories and with only the details changed. In each category, one or the other alternative question was included on the pre- or post-intervention test as determined by random selection using a coin toss. The post-intervention test included an additional set of questions asking students to indicate which lectures they had attended and which readings they had completed thus far in the course.

Writing Assignment and Exam Questions

The other instruments used to measure students' understanding of how to select sources were a writing assignment and a set of questions on the final examination. The writing assignment was a revised primary source study (RPSS), the second major writing assignment of the semester. It followed upon a primary source study (PSS), in which students chose a primary source from a list, and then used it, as the assignment sheet says, to "uncover and analyze some aspect of the society in which the source was produced." Research beyond this one primary source was neither required nor encouraged, and all the evidence used in the PSS was to come directly from this primary source. The RPSS was an opportunity for students to improve upon their PSS by revising it in accordance with comments made by the TA and resubmitting their work. This assignment also required students to write a reflection on what they had learned in the revision process and identify a secondary source that would help them better understand the primary source or the historical context in which it was written. Students were instructed to include a full bibliographic entry for the chosen secondary source, and an explanation for why this secondary source would help them understand the primary source. This source selection component of the assignment was graded by the instructor according to criteria for the good selection of secondary sources introduced in lecture (i.e., that the source be secondary, relevant, scholarly, and recent) and given a score out of 4 using the following rubric:

> Secondary: If the assignment listed any kind of secondary source, it was given 1 point. If the assignment listed a primary source with a very small amount of secondary material, it was given 0.5 of a point. If the assignment listed a primary source or no source at all, it was given no point.
>
> Relevance: If the source was relevant to the topic (either the primary source being considered or the focus the student had chosen to write about in the PSS), it was given 1 point. If the source was marginally relevant (far too broad, or in the right general area for the topic but about a different aspect), it was given 0.5 of a point. If the source was irrelevant to the primary source and the focus of the PSS, it was given no point.
>
> Academic Calibre: If the source was scholarly, it was given 1 point. If the source straddled the boundary between scholarly and unscholarly (e.g., a trade publication by an academic historian), it was given 0.5 of a point. If the source was unscholarly, it was given no point.
>
> Age: If the source was published in the year 2000 or more recently, it was given 1 point. If the source was published between 1970 and 1999, it was given 0.5 of a point. If the source was published before

> 1970, it was given no point. Exceptions were made in cases where the assignment cited a seminal work older than 2000 or 1970 but that is still current in the field.

In all instances, if the source appeared at first glance to be less than ideal in a category but the student made a compelling case for using the source, the full point for that category was awarded. For example, a student selecting a secondary source to complement her study of the letters of Marie de l'Incarnation, a nun in seventeenth-century Quebec, wrote:

> This article, though emphasizing a slightly different central theme than my own paper, provides evidence and insights which prove surprisingly relevant to my arguments and analysis; this is the first reason why this article would be useful to my research. As well, since the article was analyzing the same set of letters, it provides ample historical context about them specifically.

Students' responses to 12 questions on the final examination also were reviewed as part of this study. These were multiple choice questions taking the same form as questions on the pre- and post-intervention tests. Students in both semesters wrote the final examination for the course approximately two months after the post-intervention test.

Statistical Methodologies

The statistical analysis began after data collection was completed.

Statistical analysis was conducted using R (R Core Team, 2013). R is an open source statistical software package that consists of a base package for standard methodologies and a wealth of contributed packages for more sophisticated statistical methodologies. Ordinal logistic regression modeling, testing, and analysis were conducted using R's ordinal package (Christensen, 2012), R's MASS package (Venables & Ripley, 2002) and R's vcd package (Meyer, Zeileis & Hornik, 2013).

Ordinal logistic regression (Hosmer, Lemeshow & Sturdivant, 2013) is a statistical methodology that models the impact of predictors on a response that is measured on an ordinal scale. Unlike a linear model, an ordinal logistic regression model cannot be used to predict the response at specific values of the independent variables (predictors). Instead, ordinal logistic models estimate the probabilities associated with each level of the ordinal response at specific values of the independent variables (predictors). There are no required distributional assumptions for ordinal logistic regression. In our exploratory study, we estimated the probabilities that a student will fall into each of the five Improvement categories, given that they either had or did not have the intervention and that they fit into some combination of all the other variables described in the previous section. Graphs were created using the Minitab statistical software package (Minitab 16 Statistical Software, 2010).

Data and Analysis

Participants

A total of 498 students completed HIS101 during the 2012-2013 academic year. 280 students completed the course in the fall semester, while 218 completed it in the winter semester. A total of 297 students participated in the study (meaning that they signed survey consent forms, completed both the pre-intervention test and the post-intervention test, and earned a non-zero final mark in HIS101.) Students enrolled in the Fall 2012 section (n=211) of HIS101 Introduction to Historical Students were the intervention group, and students enrolled in the Winter 2013 section of the same course (n=86) were the control group.

Response Rates and Semester Means

As Table 1 indicates, there was a substantial difference between the response rates in the fall and winter semesters. The fall response rates ranged from 72.5 to 75.4%, whereas the winter response rates ranged from 38.1 to 39.4%. The disparity between response rates raised questions for the researchers about any biases that may have been introduced. In addition, there was a statistically significant difference in the pre-intervention test scores. In order to accommodate the disparity in response rates and pre-intervention test scores, additional statistical analyses were performed and are reported below and in the appendices. The response rate for the revised primary source study (RPSS) also was higher at 56.1% for the fall semester (n=156) compared to 30.7% (n=62) for the winter semester.

Table 1: Summary Statistics for Student Assessments, with Response Rates and p-Values, for Testing Equality of Associated Semester Means. All marks are percentages.

Variable	Semester	Sample Size (n)	Response Rate (%)	Median	Mean	St. Dev.	*p*-value
Pre-Intervention Test Mark	Fall	211	75.4	26.0	29.2	14.7	≈ 0
	Winter	86	39.4	35.2	35.7	15.3	
Post-Intervention Test Mark	Fall	211	75.4	51.8	50.3	19.3	0.013
	Winter	86	39.4	55.6	55.9	19.9	
Final Examination Mark	Fall	203	72.5	68.0	66.8	13.1	0.961
	Winter	83	38.1	67.0	66.7	10.2	
Final Course Mark	Fall	211	75.4	72.9	70.6	11.1	0.138
	Winter	86	39.4	69.6	68.4	11.8	

Initial Analysis: Linear Models

Two Independent Sample t-Test

A simple assessment of our research hypothesis is a t-test for equality of the two semesters’ mean differences between pre-intervention test and post-intervention test marks. Our data provide no evidence in support of the fundamental research hypothesis (*t* = 0.42, *df* = 154 *p* = 0.667).

Figure 1: Post-Intervention Test Mark minus Pre-Intervention Test Mark, Split by Semester (Fall semester with technological interventions and Winter semester without technological interventions)

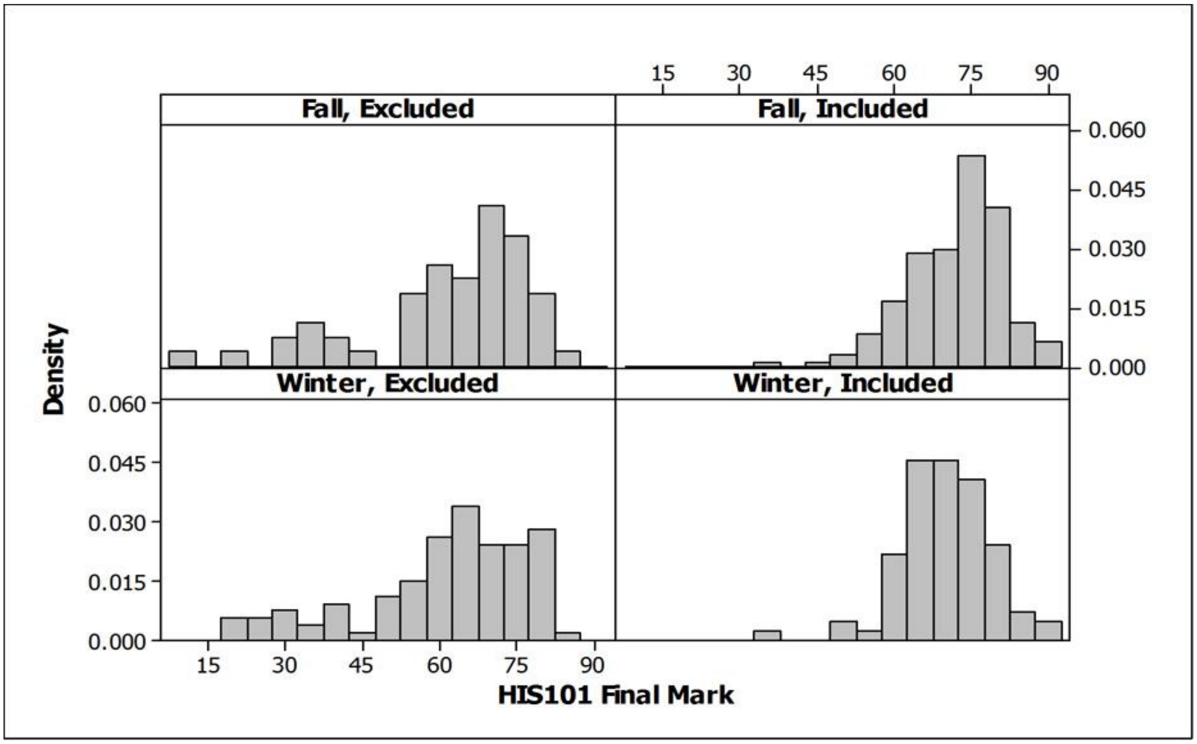


Figure 1 supports the conclusion of the t-test; there is no significant difference between the locations of the two distributions. However, Figure 1 also shows differences between the shapes of the two distributions: the winter semester looks bimodal whereas the fall semester looks quite flat over the range 13 to 40. This suggests that additional variables may need to be investigated to improve focus on the differences between the two semesters. We also notice a potential outlier in the winter semester.

In addition, we found other indications that the fall semester and winter semester could not be compared using tests based on summary statistics (means, medians, and standard deviations) and that more sophisticated tests might be required to adequately explore the effects of the Intervention. For a more detailed explanation, see Appendix 4.

Research Variables

Response Variables: Improvement and Improvement4

Each student’s post-intervention test mark was subtracted from the pre-intervention test mark. The 297 differences between pre-intervention and post-intervention test scores were partitioned into five categories defined by their quintiles (see Table 12 in Appendix 4 for ranges of mark differences in these categories). The decision to create five categories for the differences between pre-intervention and post-intervention test scores was not arbitrary, since the five levels of improvement are a rough mapping of the letter grades A, B, C, D, and Fail.

The five categories were labeled -2, -1, 0, 1, and 2, where 0 is the middle category (described as “Some” in Table 2 below). We called this ordinal variable Improvement, and used it as the response variable in our models. The lowest Improvement group, labeled -2, represents either negative improvement or virtually no positive improvement from the pre-intervention test to the post-intervention test. The remaining Improvement categories, labeled -1, 0, 1, and 2, represent increasing magnitudes of positive improvement from the pre-intervention test to the post-intervention test.

Table 2: Improvement Variable Categories

Improvement Variable	
Label	**Meaning**
-2	None
-1	Slight
0	Some
1	Good
2	Very Good

This categorization of student improvement has the advantage of removing the effects of pre-intervention test and post-intervention test outliers and unequal improvement variances.

To investigate the combined effects of the intervention and first-year status, we needed to create a new variable for Intervention. A new Improvement variable, Improvement4, was created by partitioning the 297 differences between pre-intervention and post-intervention test scores into four categories defined by their quartiles. We coded the levels of Improvement4 as -2, -1, 1, and 2 (see Table 3). This was done to compensate for the low response rate and small sample size (*n*=86) in the winter semester, which precluded our adding first-year status to a model with five Improvement categories and five Mark categories. By reducing the number of categories, it was then possible to use first-year status as a predictor of improvement.

Table 3: Improvement4 Variable Categories

Improvement4 Variable	
Label	**Meaning**
-2	None
-1	Small
1	Intermediate
2	Large

Predictor Variables: Intervention, Year1, First Year, Mark and Mark4

Intervention

In this study, students either received the intervention (fall semester students) or they did not (winter semester students).

Year1 and First Year

Two predictor variables were created for first-year student status. One of these indicates first-year status as defined by the Office of the Registrar: fewer than 4.0 completed full-year course equivalents (HIS101 is a half course) and a cumulative grade point average of at least 1.5 (on a 4 point scale). In other words, students with fewer than 4.0 completed full courses are called “first year" even if they have already spent a year or more at university. We called this variable Year1. The second first-year status indicator variable indicates that the student attended some form of high school in the academic year 2011-2012 (i.e., they were in high school during the year immediately preceding their time in HIS101). We called this latter variable First Year.

Mark and Mark4

In our model, we could not use the pre-intervention test mark as a predictor of improvement as the pre-intervention test scores indicated that the starting point was quite different for students in the fall semester compared to students in the winter semester. The mean pre-intervention test mark in the fall semester was 29.2%, and the mean pre-intervention test mark in the winter semester was 35.8%. As such, it would have been necessary to omit the main effect of the pre-intervention test mark on improvement, and instead use the interaction of the pre-intervention test mark and the intervention on improvement. Thus the pre-intervention test effect would have been confounded with the effect of the intervention. We therefore used the final mark in HIS101 as a pseudo-proxy for the pre-intervention test mark. Pre-intervention test marks measured students’ knowledge of the topics prior to instruction and experimentation, while final marks measured students’ ability in HIS101. Since there were no significant differences between the transformed final marks in the two terms, this substitution admitted a model in which the main effect of the intervention could be isolated. In addition, this substitution is allowable since the skills measured in the pre-intervention test and post-intervention test comprised less than 3% of the final course mark.

For the students included in our study, the distributions of final marks in HIS101 are not the same in the two semesters (see Figures 4 and 5 in Appendix 4). We standardized the final marks within each semester by centering each set of marks at zero and scaling each set of marks to standard deviation one. We then combined both sets of standardized marks, and they formed a single homogeneous group (see Appendix 4). We then partitioned the 297 standardized HIS101 final marks into five categories defined by their quintiles (see Table 4 and Appendix 4). One set of quintiles was calculated using all n = 297 students from both semesters of the study. This categorization produced five categories, labeled -2, -1, 0, 1, and 2, where 0 is the middle group. We have called this variable Mark; it is an ordinal measure of performance in HIS101. Table 4 below provides the Mark variable categories.

Table 4: Mark Variable Categories

Mark Variable	
Label	**Meaning**
-2	Lowest
-1	Low
0	Middle
1	High
2	Highest

To investigate the combined effects of the intervention and first-year status, we needed to create a new variable for Mark. A new Mark variable, Mark4, was created by partitioning the 297 standardized HIS101 final marks into four categories defined by their quartiles (see Table 5). We coded the levels of Mark4 as -2, -1, 1, and 2. This was done to compensate for the low response rate and small sample size (*n*=86) in the winter semester, which precluded our adding first-year status to a model with five Improvement categories and five Mark categories.

Thus, both Improvement4 and Mark4 were created because the low response rate and small sample size (*n*=86) in the winter semester precluded adding first-year status to a model with five Improvement categories and five Mark categories. By reducing the number of categories, it became possible to introduce first-year status as a predictor of improvement.

Table 5: Mark4 Variable Categories

Mark4 Variable	
Label	**Meaning**
-2	Poor
-1	Fair
1	Good
2	Very Good

Additional Variables Tested

While not central to our investigation, we suspected that many other variables would either account for variation in improvement (and thus improve focus on the intervention's effect), and/or interact with the intervention (and thus provide textured conditional conclusions about the intervention's efficacy). We collected data on and tested the following variables:

- Gender
- Age
- Year of study (as defined by the Office of the Registrar)
- Credit count (up to and including HIS101 and any courses the student took concurrent with HIS101)

- Number of clicker courses taken concurrent with HIS101
- Number of clicker courses taken prior to HIS101
- Number of history courses completed
- Number of non-history courses completed
- Total number of courses completed
- Total number of courses dropped
- Total number of transfer credits
- Total number of courses passed
- Number of history courses dropped
- Number of history courses passed
- Declared major (grouped into broad classes)

There are other variables that we would have liked to include in the study but could not since we did not have access to the data. Some of these other variables include, but are not limited to, student use of and comfort levels with technology, student perceptions regarding the use of technology in the classroom, and student self-perceptions regarding usefulness of the technologies to support their learning.

Data Analysis

We modeled the response, Improvement, using ordinal logistic regression (an extension of tests on binomial proportions) and applied the logit transform to Improvement for a proportional odds response. Our predictor of interest was the indicator for Intervention, and our model selection processes included all the additional predictors discussed in the previous section (e.g., year of study, number of history courses completed). Model selection was conducted using a backward elimination procedure, starting with a saturated model that contained all main effects and two-way interactions. Model selection progressed by consideration of individual coefficient z-tests, odds ratio confidence intervals, null and model deviances, global likelihood ratio χ^2 test, Akaike's Information Criterion (Akaike, 1974), predictor non-collinearity χ^2-tests, and diagnostic residual plots. Following a suggested guideline for exploratory research in education (Huberty, 1987), main effects and interactions with p-values below 0.15 were deemed to have a statistically significant effect on the response, and thus such terms were retained in the model. If an interaction had a significant effect on the response, then that interaction and the corresponding main effects were retained in the model regardless of the significance level of the interaction's component main effects. A main effect is the overall impact of a single predictor variable on the response. A two-way interaction is present if the impact of one predictor variable on the response is not constant over all possible values of another predictor variable.

A further constraint had an impact on our model selection process. Categorizing improvement into five groups and categorizing final marks into five groups resulted in some of the 25 combinations, particularly those involving the non-intervention class (winter semester), being sparsely populated. As such, a complex model such as this one runs the risk of being an over-fitted model. See Appendix 4 for details.

Results

Research Question 1

As noted above in the introduction, the primary research question for this exploratory study is:

1. Can the cluster of strategies for active engagement that is increasingly used in physics education improve the critical thinking skills of humanities students?

Thus, the hypothesis for the primary research question for the study is:

H_a: Student improvement in critical thinking skills, measured as the difference between post-intervention test and pre-intervention test scores, is higher with the cluster of technological interventions that appear effective in physics than it is without these interventions.

The corresponding null hypothesis is:

H_o: Student improvement in critical thinking skills, measured as the difference between post-intervention test and pre-intervention test scores, is no different with the cluster of technological interventions that appear effective in physics than it is without these interventions.

Measuring Students' Critical Thinking Skills

In this study, student improvement in critical thinking skills was measured as the difference between post-intervention test and pre-intervention test scores. As mentioned above, the differences between pre- and post-intervention test scores were partitioned into five Improvement categories (see Table 2). The model we selected as optimal for estimation of the probabilities that a student will fall into each of the five Improvement categories contains significant effects from HIS101 Mark and the interaction of HIS101 Mark with Intervention (Likelihood Ratio $p \approx 0$, Null Deviance $p = 0.402$.) Ordinal logistic regression was used to model the probabilities associated with each of the five Improvement categories as a function of the five Mark categories (the row labels in Table 4) and the two levels of Intervention (presence and absence). In the process of model selection, all of the previously mentioned predictors were considered.

The most compelling finding, as shown below in Table 6, occurred for students in Mark group 1 ("High"). In this group, there was a significantly higher probability of Slight-to-None Improvement (i.e., categories -2 or -1; $p = 0.072$) and a significantly lower probability of Good Improvement (category 1) without the intervention than with the intervention ($p = 0.051$). In other words, students with high, but not the highest, marks showed substantial improvement when the intervention is present.

The model showed that, regardless of Intervention status, students with lower HIS101 Marks were more likely to have low improvement levels than they are to have high improvement levels. More nuanced conclusions also can also be drawn as noted below.

Table 6: Combinations of HIS101 Final Marks and Improvement Groups that have significantly different improvement probabilities with and without the Technological Interventions

HIS101 Mark Group	Improvement Group	Probability		p-value	
		***No* Intervention Winter Semester**	**Intervention Fall Semester**		
-2	1	0.227	0.147	0.064	**
-1	-2 or -1	0.677	0.570	0.064	**
1	-2 or -1	0.478	0.309	0.072	**
1	-1	0.21	0.157	0.088	**
1	1	0.186	0.274	0.051	**
2	-2 or -1	0.148	0.249	0.128	*
* indicates significance at p = 0.15, ** indicates significance at p = 0.10					

Other results include findings for students who were in the lowest and highest Mark groups. For students in Mark group -2, there was a significantly higher probability of Good Improvement (i.e., category 1) without the intervention than with the intervention (*p* = 0.064), meaning that students with the lowest marks were actually *less* likely to show substantial improvement when the intervention is present. In the Mark group 1, there was a significantly higher probability of None-to-Slight Improvement (-2 or -1) without the intervention than with the intervention (*p* = 0.064). Finally, in Mark group 2, there was a significantly lower probability of None-to-Slight Improvement (-2 or -1) without the intervention than with the intervention (*p* = 0.128). These last two findings suggest that, when the intervention is *not* present, students with low, but not the lowest, marks were *more* likely to show minimal improvement, while students with the highest marks are *less* likely to show minimal improvement.

Interventions also appear promising for mid-level students (i.e., Mark categories -1, “Low”; 0, “Middle”; and 1, “High”), although statistical validation is limited. Of the ten hypothesis tests for the intervention’s effect on above average Improvement probabilities, only two showed significance at the 15% level, and of the ten hypothesis tests for differences in below average Improvement probabilities, four showed significance at the 15% level. There were no other significant differences between the fall and winter probabilities for average Improvement within any of the five Mark categories, yet this summary constitutes more statistically significant results than would be expected by chance. See Appendix 4 for complete details.

Source Selection in the Revised Primary Source Study Assignment

The revised primary source study (RPSS) assignment included a component that asked students to choose a secondary source that would help them better understand the primary source that they had earlier studied. We were interested in whether performance on this assignment could be predicted by performance on the pre- and post-intervention tests. Ordinal logistic regression models were fit for each of the four sub-components of the marking rubric (*secondary*, *relevance*, *calibre*, and *age*) and for the total RPSS score, but none of the predictors entered into the models was significant. However, there was a significant difference in the overall response rates for the RPSS assignment between the fall and winter semesters, in that a lower percentage of students completed the assignment in the fall semester (75.8%) than in the winter semester (88.4%) (*p* = 0.006).

Research Question 2

Our second research question is as follows:

2. If these strategies do help improve students' critical thinking skills, is their effectiveness constant across all levels of student ability, year of study, or previous exposure to history courses?

As mentioned in the Methods section, the cluster of technological interventions that appears promising in physics education was applied in the fall semester (intervention group). To investigate the combined effects of the intervention and first-year status (First Year), we created new variables for Intervention and HIS101 final Mark as explained above. Our low response rate and small sample size (*n*=86) in the winter semester precludes the addition of first-year status in the larger analysis above. However, by reducing the number of categories for HIS101 marks and Improvement, it became possible to use first-year status as a predictor of improvement and address our second research question.

Ordinal logistic regression was used to model the probabilities associated with each of the four Improvement4 categories as a function of the four Mark4 categories, the two levels of Intervention (presence or absence), and the two levels of First Year (presence or absence). The optimal model for Improvement4 contained significant effects for the interaction of First Year with Intervention and the interaction of First Year with Mark4. Students who had "Good" or "Very Good" (category 1 or 2) marks, who had also attended high school immediately prior to their first year at university, were likely to show greater improvement. Although the interaction effects had weak statistical significance for these two interactions, when the categories were coarsened the main effect of the intervention was not significant. This also supports the finding that the intervention was helpful for a small, specific subset of the entire class.

We found that, for non-first-year students, within each of the four Mark4 categories, there were virtually no differences in Improvement4 between the fall and winter semesters (with and without the Intervention, respectively). However, this is not the case for first-year students.

We also found that, both with and without the intervention, first-year students in the lower two Mark4 categories (-2, "Poor"; and -1, "Fair") are more likely to show little or no improvement (i.e., to be in Improvement4 categories -1, "Small"; or -2, "None") than they are to show intermediate or large improvement (i.e., to be in Improvement4 categories 1, "Intermediate"; or 2, "Large"). But this effect is more pronounced in the fall semester (F*) than in the winter semester.

Finally, we found that, both with and without the intervention, HIS101 first-year students in the upper two Mark4 categories (1, "Good"; and 2, "Very Good") are more likely to show intermediate or large improvement (i.e., to be in Improvement4 categories 1, "Intermediate"; or 2, "Large") than they are to show little or no improvement (i.e., to be in Improvement4 categories -1, "Small"; or -2, "None"). But this effect is more pronounced in the winter semester (W*) than in the fall semester. See Table 16 and Figure 12 in Appendix 4 for further information on the effects of First-Year status.

Our analysis included numerous other covariates and concomitant variables as indicated above; none of these variables had a statistically significant effect on student improvement.

Credit counts, year of study, and age were generally unrelated to any variable in the study except for engagement tasks in the fall (but only with clickers, and not quizzes). The total number of clicker attempts correlated negatively, albeit weakly, with both credit count (Spearman's rho = - 0.18, *p* = 0.01) and year of study (Spearman's rho = - 0.177, *p* = 0.011). This suggests that, in a class containing both first-year and upper-year students, the upper-year students were less likely to use clickers than the first-year students.

Table 7 highlights a few of the differences in the compositions of the fall semester and the winter semester classes. In Appendix 4 we show that of these four differences, only First Year status (i.e., definition #1) had a significant impact on our measure of student learning. The limitations of this finding are provided in our Discussion section below.

Table 7: Summary Percentages for Student Demographics, with Response Rates and p-Values, for Testing Equality of the Associated Semester Proportions

Variable	Semester	Sample Size	Response Rate	Percent	p-value
Has completed at least one other UTM history course prior to completing HIS101.	Fall	211	75.4	6.6	0.005
	Winter	86	39.4	19.8	
Has used at least one late withdrawal from a UTM course (withdrawal after a two week course shopping period).	Fall	211	75.4	8.1	0.001
	Winter	86	39.4	24.4	
First year status definition #1 (First Year): registered in an Ontario high school 2011/12 and registered at UTM in September 2012.	Fall	211	75.4	75.8	0.007
	Winter	86	39.4	59.3	
First year status definition #2 (Year1): according to the office of the registrar, < 4.0FCE completed and/or CGPA < 1.5	Fall	188	67.1	83.0	≈ 0
	Winter	82	37.6	61.0	

Another interesting observation that may be worth exploring in further analyses is that, for this study, student age and other course demographics were uncorrelated to student performance in the course or to the pre- and post-intervention test scores in the fall semester.

In the winter semester, by contrast, the credit count was positively related to the total pre-intervention test score (Spearman's rho [n=85] = .259, *p* = 0.017), to the pre-intervention test metacognition self-ranking (Spearman's rho [n=85] = .275, *p* = 0.011), and to the total score on the set of related questions on the final examination (Spearman's rho [n=82] = .264, *p* = 0.017).

Research Question 3

Our third research question is as follows:

3. Is there a positive relationship between students' level of satisfaction with the strategies and the extent to which their learning is enhanced by these strategies?

Student Satisfaction regarding Technology and Student Learning

Three questions on the fall (intervention group) course evaluation survey administered at the end of the semester asked students to assess their belief in the intervention's effectiveness at improving their success in HIS101. We grouped these beliefs into two categories: positive and negative/neutral. We also grouped both Improvement and final Mark quintile defined classes into three categories: below median (originally -2 and -1), median (0) and above median (originally 1 and 2).

The large *p*-values in Tables 18, 19, and 20 provided in Appendix 4 indicate that there is no evidence of a relationship between improvement and students' attitude to classroom technology. In other words, for this study it seems that students' belief in the efficacy of classroom technology is unrelated to their academic success.

Additional Results

Engagement and Attendance

An unanticipated development in our exploratory study was a dramatic decline in attendance during the winter semester (see Table 8 and Figure 2), a decline that was much more severe than in the fall. In week 2 of the fall semester, there were 294 students in class, which is 89% of the total enrolled (332) in the course at that time. In week 2 of the winter semester, there were 200 students in class, which is 66% of the total enrolled (301) in the course at that time. In week 6 of the fall semester, there were 242 students in class, which is 82% of the total enrolled (295) in the course at that time. In week 6 of the winter semester, there were 112 students in class, which is 44% of students enrolled (253) in the course at that time. By week 10, there were 202 students in class during the fall, which is 71% of the total enrolled (285), and 94 students in class during the winter, which is 42% of the total enrolled (226). In week 11 of the fall semester, there were 199 students in class, 70% of the total enrolled (285); in week 11 of the winter semester, there were 102 students in class, 45% of the total enrolled (225).

In 2011-2012, when the engagement strategies of online quizzes and clickers were used in both fall and winter, attendance remained fairly consistent in both semesters (see Figure 3).

Table 8: HIS101 Class Attendance over the Past Two Years

Fall Semester	% of Class, Averaged over all Weeks	Winter Semester	% of Class, Averaged over all Weeks
2012	78.0	2013	49.25
2011	79.9	2012	74.8

Figure 2: HIS101 Class Attendance in 2012-2013

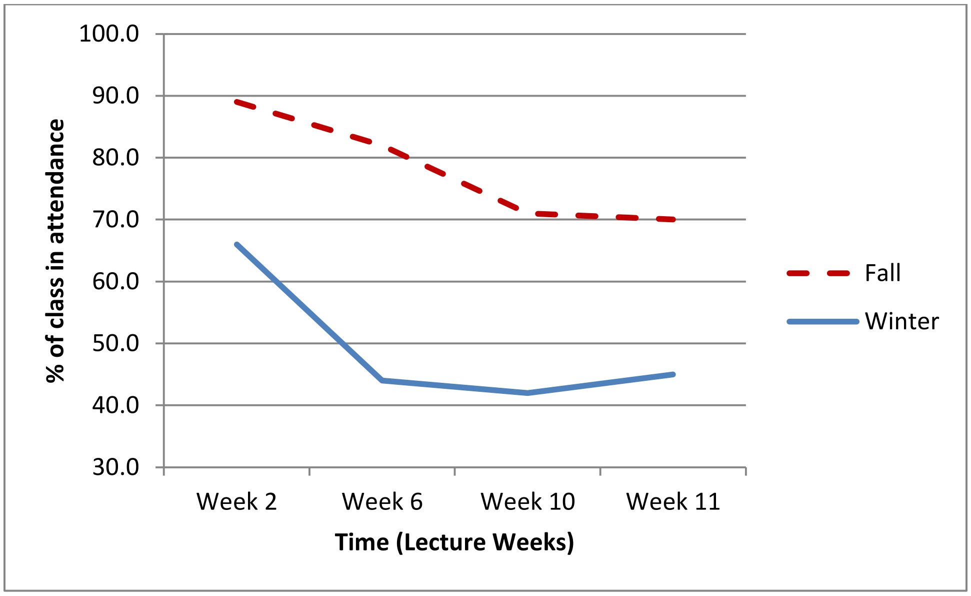
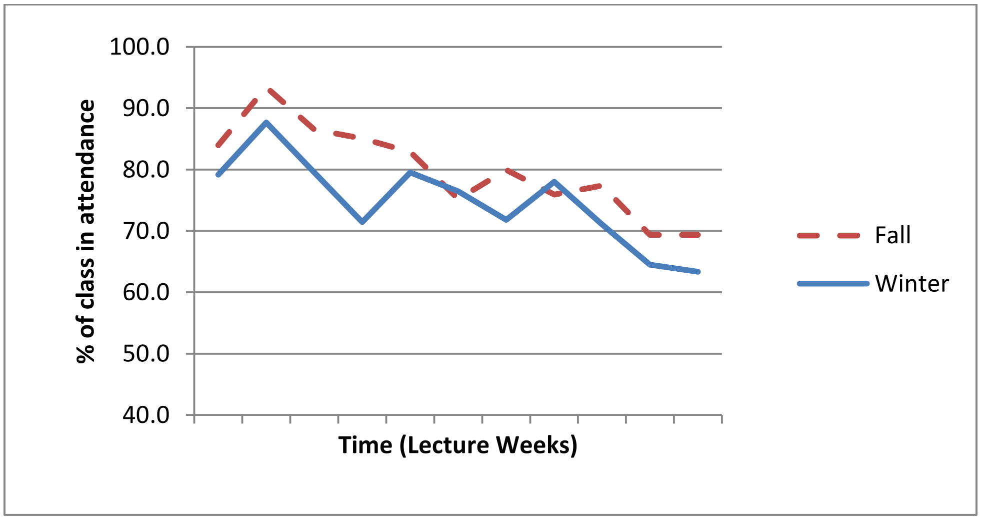


Figure 3: HIS101 Class Attendance in 2011-2012



Participation rates in online quizzes during the fall semester were generally very high (see Table 17 in Appendix 4). Over 96% of students in the fall semester completed at least one of the three quizzes that were relevant to the study, and the median student made an average of 2.3 attempts per quiz. The amount of time spent on quizzes may serve as an additional measure of engagement; when summing the time spent on initial quiz attempts only, the median student spent a total of 2.8 hours on quizzes over the course of the semester.

Preliminary Error Analysis of Pre- and Post-Intervention Tests

A preliminary analysis of exactly which foils students chose on the pre- and post-intervention tests has revealed some interesting consistent errors in the fall and winter semesters. The most common error on primary source questions was that students chose the wrong source type (i.e., they chose a secondary source when they were asked for a primary source), which happened in about 25% of all cases when students were asked to identify or choose a primary source. The most common error on questions asking students to choose a secondary source was that they chose a source that was about the wrong time or place, which happened about 27% of the time. This error was also predominant in questions where students were asked to identify a secondary source, happening about 20% of the time. On historiography questions, the most common error was that students chose a source that was about an irrelevant topic, which happened about 30% of the time. Moreover, some types of questions proved more difficult for students than others. Historiography questions were the most difficult (only about 30% of students chose the correct response), and the questions asking students to identify the kind of source from information provided in a bibliographic entry were the easiest (between one-half and two-thirds chose the correct response). The primary and secondary source application questions were in the middle (44.7% to 53.9% chose the correct response).

Discussion and Limitations of the Study

Engagement Strategies and the Critical Thinking Skills of Students in HIS101

This exploratory study found that students in HIS101 Introduction to Historical Studies improved their critical thinking skills in the area of identifying and selecting sources for historical research both with and without the engagement strategies of online quizzes to encourage reading before lecture and clicker questions to foster active engagement with authentic tasks during lecture. Although their marks on the final examination suggest similar abilities among students in both the fall and winter semesters, it is certainly possible, even likely, that students in the fall differ from students in the winter for a variety of reasons, including timetabling requirements and enthusiasm for the subject. In spite of such possible differences between the two groups, we have been able to draw some conclusions from the study. No statistically significant difference was observed between the intervention and control groups as a whole in terms of how much the students improved, but it was observed that the effectiveness of engagement strategies did differ across levels of academic achievement.

The students most likely to benefit from using the engagement strategies being tested were the somewhat above average students (those in the 1 or High Mark group). These are perhaps the students who need to work diligently at their tasks in order to succeed and who therefore benefit from an additional or an unconventional instructional approach that allows them to rehearse a skill in real time under expert guidance. Whether these students were helped most by peer instruction, scaffolding of material, repeated opportunities to test new knowledge acquired from answering questions in light of immediate instructor feedback, or merely more regular active engagement in lecture, is difficult if not impossible to say at this point.

The top students in HIS101 (those in the 2 or Highest Mark Group) improved considerably both with and without the use of engagement strategies, but they were more likely to show greater improvement without them. This pattern may be a reflection of student adaptability or aptitude: very strong students have done well in their academic careers and can probably do well under a wide variety of circumstances, but they are likely to excel in the context most familiar to them, that is, the context of conventional teaching (straight lecture without pre-reading quizzes or clicker questions) either because this is how they have learned to learn (i.e., they have been trained to learn this way) or because they are the students who are somehow predisposed to benefit from conventional instruction (i.e., they have achieved high grades because the teaching style normally used is the teaching style that suits them best).

There is some evidence of an interactive effect between First Year status and Intervention on Improvement, possibly suggesting that having at least one semester of experience at university has a greater positive impact on students' ability to learn than does this suite of instructional strategies. In terms of year of study, the students who showed the most improvement were typical first-year students in the winter, i.e., students who had completed high school the previous spring and who were taught using a conventional approach. By this point in the academic year, these students had had a full semester to acclimatize to university expectations, so it is perhaps not surprising that they were able to improve in a standard lecture format. The students who showed the least improvement, by contrast, were non-first-year students, i.e., students who had already been at university for at least a year. These were students who had had at least a year to acclimatize to university expectations, so it is perhaps not surprising that they did not improve (or did not improve much). They may also be a group of students that is somewhat less engaged with learning at university in general when compared with other students in the course, since they are students waiting until after their first year is complete before taking the only first-year history course offered in the department. Among the students in the higher Mark categories, it is the first-year students who showed the greatest improvement. Among students in the lowest Mark group, first-year students in the fall showed very little improvement, perhaps reflecting a poor adjustment from high school to university at the very start of their university careers.

Evaluation of Active Engagement Impact: Separating Student Learning from Student Attitudes

While active engagement typically has a positive impact on students' learning, an interesting finding from this study is the fact that there was no relationship between student evaluations of our active engagement strategies and individual student learning outcomes. Thus, just because a student likes a specific component of the instruction, it does not mean that student learning has occurred. This result suggests that for the purposes of evaluating the effectiveness of active engagement strategies, it may be necessary to separate "assessment of learning" (whether students have learned) from students' attitudes towards and evaluation of the learning process (whether they enjoyed the experience).

It is possible that the decline in attendance during the winter semester is a reflection of a relative lack of engagement and diminished appreciation of what was being taught when compared to the fall. If this is the case, however, it is difficult to know whether students were more engaged in the fall than they were in the winter because of the specific engagement strategies employed, because of students being more engaged in their fall courses than in their winter courses generally, or because of the marks associated with lecture participation in the fall. In the fall semester, 5% of the final grade came from lecture participation. Although this percentage is small, especially since the lowest two classes' worth of marks were dropped when calculating the final grade, nevertheless, it may have provided sufficient incentive for students to come to class even if they were not engaging particularly deeply with the material.

Limitations of the Study

We were aware of potential problems arising from the comparison of students in two different semesters for our study, especially since students in the winter semester would generally have had one more semester of experience in university when compared with students in the fall semester. We felt constrained to use this design, however, both because of the timeline for our project (two years, with the academic year of 2011-2012 for our pilot phase and the academic year of 2012-2013 for our quasi-experiment) and because of limitations for when sections of the course could be scheduled with the same instructor (the curricular priorities of the Historical Studies department determine that the course needs to be offered each fall and winter, and the teaching load of the course's instructor limits the number of sections of this course to one per semester).

Since we averaged over all student characteristics over numerous unmeasured variables, our findings may be overly coarse, and we are aware of the weak statistical relationships between the intervention and the outcomes. It is possible that our results would be refined, and possibly strengthened, by including student survey data on attitudes toward history and history education, attitudes toward technology and classroom technology, and various other demographics in the modeling process.

Conclusion

The adoption of online quizzes to encourage pre-reading before lecture and of clicker questions to encourage problem-solving during lecture did help develop students' critical thinking skills in HIS101 Introduction to Historical Studies, but so too did a more conventional approach to instruction. The strongest students in the class (Mark group 2) and those who had already gained some experience in the university environment (non-first-year students) actually improved somewhat more in the conventional setting where they were instructed to have their readings completed ahead of time, but were not given online quizzes, and where they attended lectures delivered with pauses for questions by the instructor, but did not use clickers. Above-average students (Mark group 1) and first-year students in their very first semester at university improved more when using the engagement strategies tested in this exploratory study.

In other words, students' historical and critical thinking skills improved both through conventional lecture and also through lecture enhanced by active engagement strategies. Overall ability in the course and experience at university thus far contributed to differences between the proportions of students who improved more under one set of conditions than the other.

Instructors who are thinking about adopting engagement strategies such as online quizzes and clickers should consider several factors before deciding if this is the right choice for their courses. First of all, since the engagement strategies do not appear to benefit all students equally, instructors should think about the particular students in their courses. If instructors are working at an institution where first-year courses contain "true" first-year students, then these engagement strategies look very promising. For a course populated with a significant number of students who already have some experience with conventional university courses, however, the instructor may wish to continue providing conventional reading assignments and lectures. Instructors may also wish to consider the academic strengths of the students in their courses. In a situation where an instructor is most concerned with improving the skills of above-average (though not necessarily the very top) students, then the engagement strategies examined here appear very promising. If the instructor is most concerned with improving the skills of the very best students, then conventional reading assignments and lectures may be better suited to their goals. In short, online quizzes and clicker questions help good students get even better, and they help first-year students make a smoother transition from high school to university; for such purposes, these engagement strategies look very desirable from a pedagogical standpoint.

Another factor that instructors should consider when making decisions about adopting new engagement strategies is the significant investment in money and time that is required for implementation. Instructors should weigh potential gains in student learning against a realistic assessment of what resources are needed from both students and instructors in order to implement the engagement strategies effectively. While even modest gains generated by the engagement strategies are, of course, desirable, these would come at a cost to students and instructors. Not only would students have to ensure that they have regular and reliable access to computers and the internet in order to take the quizzes, but they would also have to purchase or rent clickers in order to participate in lecture. At the University of Toronto bookstore, clickers currently sell for $41.95-$44.20 when new and $33.60-$35.40 when used. Instructors, for their part, would have to spend considerable time creating online quizzes and redesigning lectures around clicker questions, and then additional time calculating the marks earned by students through quizzes and clickers. They would also likely need to draw more heavily upon the instructional technology services at their institutions when delivering material in this manner rather than in a more conventional form.

It may well be that a richer environment for the development of critical thinking skills is in semi-structured small-group discussions like those found in seminars or tutorials. An interesting finding of this exploratory study was an interactive effect between the pre-intervention test and tutorial marks during the winter semester. Overall, there was greater improvement indicated for those students who had higher tutorial marks. The size of improvement decreases (but remains positive) as the pre-intervention test mark increases. These data indicate that tutorials play a positive role to support student learning for many students during the winter semester. Moreover, student responses to the end of semester evaluation indicated that the role of the TA was critical to the success of the tutorial. Further research regarding the role of the tutorial and the TA on student learning outcomes, such as the development of critical thinking skills, could be helpful. As well, future research could focus on the optimal integration of tutorials and TAs to support student learning in hybrid and online learning activities.

Additional aspects of this exploratory study that would benefit from further research include why students tend to pick some foils rather than others on the pre- and post-intervention tests and why they find certain kinds of questions more difficult than others; how well students know what they know (metacognition); what the longer-term effects of engagement strategies are in helping students make a transition from high school through a first-year course and into upper-year university courses; whether the same engagement strategies are effective when used to teach different skills related to critical thinking (for example, identifying unacknowledged assumptions or logical fallacies in written arguments); and even what effect non-academic covariates such as part-time employment, living in residence or with parents, or first language have on the effectiveness of engagement strategies. Finally, it would be interesting to measure the effectiveness of engagement strategies by comparing students in the fall semester of one year against those in the fall semester of the next, rather than fall against winter, to see if the results reported here are due to a semester effect, and it would be interesting to measure the effectiveness of engagement strategies when deployed for a larger proportion of each lecture, perhaps to the point of creating an “inverted” class where students spend a majority of their lecture time solving problems in small groups.

For the investigation of these and other questions relating to history education, we would hope for greater cooperation between historians and their colleagues in other disciplines so that the disciplinary knowledge of the history practitioner can be combined with the pedagogical insights of natural and social scientists. With adequate support from scholarship of teaching and learning communities of practice at institutions of higher learning, professors from history (and a wide range of other disciplines) will be able to pursue pedagogical research to infuse subject-based, discipline-specific instruction with rigorously tested pedagogical practices. Thus may history courses become as evidence-based in their delivery as good history research is in its methodology, to the benefit of historians and their students at all levels of history education.

References

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey. *Physical Review Special Topics*, *2*, 1-14.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Anthis, K., & Adams, L. (2012). Scaffolding: Relationships among online quiz parameters and classroom exam scores. *Teaching of Psychology, 39*(4), 284-287.

Balter, O., Enstrom, E., & Klingenberg, B. (2013). The effective of short diagnostic formative web quizzes with minimal feedback. *Computers and Education, 60*(1), 234-242.

Barton, K. C. (2004). Research on students' historical thinking and learning. *Perspectives on History*, *42*(7). Retrieved from http://www.historians.org/Perspectives/issues/2004/0410/0410tea1.cfm

Bartsch, R. A., & Murphy, W. (2011). Examining the effects of an electronic classroom response system on student engagement and performance. *Journal of Educational Computing Research, 44*(1), 25-33.

Bean, J. (2011). *Engaging Ideas: the professor's guide to integrating writing, critical thinking, and active learning in the classroom.* San Francisco, CA: Jossey-Bass.

Behar-Horenstein, L., & Niu, L. (2011). Teaching critical thinking skills in higher education: A Review of the literature. *Journal of College Teaching and Learning, 8*(2)*,* 25-41.

Berry, T., Cook, L., Hill, N., & Stevenson, K. (2010). An exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. *College Teaching, 59*(1), 31-39.

Bloemhof, B., & Christensen Hughes, J. (2013). *Active learning strategies in introductory financial accounting classes*. Toronto: Higher Education Quality Council of Ontario.

Brewer, C. A. (2004). Near real-time assessment of student learning and understanding in biology courses. *BioScience, 54*(11), 1034-1039.

Britten, T. (2011). Using student response systems (clickers) in the history classroom. *Teaching History: A Journal of Methods*, *36*(1), 14-21.

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

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *Life Sciences Education, 6(*1), 9-20.

Chapman, L., & Ludlow, L. (2010). Can downsizing college class sizes augment student outcomes? An investigation of the effects of class size on student learning. *The Journal of General Education*, *59*(2), 105-123.

Chapnick, A. (2010). The CHA annual meeting: Personal reflections and ideas for the future. *Bulletin, 36*(2), 30-31.

Christensen, R. H. B. (2012). Ordinal-regression models for ordinal data, R package version 2012.09-11. Retrieved from http://www.cran.r-project.org/package=ordinal/

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172-1182.

Cole, S., & Kosc, G. (2010). Quit surfing and start clicking: One professor's effort to combat the problems of teaching the U.S. survey in a large lecture hall. *History Teacher*, *43*(3), 397-410.

Counsell, C. (2002). Historical knowledge and historical skills: A distracting dichotomy. In *Issues in History Teaching* (pp. 54-71). London: Routledge.

Cowan, M., & Landon, C. (2011). The missing links in history education. *The Canadian Journal for Social Research, 4(1),* 20-30.

Crenshaw, P., Hale, E., & Harper, S. L. (2011). Producing intellectual labor in the classroom: The utilization of a critical thinking model to help students take command of their thinking. *Journal of College Teaching and Learning, 8(7),* 13-26.

Denker, K. J. (2013). Student response systems and facilitating the large lecture basic communication course: Assessing engagement and learning. *Communication Teacher, 27(1),* 50-69.

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

Elliott, A. C., & Woodward, W. A. (2007). *Statistical Analysis: Quick reference guidebook*. Thousand Oaks, CA: Sage.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Exeter, D. J., Ameratunga, S., Ratima, M., Morton, S., Dickson, M., Hsu, D., & Jackson, R. (2010). Student engagement in very large classes: The teachers' perspective. *Studies in Higher Education, 35*(7), 761-775.

Fanscali, S. E. (2011). Variable construction for predictive and causal modeling of online education data. Paper presented at the 1st International Conference on Learning Analytics and Knowledge, Banff, AB.

Fortner-Wood, C., Armistead, L., Marchand, A., & Morris, F. B. (2012). The effects of student response systems on student learning and attitudes in undergraduate psychology courses. *Society for the Teaching of Psychology, 40*(1), 26-30.

Friesen, G., Muise, D., & Northrup, D. (2009). Variations on the theme of remembering: A national survey of how Canadians use the past. *Journal of the Canadian Historical Association*, *20*(1), 221-248.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers and Education, 57*(211), 2333-2351.

Gok, T. (2011). An evaluation of student response systems from the viewpoint of instructors and students. *TOJET: The Turkish Online Journal of Educational Technology*, *10*(4), 67-83.

Golding, C. (2011). Educating for critical thinking: Thought-encouraging questions in a community of inquiry. *Higher Education Research & Development, 30*(3), 357-370.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74.

Hatteberg, S. J., & Steffy, K. (2013). Increasing reading compliance of undergraduates: An evaluation of compliance methods. *Teaching Sociology, 41*(4), 346-352.

Historical Thinking Project (2011). Available at http://historicalthinking.ca/

Hobson, E. (2004). *Getting students to read: Fourteen tips.* Idea Paper #40. Retrieved from http://www.theideacenter.org/

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. S. (2013). Applied Logistic Regression. 3rd ed. Hoboken NJ: Wiley.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 17*(4), 1623-1634.

Huber, M. T., & Morreale, S. P. (2002). Situating the scholarship of teaching and learning: A cross-disciplinary conversation. In *Disciplinary styles in the scholarship of teaching and learning.* Washington, DC: American Association for Higher Education.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher, 16*(8), 4-9.

Hunt, C. (2012). Learning in large learning spaces: The academic engagement of a diverse group of students. *Research in Post-Compulsory Education, 17*(2), 195-205.

James, M. C. (2006). The effect of grading incentive on student discourse in Peer Instruction. *American Journal of Physics, 74*(8), 689-691.

Johnson, I. Y. (2010). Class size and student performance at a public research university: A cross-classified model. *Research in Higher Education, 51*(8), 701-723.

Johnson, M., & Robson, D. (2008). Clickers, student engagement and performance in an introductory economics course: A cautionary tale. *Computers in Higher Education Economics Review, 20*, 4-12.

Jolliffe, D. A., & Harl, A. (2008). Texts of our institutional lives: Studying the “reading transition” from high school to college: What are our students reading and why? *College English, 70*(6), 599-617.

Kay, R., & LeSage, A. (2009). A strategic assessment of audience response systems used in higher education. *Australasian Journal of Educational Technology*, *25*(2), 235-249.

Keirle, P. A., & Morgan, R. A. (2011). Teething problems in the academy: Negotiating the transition to large-class teaching in the discipline of history. *Journal of University Teaching & Learning Practice, 8*(2), 1-21.

Kerr, A. (2011). *Teaching and learning in large classes at Ontario universities: An exploratory study*. Toronto: Higher Education Quality Council of Ontario.

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education, 4,* 298-310.

Koenig, K. (2010). Building acceptance for pedagogical reform through wide-scale implementation of clickers. *Journal of College Science Teaching, 39*(3), 46-50.

Leger, A., Godlewska, A., Adjei, J., Schaefli, L., Whetstone, S., Finlay, J., Roy, R., & Massey, J. (2013). *Large first-year course re-design to promote student engagement and student learning*. Toronto: Higher Education Quality Council of Ontario.

Lévesque, S. (2008). *Thinking historically: Educating students for the twenty-first century*. Toronto: University of Toronto Press.

Lo, C. C. (2010). Student learning and student satisfaction in an interactive classroom. *The Journal of General Education, 59*(4), 238-263.

Maki, P. (2010). *Assessing for learning: Building a sustainable commitment across the institution*. Second ed. Sterling, VA: Stylus.

Mandel, P., & Süssmuth, B. (2011). Size matters. The relevance and Hicksian surplus of preferred college class size. *Economics of Education Review, 30*, 1073-1084.

Matus, J., Summa, K., & Kuschke, R. (2011). An analysis of technology-enhanced pedagogy and learning student response systems (clickers) – Tool or toy. *International Journal of Business and Social Science, 2*(12), 6-13.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34*, 51-57.

McDougall, D., & Cordeiro, P. (1993). Effects of random-questioning expectations on community college students’ preparedness for lecture and discussion. *Community College Journal of Research and Practice, 17*, 39-49.

McKeown, M. G., & Beck, I. L. (1994). Making sense of accounts of history: Why they don’t and how they might. In *Teaching and Learning in History* (pp. 1-26). Hillsdale, NJ: Erlbaum.

Metz, A. M. (2008). The effect of access time on online quiz performance in large biology lecture courses. *Biochemistry and Molecular Biology Education, 36*(3), 196-202.

Meyer, D., Zeileis, A., & Hornik, K. (2013). *VCD: Visualizing Categorical Data*. R package version 1.3-1.

Minitab 16 Statistical Software (2010). State College, PA: Minitab, Inc. Retrieved from www.minitab.com

Moulding, N. T. (2010). Intelligent design: Student perceptions of teaching and learning in large social work classes. *Higher Education Research and Development, 29*(2), 151-165.

Mulnix, J. W. (2010). Thinking critically about critical thinking. *Educational Philosophy and Theory*, *44*(5), 464-476.

Nakayama, M., Yamamoto, H., & Santiago, R. (2009). Relationships between earners' characteristics and learned behaviour of Japanese students in blended learning environment: A three-year study. Proceedings of International Conference on e-Learning, 2009, 377-385.

Neter, J., et al. (1996). *Applied Linear Statistical Models*. Chicago, Il: Irwin.

Osborne, K. (2003). Teaching history in Schools: A Canadian debate. *Journal of Curriculum Studies, 35*(5), 585-626.

Pace, D. (2004). The amateur in the operating room: History and the scholarship of teaching and learning. *The American Historical Review, 109*(4), 1171-1192.

Peace, R. (2010). Cultivating critical thinking: Five methods for teaching the history of U.S. foreign policy. *The History Teacher, 43*(2), 265-273.

Peck, C., & Seixas, P. (2008). Benchmarks of historical thinking: First steps. *Canadian Journal of Education, 31*(4), 1015-1038.

Perfetti, C. A., Britt, M. A., Rouet, J.-F., Georgi, M. C., & Mason, R. A. (1994). How students use texts to learn and reason about historical uncertainty. In *Cognitive and Instructional Processes in History and the Social Sciences* (pp. 257-283). Hillsdale, NJ: Erlbaum.

Poole, D. (2012). The impact of anonymous and assigned use of student response systems on student achievement. *Journal of Interactive Learning Research, 23*(2), 101-112.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Redish, E. (2003). *Teaching physics with the physics suite.* Hoboken, NJ: Wiley and Sons.

Reichard, D. A. (2006). How do students understand the history of the American West?: An argument for the scholarship of teaching and learning. *The Western Historical Quarterly, 37*(2), 207-214.

Saltmarsh, D., & Saltmarsh, S. (2008). Has anyone read the reading? Using assessment to promote academic literacies and learning cultures. *Teaching in Higher Education, 13*(6), 621-632.

Seixas, P. (2006). “Ropes and pulleys”: Reflections on a conference on the preparation of history teachers. *Bulletin, 32*(3)*,* 27-29.

Seixas, P., Fromowitz, D., & Hill, P. (2005). History, memory and learning to teach. In *Understanding history: Recent research in history education* (pp. 107-123)*.* London: Routledge.

Seixas, P., & Morton, T. (2012). *The big six historical thinking concepts*. Toronto: Nelson.

Snowball, J. D., & Boughey, C. (2012). Understanding student performance in a large class. *Innovations in Education and Teaching International, 49*(2), 195-205.

Stowell, J. R., Oldham, T., & Bennett, D. (2010). Using student response systems (“clickers”) to combat conformity and shyness. *Teaching of Psychology*, *37*, 135-140.

Stull, J. C., Majerich, D. M., Bernacki, M. L., Varnum, S. J., & Ducette, J. P. (2011). The effects of formative assessment pre-lecture online chapter quizzes and student initiated inquiries to the instructor on academic achievement. *Educational Research and Evaluation*, *17*(4), 253-262.

Tolley, L. M., Johnson, L., & Koszalka, T. A. (2012). An intervention study of instructional methods and student engagement in large classes in Thailand. *International Journal of Educational Research*, *53*, 381-393.

Trees, A. R., & Jackson, M. H. (2007). The learning environment in clicker classrooms: Student processes of learning and involvement in large university-level courses using student response systems. *Learning, Media and Technology, 32*(1), 21-40.

University of Toronto Mississauga (2012). 2012-2013 Academic Calendar. Retrieved from https://student.utm.utoronto.ca/calendar/images/FinalCalendar20122013v3.pdf

University of Toronto Mississauga (n.d.). Guidelines for University of Toronto Mississauga Undergraduate Degree Level Expectations. Retrieved from http://www.vpacademic.utoronto.ca/Assets/VP+Academic+Digital+Assets/DLE/UTM_DLE.pdf

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Fourth ed. New York, NY: Springer.

Warren, W. J., Memory, D. M., & Bolinger, K. (2004). Improving critical thinking skills in the United States survey course: An activity for teaching the Vietnam War. *The History Teacher, 37*(2), 193-209.

Wineburg, S. (2001). *Historical thinking and other unnatural acts*. Philadelphia, PA: Temple University Press.

Zurmehly, J., & Leadingham, C. (2008). Exploring student response systems in nursing education. *Computers, Information, Nursing*, *26*(5), 265-70.

Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario