Teaching and Learning with Online Courses

Richard Scheines, Gaea Leinhardt,
Joel Smith, and Kwangsu Cho

July 8, 2002

Technical Report No. CMU-PHIL-135

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Teaching and Learning with Online Courses[1]

Richard Scheines,[2] Gaea Leinhardt,[3]
Joel Smith,[4] and Kwangsu Cho,[5]

Abstract

With the increasing use of online technology to deliver, support, or extend college level instruction comes a new range of demands on faculty as instructional designers and students as learners. Faculty are challenged to provide effective materials in the online rather than face-to-face setting, and students are challenged to adapt their learning strategies to the online environment. In a series of 5 experiments in 2000 and 2001, several hundred students at two different universities with three different professors took a semester long course on causal and statistical reasoning in either standard lecture/recitation or online/recitation format. In this paper we compare the pre-post test gains of these students, we identify features of the online experience that were helpful and elements that were not, and we identify student learning strategies that were effective and those that were not. Students who replaced going to lecture with doing online modules did as well or better than those who went to lecture. Simple strategies like incorporating frequent interactive comprehension checks into the online material (something that is difficult to do in lecture) proved effective, but online students attended face-to-face recitations less often the lecture students and suffered because of it. Supporting the idea that small, interactive recitations are more effective than large, passive lectures, recitation attendance was three times as important as lecture attendance for predicting pre-test to post-test gains. For the online student, embracing the online environment as opposed to trying to convert it into a traditional print-based one was an important strategy, but simple diligence in attempting "voluntary" exercises was by far the most important factor in student success. In this paper we detail the studies we performed, summarize the results, and describe ongoing and future studies.

---

[1] This research was funded by the A.W. Mellon Foundation's Cost-Effective Uses of Technology in Teaching program (http://www.ceutt.org/).
[2] Dept. of Philosophy and Human-Computer Interaction Institute at Carnegie Mellon University.
[3] Learning Research and Development Center, and School of Education, University of Pittsburgh
[4] Office of Technology for Education, and Chief Information Officer, Carnegie Mellon University.
[5] Learning Research and Development Center, University of Pittsburgh

1. Introduction

Because courses given entirely or in part online have such obvious advantages with respect to student access and potential cost savings, their development and use has exploded over the last several years.[6] Although we now know quite a lot about online learning, e.g., how faculty and students respond subjectively to it, and what strategies have proven desirable from both points of view [Hiltz et al., 2000; Kearsley, 2000; Sener, 2001; Wegner, et al., 1999; Clark, 1993; Reeves and Reeves, 1997], we still know far too little about how online course delivery compares to traditional course delivery with respect to objective measures of student learning. Although many studies have reported no significant difference in learning outcomes between delivery modes [Barry and Runyan, 1995; Carey, 2001; Cheng, Lehman, and Armstrong, 1991; Maki, Maki, Patterson, & Whittaker, 2000; Russell 1999], and some have shown that online students fared worse [e.g., Brown and Liedholm, 2001], few have compared entire courses and still fewer have managed to overcome the many methodological obstacles to rigorous contrasts [Phipps, et al., 1999; Carey, 2001; IHEP, 1999].

There are extensive bodies of solid research showing that educational technology that involves intelligent models of the student and/or cognitive theory does improve objective learning outcomes, e.g., Carnegie Mellon's Cognitive Tutors built on John Anderson's ACT-R theory of cognition [Anderson, et al., 1995, Koedinger & Anderson, 1998], or VanLehn's physics tutor [VanLehn et. al., 2000], or the Jasper Project's mathematics tutor [Cognition and Technology Group at Vanderbilt, 1997], but studies of this type focus on the impact of the cognitive theory and do not systematically compare delivery modes. Further, cost is an issue. While many might be able to put a course online in a form that approximates their own lecture pedagogically, few have the resources or training to build a cognitive tutor.

Although a small percentage of college faculty are truly great lecturers who inspire hundreds of students each year and should continue to do so, many who give large lectures for high enrollment introductory courses would perhaps provide a greater educational service in some other way, either by spending more time meeting small groups of students, or by writing fresh interactive material. Although some faculty may believe this, few are willing to stop being the 'sage on the stage' and move to being the 'guide on the side'. Here we believe the computer really does have vast but largely untapped potential. Instead of just capturing decent lectures on video and thus delivering

[6] See, for example, the many efforts described or cited in [Bourne and Moore 2000]

them at a lower bandwidth than the live setting (the default for early efforts at "distance education"), the computer provides the capability to deliver the "content" covered in a lecture interactively: it can provide frequent pauses for comprehension checks, interactive simulations/animations along the way, almost continual feedback on the student's progress, experiments performed by the students as a group. Well designed interaction between the student and the computer has the potential, we believe, to improve the experience that most students are getting in large introductory lecture courses.

We are not arguing that face-to-face time between students and teachers should be replaced by the student-computer interaction – we believe no such thing. All of the students in our experiments attended small face-to-face recitation sections at least weekly. The question at issue is the effect of replacing large introductory lectures with interactive, online courseware. In this paper, therefore, our priority is to address the simplest question about online courseware: can it replace face-to-face lectures without doing any harm to what the students objectively learn from the course.

Whatever the answer to this question in our study or others, we do not yet generally know how to design online learning environments in order to maximize objective student learning, nor do students know much about what strategies they must adopt to learn effectively in "distance education" settings or in online courses. Thus the second goal of this paper is to begin the process of identifying the features of online environments that are pedagogically important, and the student strategies that are adaptive in the online setting and those that are not.

Because online technology makes it possible to automatically collect a lot of data, it presents as large an opportunity to learn about teaching and learning as it does to increase access and to reduce costs. We should be able to answer questions like: in an online course, where are students really spending their time? Where are they really having trouble? When and for what sorts of activities do they need face-to-face time, and when will online learning suffice? We do not answer all of these questions in this paper, but we begin the process.

Seeing the potential for using online courseware to learn about learning, we recently began to turn our online course project,[7] originally designed merely to deliver educational material, into an environment for studying the human-computer interaction of online learning. It is far from a sophisticated laboratory yet, however, and we emphasize that

[7] See www.phil.cmu.edu/projects/csr. Our material was first offered to large numbers of real students in the spring of 2000.

the results we report here are the tip of the iceberg. Nevertheless, we locate our efforts in the second part of the paper midway along a continuum, on one end of which are extremely fine-grained studies performed on tiny segments of a course and on the other end of which are large 'black box' evaluations of pre-test post-test gains.

The paper is organized as follows. In the next section, we briefly describe the online course material. In section three we describe the experiments we carried out to study it. In section four, we discuss the evidence for the claim that replacing lecture with online delivery did no harm and probably some good, and we discuss which features of the online environment helped and which seemed to hinder student outcomes. In section six we discuss the student strategies that were adaptive and those that were not, and in the final section we discuss some of the many questions left unanswered and the future research that will address them.

2. Online Courseware on Causal Reasoning

Although Galileo showed us how to use controlled experiments to do causal discovery more than 400 years ago, it wasn't until R.A. Fisher's [1935] famous work on experimental design that further headway was made on the statistics of causal discovery. Done well before World War II, Fisher's work, like Galileo's, was confined to experimental settings in which treatment could be assigned. The entire topic of how causal claims can or cannot be discovered from data collected in non-experimental studies was largely written off as hopeless until about the mid 1950s with the work of Herbert Simon [1954] and the work of Hubert Blalock seven years later [Blalock, 1961]. It wasn't until the mid 1980s, however, that artificial intelligence researchers, philosophers, statisticians and epidemiologists began to really make headway on providing a real theory of causal discovery from non-experimental data.[8] Convinced that at least the qualitative story behind causal discovery should be taught to introductory level students concurrent with or as a precursor to a basic course on statistical methods, and also convinced that such material could only be taught widely with the aid of interactive simulations and open ended virtual laboratories, a team at Carnegie Mellon and the University of California, San Diego[9] teamed up to create enough online material for an entire semester's course in the basics of causal discovery. At the time of this

---

[8] See, for example, Spirtes, Glymour and Scheines (2000), Pearl (2000), Glymour and Cooper (1999).
[9] In addition to Scheines andd Smith this includes Clark Glymour, at Carnegie Mellon and the Institute for Human-Machine Cognition (IHMC) in Pensacola, FL and David Danks, now at IHMC, Sandra Mitchell, now at the University of Pittsburgh, Willie Wheeler and Joe Ramsey, both at Carnegie Mellon.

writing (the spring of 2002), over 1,600 students at eleven different universities have taken our online course.

Our online material (www.phil.cmu.edu/projects/csr) involves three components: 1) 17 lessons, or "concept modules," 2) a virtual laboratory for simulating social science experiments, the "Causality Lab" and 3) a bank of over 100 short case social science and medical studies taken from news service reports of "studies," e.g., "link found between day care and poor social adjustment." Each of the concept modules contains approximately the same amount of material as a text-book chapter or a 90 minute lecture, but also includes interactive simulations, questions with feedback, in some cases more extended exercises to be carried out in the Causality Lab, and frequent comprehension checks, e.g., two or three multiple choice questions after approximately every page or so of text. At the end of each module is a graded online quiz.

The material is intended to replace lectures and constitute some part of the homework in a class, but is not intended be used as an entirely stand alone course. It still requires teaching assistants to lead recitation sections and discussions. Although many of the basic concepts in causal reasoning (and many other subjects) can be delivered by computer, integrating these concepts and especially applying them to real cases requires the kind of subtlety and flexible intelligence well beyond even the best computer tutors.

3. The Experiments

The Treatments

In order to test the relative efficacy of delivering our material online, we created two versions of a full semester course, one to be delivered principally online and one principally by lecture. The two versions were as identical in all respects save delivery format as we could make them. In the lecture version of the course, the class consisted of two lectures per week and one recitation section. For reading, the online modules were printed out (minus, of course, the interactive simulations and exercises) and distributed to the students. The lectures essentially followed the modules.

In the online version of the course, students got the material from the online concept modules instead of lecture (they were required to complete one module each time a lecture was given on the same topic), and in fact were not allowed to go to lecture. Since the online version of the modules involved interactive simulations and exercises not included in the readings passed out to lecture students, extra assignments and more traditional exercises were given out to lecture students to equalize the opportunity to learn.

Both versions of the course included one interactive recitation section per week in which problem sets and case studies were gone over and discussed. Students in both conditions went to identical recitations. In three of the five experiments online and lecture students were mixed in the same recitations, but the results were indistinguishable to experiments in which online and lecture students were separated in recitation. All students took identical paper and pencil pre-tests, midterms, and final exams, at the same time in the same room.

We compared both delivery formats in the 2000 winter and spring quarters at UCSD (University of California, San Diego) on over 300 students, in the 2001 winter and spring quarters at UCSD on almost 280 students, and in one full semester during the spring of 2001 at the University of Pittsburgh on over 80 students,.

Treatment Assignment

Allowing students to choose which delivery format they receive is desirable from the student's point of view, but clearly invites a selection bias from our point of view, which is a disaster for causal discovery. In fact most of the studies comparing online to lecture delivery that we are aware of did not randomize treatment assignment, even partially.[10] There are two simple ways to deal with treatment selection bias: randomly assign treatment or identify the potential source of the bias and then measure and statistically control for it. In 2000 we used a semi-randomized design, which employed both strategies.

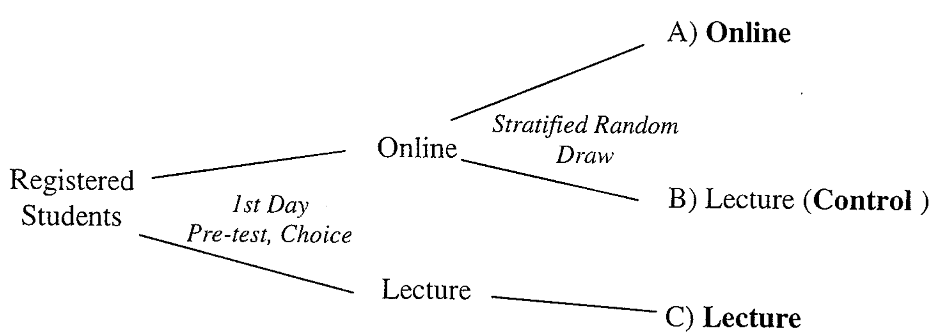

Figure 1. Semi-Randomized Design for Experiments in 2000

---

[10] See, for example, Maki et al., 2000, and Carey, 2001.

In our semi-randomized experiments, conducted in the winter and spring of 2000 in a large introductory level Philosophy course on scientific reasoning given at UCSD, we did not advertise the course as having an online delivery option. On the first day of class we administered a pre-test and informed students that they had the option to enter a lottery to take the course in online format. All students who wanted traditional lecture format (condition C) got it. We then took all the students who opted for the online delivery condition, ranked them by pre-test score, and then did a stratified random draw to give 2/3rd of the students who wanted online delivery their choice: A) Online – wanted and got the online condition, and B) Control – wanted online but got lecture. Although this design leaves out one condition: students who wanted lecture but were assigned online delivery – we felt that such an assignment was unethical given how the course was advertised and given we did not yet know how the two groups would fare in the learning outcomes. We assured both groups that if there were any differences in the mean final course scores we would adjust the lower up by the difference in means.

In 2001, both at UCSD and at the University of Pittsburgh, students were again informed of the two options on the first day of class as well as how the previous year's groups had done, but were then given whichever treatment they chose. First, the results from 2000 showed little difference between the groups, second, we observed very little treatment selection bias, and finally, we improved our pre-test and were confident we had a reasonable measure of pre-course ability.

4. Results

We present the results from these five experiments roughly chronologically, for several reasons. First, as with any experience that repeats, we learned things in early versions of the study, which we used to change later versions, and in several instances the lessons learned are worth recounting. Second, the scope and quality of the data collection effort improved steadily over time. We had a richer set of measures to analyze in 2001, especially at Pitt, and our record keeping improved over time. Finally, although presenting five studies sequentially may seem a little redundant, the fact that the results were approximately replicated over five slightly different versions of a course involving three different professors, six different teaching assistants, two different treatment assignment regimes, and two locations separated by over 2,000 miles convinced us far more than p-values that we were not seeing a statistical mirage. In what follows we slightly vary the format of our presentation of the results, mostly in response to the data available for the study reported on.

UCSD: Winter and Spring 2000

In the semi-randomized design used at UCSD in the winter and spring quarters of 2000 (Figure 1), two comparisons are in order: 1) the Online vs. Control comparison, and 2) the Control vs. Lecture comparison. Comparing Online vs. Control gives us the treatment effect among students who are disposed to do online courses, and comparing Control vs. Lecture gives us an estimate of the treatment selection bias, as these groups both received the same treatment (lecture delivery) but differed as to what delivery they chose. Figure 2 displays the means[11] for each group on the pre-test, midterm and final exam and thus summarizes the results for winter 2000. Pre-test means were indistinguishable across groups, and although Online students outperformed Control and Lecture students, the differences were not significant at $\alpha = .05$.[12]

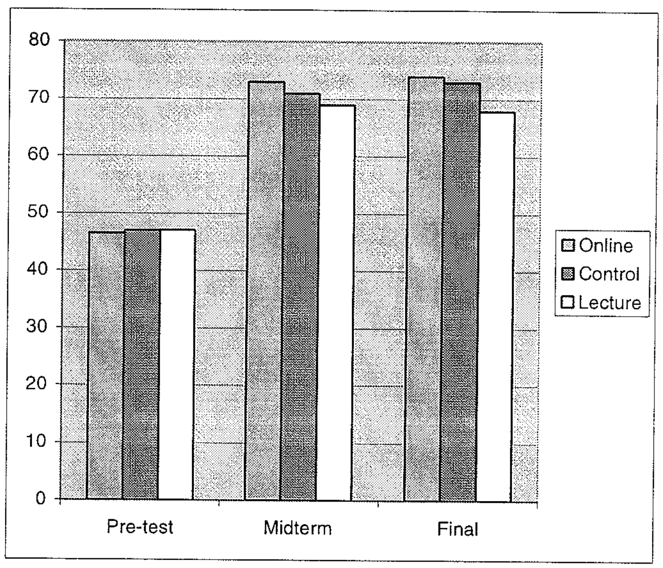

Figure 2. Winter Quarter 2000 (N = 180)

Interestingly, although the Control and Lecture conditions showed literally no pre-test difference, Control students did consistently slightly outperform the Lecture condition – especially on the final exam (p = .2). We took this as suggestive evidence that there *was* a small selection bias of approximately 2-3% that our pre-test did not pick up. This is consistent with other studies comparing online vs. lecture treatment in which treatment

[11] All sample distributions were approximately normal.
[12] Considering only items common to the pre-test and final exam, the online students did outperform the control group at p = .015.

was selected by the students and not assigned; see [Maki & Maki, 1997 and Maki, R. H., Maki, W.S., Patterson, M., & Whittaker, 2000], for example.

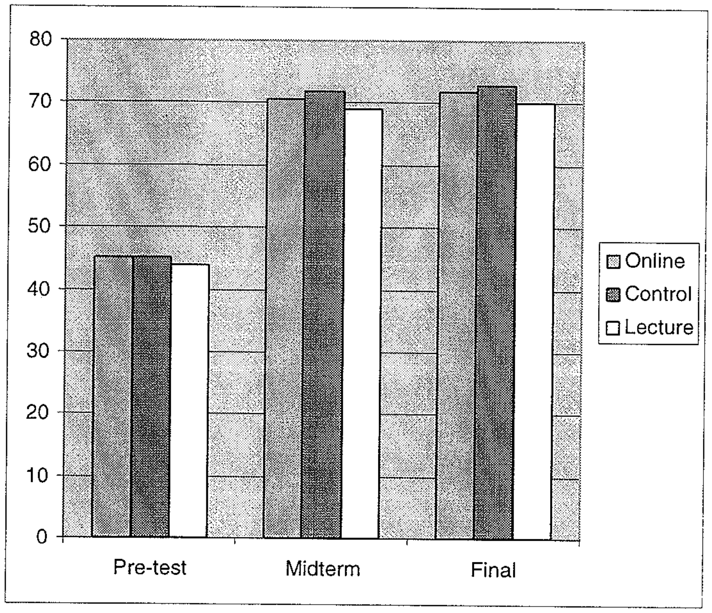

Figure 3: Spring Quarter 2000 (N = 130)

In the spring quarter, we repeated the experiment (Figure 3). Again, there was a small (~3% selection bias), but unlike the winter quarter, in the spring quarter the Control condition consistently (albeit insignificant statistically) outperformed the Online condition. Upon examining the attendance records, a promising explanation emerged. In the winter quarter, all students in lecture delivery were required to attend recitation to take a weekly quiz (the online students did their quiz on the computer). Over the winter semester, the average lecture students attended 85% of the recitations, but the average online student attended only 20% for the online students in the winter quarter. In the spring, however, average recitation attendance among lecture students stayed at almost exactly 85%, but online students attended an average of fewer than 8% of recitations!

As a result of these experiments, we made three major modifications for 2001. First, we improved the pre-test by incorporating GRE items from the analytic ability section relevant to our subject. Second, we allowed all students to choose their method of delivery, and third, we required recitation attendance of both online and lecture students. We again ran the experiment at UCSD in both winter and spring quarters of 2001, and also added a class in the spring semester of 2001 at the University of Pittsburgh.

UCSD: Winter and Spring 2001

The results in the winter quarter for 2001 at UCSD were quite similar to those in 2000, but in the spring quarter Online students showed a larger selection bias (3.3%) and larger performance advantage as well.

Table 1: Winter and Spring 2001: Online vs. Lecture

Percentage Difference in Means: Online - Lecture				
	Pre-test	Midterm 1	Midterm 2	Final Exam
Winter 2001 (N = 157)	1.9 (p = .51)	3.6 (p = .09)	-0.35 (p = .866)	0.4 (p = .801)
Spring 2001 (N = 121)	3.3 (p = .26)	**9.8 (p = .001)	**11.2 (p = .001)	*6.08 (p = .014)

Unfortunately, the connection between individuals and pre-test scores was corrupted in the winter 2001 data for UCSD, as was the attendance records, so only summary statistics are available. In the spring quarter, however, the Online students averaged 4.42% higher on the final exam than the Lecture students, after controlling for pre-test. Regressing Final exam on pre-test and a dummy variable to encode treatment condition (Online: 1= online, 0 = lecture), with standard errors in parentheses and p-values below gives the following results.

$$\textbf{Final (\%)} = 53.4 + \underset{\substack{(2.42)\\ 0.073}}{4.42}\ \textit{Online} + \underset{\substack{(0.087)\\ {}^{**}0.001}}{0.315}\ \textit{pre-test}$$

University of Pittsburgh 2001

For several reasons, our best data come from the spring semester at the University of Pittsburgh. First, we were present to supervise data collection efforts. Second, the principle author of the online material (Scheines) was also the lecturer for the course, presumably thereby giving the lecture section the best chance to shine. Third, and perhaps most importantly, we logged student behavior on a few important variables - how often they printed out modules to study, how often they attempted the voluntary comprehension checks inserted every page or two in the online modules, and how well they did on each post-module quiz.

As in the UCSD experiments performed in 2001, students were told the Online and Lecture options on the first day of class and then allowed to freely choose their treatment condition. At the University of Pittsburgh, 35 students chose Online and 50 chose Lecture. First, the difference in pre-test means between the Online and Lecture conditions was just over 1%, statistically insignificant. Second, gender was independent of virtually every quantity we measured, including pre-test, treatment preference, exam performance. Third, dropout, which averaged around 10-15% across our experiments, was nearly independent of treatment condition and thus had little or no effect on any of estimates of treatment effect.

After controlling for pre-test and recitation attendance, online students averaged 5.3% higher on the final exam. Regressing Final exam score (in percent) on pre-test, the percentage of recitations attended, and a dummy variable for treatment condition gives the following results:

Final = 42.1 +	0.220 *pre-test* +	0.233 *rec* +	5.26 *Online*
	(.092)	(.069)	(2.73)
	*.021	**.001	.059

Consistent with the UCSD experiment in spring 2001, Online is significant at .1 but not at .05. It also shows that, as we had suspected from the UCSD experiments, recitation attendance strongly predicts performance. The expectation of Final exam score rises almost a quarter point (.233) for every extra percent of recitation attendance, pre-test and treatment condition equal. Since there were only 13 recitations, each accounting for almost 8% of total recitation attendance, each extra recitation attended increases the expectation for the Final exam by almost 2%.

To get a further handle on the importance of recitation attendance, we compared the relative importance of recitation vs. lecture attendance among only Lecture students in the Pitt experiment. These students were supposed to attend lecture twice a week and recitation once, but attendance at recitation was over four times more predictive than attendance at lecture in a regression with Final as the dependent variable:

0.317 *rec%* +	0.078 *lec%*
(.117)	(.101)
**.010	.448

We take this as evidence, found by many others, that students learn more from small sessions in which they are active and engaged as opposed to large sessions in which they are lectured at.

Although the percent of recitations attended among online students rose from an embarrassing average of 8% in the spring of 2000 at UCSD to an average of 71% in the spring of 2001 at Pitt, it still trailed recitation attendance among Lecture students (81%) by 10% ($p = .05$). We hypothesize that this discrepancy is a result of the greater aversion among online students to attend scheduled educational gatherings. It might, however, be the result of reduced weekly contact, or the greater independence required of online students. We do not yet know. If being in the online condition caused students to attend fewer recitations, then that probably has an adverse indirect effect on performance.

Path Analysis

Since there might well be two mechanisms through which treatment effects learning outcome, one direct and the other indirect, we used path analysis [Wright, 1921; Bollen, 1989] to estimate the strength of each mechanism. Table 2 shows the sample correlations among the four variables, with an "*" attached to correlations significant at $\alpha = .05$:

Pre: pre-test%,

Online: (1 = yes, 0 = no),

Rec: % recitations attended

Final: final exam %

Table 2: Correlations, Means, SDs (N = 83)

	Pre	*Online*	*Rec*	Means	SD
Pre				33.83	13.15
Online	0.023			-	-
Rec	-0.004	*-0.255		78.45	19.51
Final	*0.287	0.182	*0.297	70.23	11.14

The path model we used to estimate the relations among these variables is shown in Figure 4, along with the path coefficients, estimated not from the correlations but from the raw data to connect easily to regression results above. The path coefficients on the edges going into Final are almost identical to the regression estimates shown above.

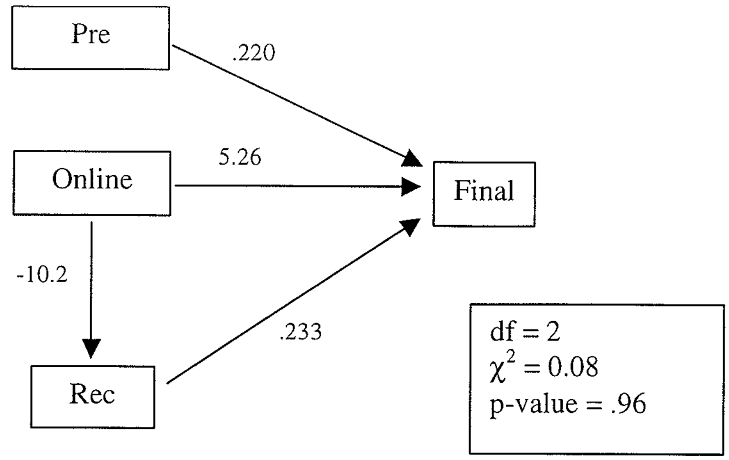

Figure 4: Two Paths from Online to Final

The path model as a whole contains two important pieces of information. First, the fact that there are no edges connecting Pre to either Online or Rec is important, as it signifies that ability as measured by the pre-test has no influence on treatment selection, and no influence on whether a student attends recitation. Second, there are two paths from treatment (Online) to Final. The direct path indicates that, controlling for pre-test and recitation attendance, online students tend to average 5.26% higher on the Final than Lecture students. The indirect path: Online → Rec → Final, however, indicates first that Online students attend 10.2% fewer recitations on average, but that each extra percent of recitation attendance increases a student's average final exam score by .23%, meaning the indirect effect of Online on Final through Rec is to reduce Final exam scores by an average of 2.38 percent. Thus, if the path model above is correctly specified, the approximate overall effect of Online on Final exam score is 5.3 - 2.4 = 2.9, or about a 1/3 of a grade.

The standard approach to estimating the strength of the relationships between variables like these, is to first specify a statistical model, and then calculate p-values relevant to the existence of particular relationships. This sort of statistical inference, however, is *conditional* on the model specification, a fact that is appreciated in theory but widely ignored in practice. Put another way, coefficient estimates and standard errors will vary considerably with the model specification, so unless one has high confidence in the model specification, the statistics are illusory. With a p-value of .96, the path model Figure 4 fits the data extremely well,[13] and in an exhaustive search of all possible

[13] In path analysis, higher p-values mean better overall fit. See [Bollen, 1989].

alternative path models consistent with the time order among the variables in this model,[14] no alternative fit as well.

Path models are limited in that they do not allow for the possiblity of unmeasured confounders. In this case, the significant negative correlation between Online and recitation attendance might be due to an unmeasured confounder and not the result of a direct cause, but we modeled it as a direct cause because if anything this specification *weakened* the case for Online being the better treatment condition.

5. The Good Online Student

Up to this point we have compared the learning outcomes of online vs. lecture students. In this section we begin the process of analyzing the sorts of student behaviors that support or restrict objective learning in the online setting.

As with face-to-face instruction in colleges and universities there is a presumed set of student behaviors and an enacted set of behaviors. The presumption is that students will want to maximize their learning outcomes in a given course and so attend classes, do the suggested readings at a fairly steady pace, do the homework as assigned, study for tests and quizzes in a way that integrates new pieces of information together in a coherent and flexible fashion. In other words the student is expected to become engaged with both the process and substance of a course. Becoming engaged is somewhat more ill defined in the online setting. One might hypothesize that the skills of studentship are the same online as they are in a face-to-face setting. It might be the case, however, that in the online setting students adopt a more passive 'just follow the directions' stance, or it might be that online course work requires a more engaged and active student - one that moves around flexibly in the virtual world as opposed to linearly in a textbook world. The good online student might want to become engaged, but it might not always clear what they are to be engaged with. To date there has been very little in the way of behavioral research on the topic of 'what makes a good online student?' especially in terms of their behavior over an entire term.

It is highly likely that what makes a good online student depends on what the student is trying to learn and in which community of practice the course is trying to involve the student. For example, an online course in physics is likely to be built around a series of problems that make use of an increasingly sophisticated application and combination of

---

[14] Pre-test was prior to treatment selection, which was prior to recitation attendance, which was prior to final exam.

core principles [e.g., Ur & VanLehn, 1995]. The work of the course designers is to trace the student's 'moves' and to anticipate the confusions that might occur about principles and their applications. The activity of students is that of creative and flexible problem solving and a continuous updating of the representation and meaning of the core principles. On the other hand, in a course on causal and statistical reasoning the course may be built around a core set of terms and concepts that are applied to increasingly complex real world situations. The work of the students is to learn the terminology and definitions and be able to apply them in increasing complex environments. The activity of the student is to keep track of the newly evolving language systems and their subtle distinctions.

We address the issue of the 'good' student by examining the practices of the 'successful' student, where success is test performance. Thus, final exam performance is a proxy for success and success is a proxy for good. We address the problem in this way only as a first approximation. Our central question then is, what do successful students do when taking a course online? More specifically, what do successful students do when taking a particular course on causal reasoning in an online environment.

The investigation of student learning performance has a rich history, but much of it has taken place in the last 35 years. In that time, for example, we re-discovered the Socratic notion that students should not merely be present but that they should be actively engaged in their learning [Leinhardt, 1980]. Likewise we rediscovered the idea of apprenticeship by noticing that it applies to cognitive and not just physical skills [Brown & Campione, 1994]. Finally, we came to understand the ways in which students could be seen as entering a community of practice by gradual participation in the activities of that practice; most specifically by engaging in the discourses of those practices [Greeno, 1998]. In the context of considering learning as the increased participation and competence in a particular set of activities it is especially important to realize that the online student is participating in two communities at all times – the community of online learning and the community of the specific course and its content.

What do we know about the activity of learning material online? We know little but are developing an emerging picture. For example, in order to apply the normal and successful practices of traditional learning the good student often extracts the material from the screen and places it on paper. The paper version can be marked up, shuffled around, carried, and studied in a variety of environments. While analogous activities can be carried out on the computer they are more effortful and often less satisfactory – scribbling in the margins and drawing small diagrams does not require opening new windows or highlighting and dragging. McIsaac and Gunawardena [1993] suggest that print is a critical support for distance learners in current online learning systems.We also know that a good indication of engagement is that students actually do the embedded problems of the course material [Pressley & Ghatala, 1990] and do not simply flounder by clicking answers until they find the right ones.

What is the task of the learner online? First, the learner must determine a way of becoming engaged. Engagement no longer means listening and taking notes and asking questions, it means something else. But what? Second, the student must determine both what the course author deems important and what the student him or herself recognizes is necessary—what is to be studied. Finally, the student needs to assess how to study, practice, and rehearse ideas and skills. What are the challenges? Figuring out how to become a good student, what to do and how to do it has become less visible. Students are less likely to be getting regular reality checks on their own behaviors during a course than they do in a face-to-face environment. Even though the 'programs' that the student is involved with may have more frequent checks built in the form of tests the normal casual give and take in the classroom setting has been removed and accommodating to the delivery medium is difficult. What are the affordances? There is the powerful feature of animation and dynamics in course presentation. Dynamic presentations can enact principles more clearly than static ones in which the 'motion' must be imagined [Mayer & Anderson, 1991; Mayton , 1990]. Clear mappings between representations can be more forcefully and economically presented in the online environment this can allow for easy movement between qualitative understandings and more formal ones. Finally, the potential for continuous self-assessment through deliberate feedback and sample test material is available. But, as with all affordances students have to have the knowledge that there is something in the environment that they can make use of.

A good student in this course would need to find a way to access and notate the material, study the examples carefully, take all of the embedded questions and note them, study the materials sufficiently carefully to pass the quizzes at the end of each module. One way to access and notate the materials is to print them out. However, when the materials are printed out the embedded questions disappear. Thus the student must read/study the materials off line and then take the embedded questions and run the simulations or study the materials online and for go the printing out.

We began to study some of these issues in the spring of 2001 by recording more about student behavior than just attendance, pre-test, post-test, etc.

Population

The study was conducted in the first half of 2001 and involved two groups of online students, one taking the course in the winter quarter at UCSD and the other at the University of Pittsburgh in the spring. Out of the 75 students who decided to take the course online and who stayed in the course for the entire semester 68 records were obtained, 52 of which were complete and used here.

Measures

Pretest. Students took a short multiple-choice pretest that included items similar to those that they would take as final exam questions and embedded questions. The items included vocabulary based ideas, graphical representations of material, and concrete as well as abstract representations of core ideas. The score is the percent correct.

Printout usage. As a feature of the courseware each module has available a "print" feature/link. If a student clicks on this button then this links the student to a 'printable' page made up of the entire module and its headings. Therefore, whenever a student made use of this feature a record of the behavior was available to us. The 'printout' measure consists of a ratio of the total number of clicks to this button divided by the total number of modules accessed by the student. Printing out the module is a mixed signal, as it indicates a level of engagement, but perhaps a resistance to using the modules online, as they were intended. Obviously printing in and of itself does nothing to the acquiring of knowledge.

Question missing. As we described above each module contains a set of embedded questions. The questions usually deal with the material introduced in the previous two or three pages. Sometimes the questions follow an active simulation. Ideally a student would run all of the simulations in the module and would answer all of the embedded questions. However, a student might choose to skip the questions, not do the simulation, answer the questions by scanning all of the possible answers, etc. More important even than these problems is the fact that if a student is working from a printout version then there is a press in the direction of NOT doing the questions. We chose the 'measure' of answering embedded questions to be NOT doing the question. Not doing the questions is the ratio of all unopened questions divided by the total number of embedded questions. (If a student 'skipped' a module then the student was counted as having missed the question.)

Quiz score. Each module ends with a quiz. The students take the required quiz online and they received a percentage correct score. The quiz score contributed to their course grade. We summed the percentage correct divided by the number of quizzes taken over the entire course to construct a measure of total quiz score.

Final exam score. This is the student's percentage correct on the final exam.

Path Analytic Models

As above, we used path analytic models to study the causal relationships among these variables. The correlation matrix among these variables is given in Table 3.

Table 3: Correlation, Mean, SD (N = 52)

	Pre	*Print*	*Miss*	*Quiz*	*Final*	M	SD
Pre	1.000					22.2/30	16.6
Print	0.301*	1.000				0.5	0.60
Miss	0.258	0.421*	1.000			0.4	0.30
Quiz	-0.112	0.419*	-0.774*	1.000		0.5	0.20
Final	0.164	-0.259	-0.346*	0.399*	1.000	0.75	0.10

Prior to path analysis, the variables were all standardized to have mean of 0 and standard deviation of 1 and with abbreviations as follows:

pre: pretest

print: Printout usage

miss: Questions missing

quiz: Quiz score

final: Final exam score

Unlike the path model in Figure 4, where we had good scientific reason to prefer a model specification we could then test and compare against alternatives, in the case here, even after assuming the relationships are approximately linear,[15] we do not have

[15]All variables reasonably approximate a normal, or truncated normal distribution.

sufficient domain knowledge to specify a unique path model among the five variables above. A variety of approaches exist to handle specification uncertainty. One can articulate a list of plausible models, assign a degree of belief to each, and then model average to compute the appropriate estimates and confidence intervals. This is only feasible for a small set of alternative specifications over which one has coherent degrees of belief, again a luxury we do not have here. One can also search among the model specifications considered equally plausible, and report on features shared by those models which best fit the data. We take this approach.

The variables above were measured in the same order in which we list their abbreviations, so we searched the 2^{10} path analytic models consistent with this time order, using the PC algorithm [Spirtes, Glymour, & Scheines, 2000], and a genetic algorithm search based on the Bayes Information Criterion [Harwood & Scheines, 2002]. The model in Figure 5 is the clear favorite. With a p-value of .42, which is a measure of goodness-of-fit in path models and thus better when higher, this model fits the data quite well.

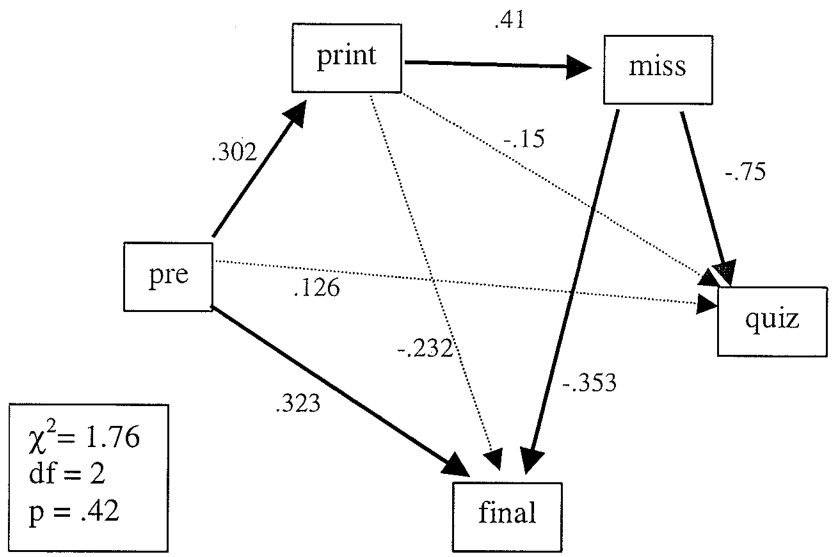

Figure 5: Best Path Model (marginally significant edges dashed)

The most important coefficients, those expressing the direct influences on *final*, are as follows, with standard errors, t-tests, and p-values:

[16] All variables reasonably approximate a normal, or truncated normal distribution.

Predictor	Coef	SE Coef	T	P
pre	0.323	0.136	2.38	0.022
print	-0.227	0.144	-1.57	0.122
miss	-0.353	0.142	-2.48	0.017

Coefficients representing the relationships between the same predictors, but with *quiz* as the dependent variable, are as follows:

Predictor	Coef	SE Coef	T	P
pre	0.126	0.094	1.33	0.189
print	-0.148	0.10	-1.47	0.147
miss	-0.75	0.099	-7.58	0.000

Other models that do well in the search are mostly variants of Figure 5 that simplify the model by removing edges that are marginally significant (represented as dashed lines in Figure 5). None of the estimates on the edges that remain change dramatically, which inspires confidence. We list several of the top models in Table 4 by indicating which edge in was dropped from the model in Figure 5, and the corresponding change in model fit statistics.

Table 4: Alternative Models

Edge Removed from Figure 5	df	χ^2	p-value
pre → *quiz*	3	3.61	0.31
& *print* → *quiz*	4	5.09	0.28
& *print* → *final*	5	7.66	0.18

It is the set of features that are shared by the top models that warrant confidence. Several are worth noting. First, skipping optional questions (*miss*) lowered a student's average quiz scores dramatically. The coefficient representing this relationship is estimated at from -.7 to -.8 for all the top models. This means that the percentage of optional questions skipped accounts for approximately 2/3 of the variance in quiz score, even controlling for pre-test.[17]

[17] One might reasonably ask whether all or some of the arrows in these models, which represent direct causal influence, could be replaced by latent common causes. In the case of the edge from *miss* to *quiz*, there are grounds for denying this possibility. All of the top models that do not include the *pre->quiz* and

Second, the effect of the pre-test on the final exam is quite stable in all the top models. Estimates are all significant, and range from approximately .25 to .35. Clearly, pre-course knowledge and ability affects final performance. Third, no top model postulates a direct connection from quiz to final. In each case, the association between *quiz* and *final* (.399, $p < .05$), is mediated mostly by *miss*, in others by *miss* and *pre*, and in a few by *miss*, *pre*, and *print*. Fourth, estimates of the effect of *miss* on *final* range from -.35 to -.44, and are significant in each case.

What, from the perspective of trying to characterize the good online student, do these results mean? First, the good student takes advantage of the frequent voluntary comprehension checks (with feedback) embedded every page or two in the online modules. Printing the modules is in conflict with engaging the interactive exercises, which means it has at least an indirectly negative effect on both quiz and final exam performance. Its direct effect, however, is harder to gauge. The literature suggests, and our data supports, that good students (as indicated by pretest score) do print out textual material originally available online. Our data, however, support a more complicated story. Even after controlling for *pre* and *miss*, the effect of *print* on *quiz* and *final* is negative, although not significantly so ($p=.15$ and .12 respectively). We hypothesize the following mechanism. Students who choose to print often are engaged and enthusiastic, yet are probably taking a different strategy for studying for the quizzes and exams. They most likely make notes on their printouts and consult these notes and the printed text while studying. They may also have a different pattern of studying, one that is more like that of cramming the material all at once which would account for the weak but consistent negative link between it and final test. The students who did not print the modules probably studied by revisiting the interactive questions with feedback, re-doing the simulations, a behavior which we unfortunately did not record. There are two plausible pathways from printing to performance that do not go through *miss* or *quiz*: one through note taking and highlighting, the other through revisiting the interactive exercises. Printing encourages the first and discourages the second, thus the overall effect for this version of the online material was mildly negative. Revisiting the interactive exercises is the instructor selected emphasis, but that may not be the student's choice.

---

the *pint->quiz* edges entail that the association between *print* and *quiz* vanishes conditional on *miss*. Along with this statistical claim, and the supposition that *print* is prior to *miss*, *print* is an instrumental variable (Bowden and Turkington, 1984; Scheines et al., 2001) for the effect of *miss* on *quiz*, making the path analytic coefficient of *miss* on *quiz* an unbiased estimate of the causal effect, even if *miss* and *quiz* are confounded (Pearl, 2000).

6. Conclusions

After five different experiments involving over five hundred students, three different lecturers, and two different locations, we are convinced that replacing lecture with online material did no harm and probably some good. Given that our online material is far from optimal, and having just begun to systematically collect data on student behavior and performance that will help us improve it, we believe that we will soon be able to show a significant gain from using interactive online material in place of lecture.

Although students who gave up attending lecture to learn online did as well, face-to-face contact at a weekly recitation section was clearly essential. For every recitation attended, a student can expect an extra 2% on the final exam. Among those who received traditional lecture/recitation delivery format, recitation attendance was four times more predictive of final exam score than lecture attendance. At least in our experience, however, online students were less likely to attend recitation. Even when it was the only face-to-face contact online students engaged in for a whole week but one of three face-to-face contacts for lecture students, online students still attended an average of 71% of recitations compared with 81% for lecture students. We do not yet know how to explain this difference.

The online environment is different than the traditional one, and it is not immediately clear from the student's perspective how best to learn within it. Unfortunately, most students put effort into only those activities they consider causally efficacious for their grade. In our case, students were asked to read the modules, do the frequent comprehension checks as well as the simulations/animations/labs within the modules, and come to recitation to work through problems and case studies. As it turned out, taking full advantage of the frequent but voluntary comprehension checks embedded in our modules was crucial, but the students didn't realize it, as the average student attempted only 50% of them. We don't yet know how important voluntary use of the simulations/animations/labs proved to be, but we suspect their use was highly correlated with the use of the comprehension checks. Some researchers [see, for example, Pane, Corbett, and John, 1996] have found that presenting animations or dynamic simulations has little effect on objective learning, however, so we are unsure of the effect we will find.

A service we provided but did not anticipate making much of a difference was the ability to print out the modules stripped of comprehension checks and all interactive material. We believe that good students saw this as an opportunity to further engage the material, but as it turns out printing came at a price: students who printed out the modules frequently tended not to go back and do the comprehension checks and/or the simulations/animations/labs, and their performance on quizzes and final exams suffered accordingly.

We need to build online environments that support students, not only with content and interactivity, but also in how they are using the environment itself. In future versions of our course, for example, the computer will inform the student if he or she is exhibiting certain behaviors we are convinced are not adaptive. If, for example, a student is printing out the modules but not doing the voluntary comprehension checks and also not doing well on the post-module quizzes, then he or she should be informed that such a strategy is counterproductive, not in terms of printing, but in terms of forgoing the comprehension checks and interactive material.

In general, the type of research we report on here should be useful to students as well as professors. Many students are quite interested in self-assessing both with respect to what they know and with respect to their learning strategy for the course. Information that feeds back into their learning strategies should help.

Future Studies

Having established to a considerable degree that the online version of the course is comparable and in some ways better than a face to face version; and further, having established that making use of the pedagogical devices embedded in the material are useful, we are now in a position to ask a series of deeper questions about the material. Part of what we want to know is how different components of online instruction work – what parts buy what understandings. To do so we will pursue two strategies, one at a medium grain size and one fairly fine.

By upgrading the ability of our courseware to automatically collect data on student behaviors, for example the time students spend actually reading, doing exercises, doing interactive labs., we will gain a much more detailed view of where students are spending their time. By combining this with data collection efforts about how they collaborate, how they study for tests, and by analyzing the relationships between these variables and performance, we should get a much clearer picture of what gross level behaviors are really causing learning.

We also need to work toward an understanding of the student's mental model as they gradually acquire an understanding and competence of the material, however. To accomplish that we will need to monitor a small group of students actually going through sections of material with a focus on what the material is, how they are interacting, and what they are actually thinking as they engage in the course material. This requires detailed log files, video of activity, and think aloud protocols coupled with expanded interview/tests.

We also need to understand in a more fine-grained way than we currently do the designed emphasis of core concepts. For each major concept in the course we need to understand the level of presentational coverage (text presentation), the level of exploratory coverage (causality labs), the level of comprehension monitoring and testing (short answer questions and quiz items). Armed with that sort of map we can determine if an increase in presentation and practice leads to an increase in competence and performance; we can also determine where the payoff for increasing presentation ends. In this we follow the examples of VanLehn and his colleagues (VanLehn, et al, in press)

In general, we cannot stress heavily enough that an unexpected but potentially enormous benefit from online learning environments is the data about teaching and learning such environments can generate. The field of human-computer interaction is new, but it has thrived largely because it has taken seriously the idea that designers are not users, and that it is only through continual data collection about the user's experience and iterative redesign that products that *work for the user* are built. In our case, the online learning environment can work, but not without continual data collection about what students really need in order to really learn.

7. References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. The *Journal of the Learning Sciences,* 4 (2), 167-207.

Barry, M., and Runyan, G. (1995). A review of distance learning studies in the U.S. military. *The American Journal of Distance Education*, 9, 3, 37-47.

Blalock, H. (1961). *Causal Inferences in Nonexperimental Research.* University of North Carolina Press, Chapel Hill, NC.

Bollen, K. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Bourne, J. and Moore, J. (2001). *Online Education: Proceedings of the 2000 Summer Workshop on Asynchronous Learning Networks, Volumes I and II,* Sloan Center for Online Education.

Bowden, R. and Turkington, D. (1984). *Instrumental variables*. Cambridge University Press, NY

Brown, A.L., & Campione, J,C,(1994) Guided discovery in a community of learners. in *Classroom Lessons: Integrating Classroom Practices*, K. McGilly, ed. Cambridge, MA. MIT Press

Brown, B., and Liedholm, C. (2002). Can Web Courses Replace the Classroom in Principles of Microeconomics, *American Economic Review*, May.

Carey, J. (2001). Effective Student Outcomes: A Comparison of Online and Face-to-Face Delivery Modes. *Distance Education Online Symposium*

Cheng, H., Lehman, J., and Armstrong, P. (1991). "Comparison of performance and attitude in traditional and computer conferencing classes," *The American Journal of Distance Education*, 5, 3, 51-64.

Clark, T. (1993). Attitudes of higher education faculty toward distance education: a national survey," in American Journal of Distance Education. 10, 2, 4-36.

Cognition and Technology Group at Vanderbilt, (1997). *The Jasper Project: Lessons in Curriculum, Instruction, Assessment, and Professional Development.* Lawrence Erlbaum Associates, New Jersey.

Harwood, S. & Scheines, R. "Genetic Algorithm Search Over Causal Models," (2002), *Technical Report No. CMU_PHIL-131,* Dept. of Philosophy, Carnegie Mellon University, Pittsburgh, PA, 15213

Fisher, R.A. (1935). *The Design of Experiments.* Hafner Publishing Co., New York

Glymour, C., and Cooper, G. (1999). *Computation, Causation, and Discovery.* AAAI Press and MIT Press.

Greeno, J.G. (1998). "Where is Teaching," *Issues in Education*, V. 4, No. 1, 125-132.

Hiltz, S.R., Benbunan-Fich, R., Coppola, N., Rotter, N. & Turoff, M. (2000). "Measuring the Importance of Collaborative Learning for the Effectiveness of ALN: A Multi-Measure, Multi-Method Approach. Journal of Asynchronous Learning Networks, 4, 2 (www.aln.org/alnweb/journal/jaln-vol4issue2-3.htm)

Institute for Higher Education Policy. (1999). What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education. Available at http://www.ihep.com/Pubs/PDF/Difference.pdf

Kearsley, G. (2000). *Online Education: Learning and Teaching in Cyberspace*, Wadsworth, Belmont, CA.

Koedinger, K. R., & Anderson, J. R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. In *Interactive Learning Environments*, 5, 161-180.

Leinhardt, G. (1980). Modeling and Measuring educational treatment in evaluation. *Review of Educational Research*, 50(3), 393-420.

Maki, R. H., Maki, W.S., Patterson, M., & Whittaker, P.D. (2000). Evaluation of a Web-based introductory psychology course: I. Learning and satisfaction in online vs. lecture courses. *Behavior Research Methods. Instruments & Computers*, 32. 230-239.

Maki, W.S and Maki, R. H. (2000). Evaluation of a Web-based introductory psychology course: II. Contingency management to increase use of online study aids. *Behavior Research Methods. Instruments & Computers,* 32. 240-245.

Maki, W.S., & Maki, R. H. (1997). Learning without lectures: A case study, *IEEE Computer*, 30, 107-108

Mayer, R. E. & Anderson, R. B. (1991) Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology* 83, 484-490.

Mayton, G. B. (1990). The effects of the animation of visuals on the learning of dynamic processes through microcomputer-based instruction. Ohio State University: *PhD thesis*.

McIsaac, M. S., & Gunawardena, C. N. (1993). Distance education. D. H. Jonassen (ed.), *Handbook of research; Educational communications and technology*. New York: Simon & Schuster Macmillan.

Pane, John F., Corbett, Albert T., John, Bonnie E., (1996). "Assessing Dynamics in Computer-Based Instruction", School of Computer Science, Carnegie Mellon, Pittsburgh, PA, http://www-2.cs.cmu.edu/%7Eacse/chi96_electronic/

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press. Cambridge, UK.

Phipps, Merisotis, and O'brien (1999). "What's the Difference?: A Review of Contemporary Research on the Effectiveness of Distance Learning in Higher Education," Washington, DC: The Institute for Higher Education. (http://www.ittheory.com/difference.pdf)

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist*, 25(1), 19-33.

Reeves, T. and Reeves, P. (1997). "Effective Dimensions of Interactive Learning on the World Wide Web," in *Web-based Instruction*, Badrul H. Khan (ed.), Educational Technology Publications, Englewood Cliffs, NJ.

Russell, T. (1999). *The No Significant Difference Phenomenon*. Office of Instructional Telecommunnications, North Carolina State University.

Scheines, R., Cooper, G., Changwon Yoo,Tianjiao Chu (2001), "Piece-wise Linear Instrumental Variable Estimation of Causal Influence," in *Proceedings of Eighth International Workshop on Artificial Intelligence and Statistics*, Morgan Kauffman.

Sener, J. (2001). Bringing ALN into the Mainstream: NVCC Case Studies, in *Online Education: Proceedings of the 2000 Summer Workshop on Asynchronous Learning Networks, Volume II*, Bourne, J. and Moore, J. (eds.), Sloan Center for Online Education.

Simon, H. (1953). Causal ordering and identifiability. Studies in Econometric Methods. Hood and Koopmans (eds). 49-74.Wiley, NY.

Spirtes, P., Glymour, C., Scheines R., (2000). *Causation, Prediction and Search*, 2nd Edition, MIT Press, Cambridge, MA.

Ur, S., & VanLehn, K. (1995). STEPS: A simulated, tutorable physics student. *Journal of Artificial Intelligence and Education*, 6(4), 405-437.

VanLehn, K., Siler, S., Murray, C, Yamauchi, T. & Baggett, W. B. (in press). Human tutoring: Why do only some events cause learning? *Cognition and Instruction*.

VanLehn, K., Freedman, R., Jordan, P., Murray, C., Osan, R., Ringenberg, M., Rose, C., Schulze, K., Shelby, R., Treacy D., Weinstein, A., and Wintersgill, M. (2000). Fading and Deepening: The next steps for Andes and other model-tracing tutor. in C. Frasson (ed.), *Proceedings of ITS 2000*. Berlin: Springer-Verlag.

Wegner, S, Holloway, K., and Garton, E. (1999). "The Effects of Internet-based instruction on student learning," Journal of Asynchronous Learning Networks, 3, 2.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat.* 5, 161-215.