RESULTS FOR A MULTI-CENTER INVESTIGATION OF THE EFFECTS OF NETWORK LATENCY ON PEDAGOGIC EFFICACY

James C. Squire, P.E., Ph.D.[1]
H. Francis Bush, Ph.D. [1]
Vonda K. Walsh, Ph.D. [1]
Gerald A. Sullivan, P.E., Ph.D. [1]
Anthony English, Ph.D. [2]
[1.] Virginia Military Institute, Lexington, VA
[2.] University of Tennessee, Knoxville, TN

Abstract

Interactive web-based learning tools, such as engineering simulations, are becoming increasingly common. Universities find them cost-effective, and students find them convenient. Professors find web-based simulations effective to intuitively convey the complex cause-and-effect relationships that are central in engineering education. For example, moving a slider can be used to interactively see how changing a resistor's value changes current flow through a current divider. There are many studies investigating the effectiveness of interactive web-based learning materials, yet, little systematic investigation of the pedagogic impact of network delay. This paper, therefore, seeks to quantify the relationship that network latency, or delay, has upon student enjoyment and student comprehension.

An interactive software application was designed purportedly to teach Fourier Analysis concepts, but embeds a secret delay between the time a student moves one of the interactive controls and the time that the screen updates. Different versions of the application were designed, each identical except for the delay. Students were randomly assigned application versions, ensuring double-blind test conditions. Students used the application while completing a short guided lesson that used the Socratic Method to intuitively teach Fourier Analysis. After completing the tutorial questions, which provide an objective assessment of student comprehension, students self-rated their comprehension and enjoyment, and recorded their program version number which encoded the delay. The data was least-squares fit to several different functions with varying degrees of freedom and residuals were computed.

Data involving 281 students from four universities and one high school using eight equally-spaced delays from 0 to 420 ms was analyzed. A two-part piecewise linear function was found to have both a low number of degrees of freedom and low sum of residuals that suggest a "knee" in pedagogic efficiency exists. One knee at a 300ms delay describes self-rated comprehension and self-rated enjoyment tolerance to delays. A second knee exists at 60ms ± 30ms and describes objective comprehension.

The difference in knee location suggests that our learning is maximally effective for cause-and-effect relationships when delay is minimized, but that our psychological tolerance for delay is much higher. This conflict between competence perception and objective reality impacts university information technology infrastructure and pedagogical software design. This is especially the case for the emerging field of long-distance web education. These studies expose flaws in perception-based assessment of these areas. Continued studies are planned to assess category-specific differences such as age, gender, and major.

Introduction

The use of web-based learning tools is continuing to increase today as well as the promotion of long-distance learning and assessment[1]. Many standardized tests, such as

the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE) can be taken online. Universities are similarly increasing online course offerings, and some have offered distance-learning degrees for more than fifteen years.[2] Both the increasing demand for higher education and the increasing capabilities of technology combine to ensure the growing use of web-based learning tools.

Learning is not a single process but rather a series of processes that the learner completes in a successful sequence. Included in this sequence is "attention," "selective features of perception," and "semantic encoding."[3] Each of these processes is affected by the medium use to deliver the information to the learner. A key component of the computer-based learning environment is the time between the learner's input and feedback from the system received via the computer screen latency.

Research has been conducted examining the effects of delay on understanding as early as 1910 when the nascent telephone industry began to design echo suppression circuitry to improve speech comprehension.[4] More recent work by Bell Systems shows that there is not a simple inverse relationship between network latency, or delay, and comprehension.[5] Instead, the relationship can be characterized as a two-segment piecewise linear function, with small latencies unaffecting comprehension, but latencies larger than a critical value causing a rapid decline of intelligibility.

Other studies have examined how delay affects fundamental learning processes, rather than examining it in the specific context of network information transfer. Maddox et al.[6] investigated the effects of delaying feedback on ruled-based and information-integration learning. Rule-based skills require the learner to apply an explicit reasoning process, whereas information-integration skills require the learner to integrate existing knowledge, for example to infer the results of decreasing a resistor's value given Ohm's Law and the power equation. They reported that feedback delay did not appear to affect the rule-based learning but significantly hindered information-integration learning, such as engineering simulation software seeks to convey.

Findings that increased feedback delay lead to reduction in learning efficiency do not mean that increased feedback delays lead to reduction in performance when there is no learning component. Pfordresher studied auditory stimuli in an experiment in which pianists were asked to perform short pieces[7]. Pfordresher set up the experiment such that subjects were randomly assigned to groups with varying time between touching the key and hearing the tone. He demonstrated that the delayed auditory feedback disrupted the timing of the musical piece but did not increase the overall number of errors in comparison to those subjects receiving traditional feedback timing. Although the overall error rate was comparable between subjects from the control group and those subjects that had experience a delay, Pfordresher noted that error rates did vary with phase shift in the timing of the auditory feedback.

The learning process is clearly influenced by feedback delay, and web-based learning that imposes network-related delay is becoming ubiquitous. Yet, there has been no research rigorously examining the relationship between network latency and learning efficiency in the context of computer-based simulations that are commonly used to teach cause-and-effect concepts. Based upon the seminal studies at Bell Labs[5], we hypothesized that learning efficiency is relatively insensitive to small network delays, but exhibits a sharp downwards knee for delays in excess of a few hundred milliseconds. We are typically intolerant of long web page load events, for example.

Methods

We specifically sought to quantify the influence of screen update latency, referred to simply as "latency" from here on, and three aspects of pedagogical importance: objective

comprehension as measured by multiple choice examination, student self-reported subjective comprehension, and self-reported enjoyment.

An interactive software application was designed purportedly to teach Fourier Analysis concepts, but actually tested the above hypothesis by embedding a hidden delay between the time a student moves one of the interactive controls and the time that the screen updates. A screenshot of the application is shown in Figure 1, and the program and tutorial are available for download at http://academics.vmi.edu/ee_js/Research/Fourier_Synthesis/Fourier_Synthesis.htm.

The application was programmed entirely in C# and consists of a single executable file; it does not require an installation program to simplify use and encourage student participation in the testing procedures. Different versions of the application were designed, each identical except for the delay.

Pilot testing was performed with 48 test subjects to determine what range of latencies should be examined. Based upon that data, it was decided that an upper latency limit of about 400 ms would be sufficient. This range also seemed reasonable from our personal experience; a delay of nearly half a second seemed intolerable to several of the authors. It was unclear how many different discrete latencies should have been tested within that range. If too few bins were chosen (e.g. 0, 200, and 400 ms) then the best estimate of the critical knee latency would be correspondingly coarse or might be missed entirely. Too many bins (e.g. 0, 1, 2, …, 400ms) would create so many unique latencies, and therefore different test applications, that each bin would only hold a single observation point. This would eliminate the ability to average out measurement noise associated with variance inherent to the tester

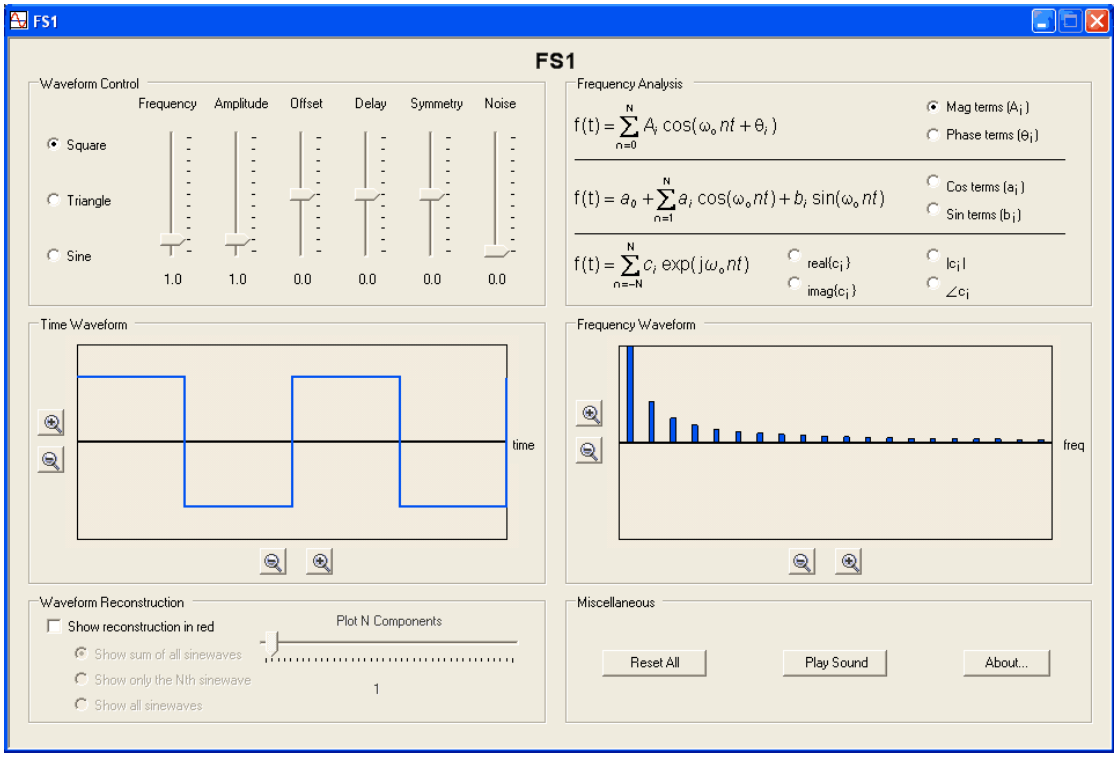

Figure 1. The Fourier Synthesis application program. This program appears to teach how arbitrary periodic functions can be synthesized from sums of sinusoids, but actually tests how learning and enjoyment is affected by delays between user interaction with controls and screen update. Eight different versions of this application were made, each with a different hidden delay. The amount of delay is coded in the title bar; this FS1 program corresponds to a 60ms delay.

rather than systemic to the latency. As a balance to these conflicting design issues, eight evenly-spaced latencies were chosen based on the pilot testing and the size of our estimated test population, at 0, 60, 120, 180, 240, 300, 360, and 420 ms.

A total of 281 students, none from the initial pilot study, from four universities and one high school were randomly assigned a number from one through eight that encoded their delay setting. The testing was thus blinded to the student. Since the students were not aware of the latency-testing aspect of the assignment and they were only aware that they were learning about Fourier Analysis, the testing was blinded to the student. Furthermore, the scoring was done by computer to effectively blind the scoring from the evaluator. Authorization was obtained from the human subjects testing board waiving the usual requirement to inform students of the test since the tutorials were completed anonymously, the assignment was administered as an actual pedagogical tool as part of the academic curriculum, it would not impose an undue time burden, and there could be no adverse effects from partaking in the study.

A self-guided tutorial was developed that initially asked six demographic questions that included class year, age, gender, major, instructor, and university. It requested the number 1-8 assigned by the instructor to the student and explained how to download the correct version of the Fourier Analysis program given that number. Next there were ten blocks of a theory paragraph followed by a multiple-choice question that required the student to use the Fourier Analysis program. The final two questions asked the student to self-report how much they enjoyed the assignment and how much they felt they learned about Fourier Analysis from it. The students' raw responses were entered into a master spreadsheet. The ten objective multiple choice questions were used to assess objective student comprehension, and the students' self-reported scores to the final two questions were used to assess subjective comprehension and enjoyment.

A program was coded in Matlab that performed three types of analysis. It automatically graded the objective portion of the student assignment and plotted the means and the standard deviations of any of the three test measures that included objective comprehension, subjective comprehension, and enjoyment against the latency. It could also fit two piecewise continuous lines or a single line to the data and calculate the residuals. Lastly, it could also determine the standard deviation of the knee, where the "knee" refers to the latency at which the two piecewise continuous lines intersect. Since the operation to find the best-fit piecewise continuous lines is nonlinear, it was not possible to directly calculate the standard deviation of the knee location. Instead, the standard deviation of the knee was estimated using Monte Carlo analysis techniques by generating many faux data sets for each latency bin, each having the same mean and standard deviation as the experimental data. Each one was fit to the piecewise continuous lines, and the standard deviation of the many resulting faux knees were calculated.

Results

Aggregate results for the experiments are shown that describe the observed relationship between latency and either enjoyment, objective comprehension, or subjective comprehension. In each figure, error bars are drawn to show the range of one standard deviation of each latency sample population from the mean. The best-fit horizontal, linear, and bilinear (two piecewise continuous lines, one of which are horizontal) lines are superimposed on the data histograms. Two types of bilinear segments were calculated, one starting with a horizontal segment and one ending with a horizontal segment; the one with the smallest residuals is shown.

Enjoyment vs. Latency

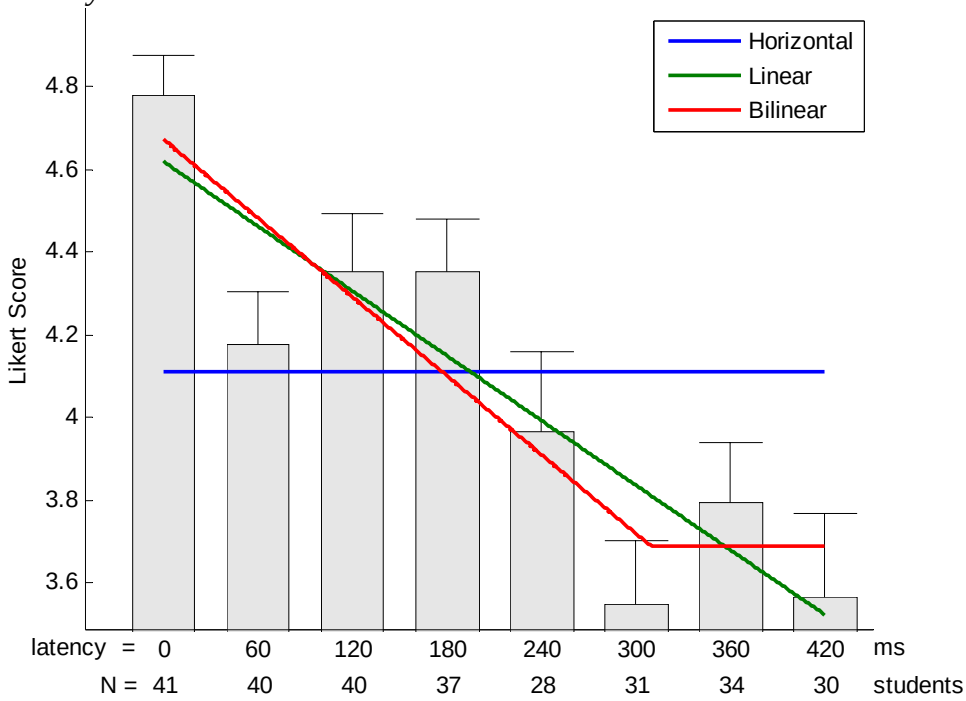

Figure 2: The relationship between student-reported levels of enjoyment of an interactive teaching software application and screen-update latency, with three best-fit lines describing the data.

Figure 2 shows that student enjoyment decreases with increasing screen update latency as expected. For clarity, only three of the best-fit lines are superimposed on the raw data bins, although five were calculated. Table 1 reports summary data for all five fits, including the type of fit, the degrees of freedom (DOF) for each type of fit, and the sum of the residuals indicating the error associated with each fit. The bilinear fit with four degrees of freedom refers to a dual piecewise-linear line segment with each segment having arbitrary slope. The three degree of freedom bilinear fit constrains one of the segments to be horizontal, choosing the segment to result in the smallest sum of residuals.

Table 1: Student enjoyment vs. screen update latency.

Fit name	DOF	Residuals
Horizontal	1	249.58
Line	2	212.95
Bilinear, one side horizontal	3	212.42
Bilinear, unconstrained	4	212.15
Spline, horizontal start	7	212.44

Five distinct functions were fitted to the data describing student enjoyment vs. screen update latency. Relatively small decreases in residuals for fits with more than 2 degrees of freedom suggest that a declining line is a reasonable model.

Objective Comprehension vs. Latency

The data describing student objective comprehension versus screen update latency shown in Figure 3 shows a clear differentiation between instant screen updates and delays as small as 60ms. Surprisingly, the difference even a small delay makes in objective comprehension is far greater than in student-reported enjoyment.

Comparisons of the residuals among fit types suggest that an angled line followed by a horizontal line provides a good model (Table 2). The raw histogram illustrates that because of the coarseness of the experimental latency sampling. The slope of the initial line cannot be accurately determine, but only that the knee exists prior to 60ms.

Table 2: Subjective comprehension vs. screen update latency.

Fit name	DOF	Residuals
Horizontal	1	7.30294
Line	2	7.26965
Bilinear, one side horizontal	3	7.17020
Bilinear, unconstrained	4	7.16886
Spline, horizontal start	7	7.28731

The residuals show small decreases for fits with more than 3 degrees of freedom, suggesting a "knee" type fit accurately models the data. The sharp decrease in objective learning is poorly modeled by the smooth spline.

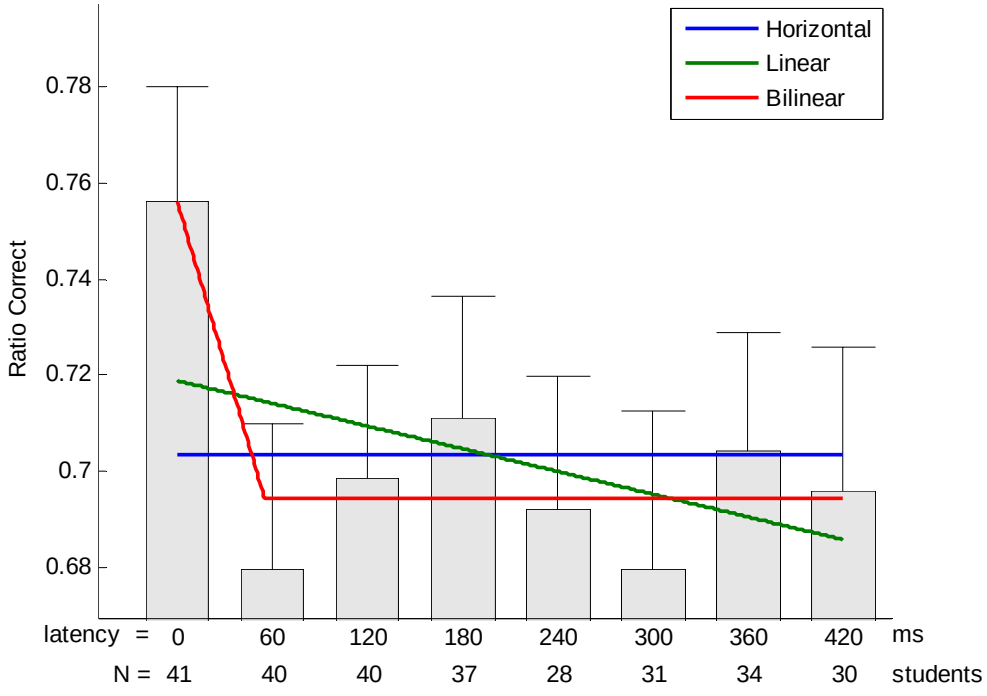

Figure 3: The relationship between objectively-scored measures of student comprehension of an interactive teaching software application and screen-update latency. Delays as small as 60ms noticeably impact comprehension.

Subjective Comprehension vs. Latency

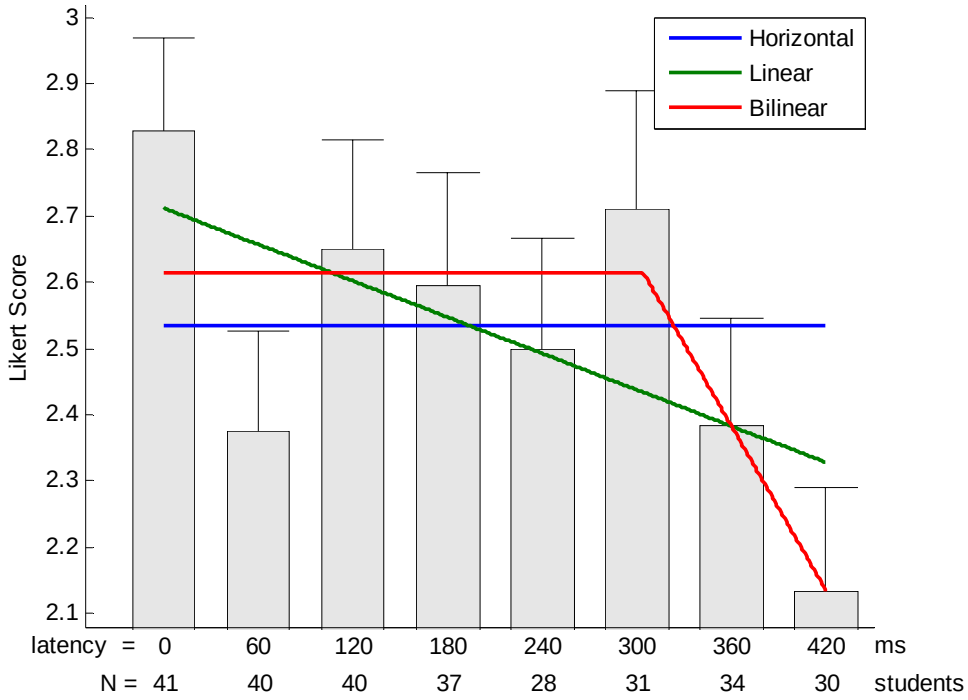

Figure 4: The relationship between student-reported comprehension of an interactive teaching software application and screen-update latency. Although the general downward trend is preserved, the data shows noticeably more variance than previous graphs showing other factors not included in the experiment, such as personal traits, have a strong influence on self-assessed comprehension.

The graph displaying self-assessed comprehension as a function of screen update latency shown in Figure 4 is difficult to fit and has high residuals (Table 3). This suggests that a student's subjective assessment of comprehension is not strongly correlated with the amount of screen update latency. The high degree of variability among latency populations are likely caused by factors not considered in this model. This is in contrast to the strong correlation shown between latency and objective measures of comprehension.

Table 3: Subjective comprehension vs. screen update latency.

Fit name	DOF	Residuals
Horizontal	1	263.929
Line	2	259.435
Bilinear, one side horizontal	3	256.981
Bilinear, unconstrained	4	256.917
Spline, horizontal start	7	257.832

All the residuals are relatively high when attempting to fit a function to the data describing self-reported comprehension as a function of screen update latency. A piecewise linear function consisting of a horizontal line followed by a sloped line provides a reasonable model, although the location of the knee is sensitive to the variability of the data.

Discussion

The data obtained from the interactive learning experiments suggest several guidelines for the design and implementation of pedagogical software that teaches cause-and-effect relationships. Comparison of the sum of residuals with degrees of freedom in Table 1 shows student enjoyment can be accurately modeled as linearly-decreasing line with increasing latency. Since students are more likely to use an interactive computer program that they enjoy, for instance a computer game, it is important to understand at what point computer latency makes the learning task unpleasant. If a score of 4 out of 5 on the Likert scale is considered to be the threshold for an "entertaining" learning experience, the model of enjoyment versus latency indicates that a delay of about 250 ms is the limiting time delay for enjoyment. It is important to note that any decrease in latency is associated with a corresponding increase in student enjoyment, and that in general students were tolerant of noticeable ¼ second delays that correspond to typical webpage load events.

In the case of objective learning, there is a distinct drop in the ratio of correct answers observed at delays as short as 60 ms. The bilinear model fits this data well; however, because we chose *a priori* to test at 60 ms intervals, it is not possible to determine if enjoyment continues to increase as latency decreases from 60 ms to zero. For delays of 60 ms or greater, objective comprehension drops by 10%, but stays fairly constant thereafter up to the maximum tested delay of 420 ms. This means that even small delays in screen update impacts the pedagogical effectiveness of cause/effect simulations, suggesting such applications should either programmed as a thick-client application, where the computation and screen updates are done on the client computer, or be released only as a stand-alone software product and not as a web-based application. Additionally, this data shows that decreasing the time delays inherent to a network or software application does not improve learning in a proportional manner. From a cost-benefits point of view, only changes that result in the near-elimination of network delays to levels less than 60 ms are worthwhile.

Students' subjective comprehension are harder to model than either objective comprehension or enjoyment ratings, possibly because a student's self-confidence is more a function of personality traits rather than reflective of the learning experience. The graph is, therefore, influenced by factors not present in our model. Comparison of the residuals among the five numeric models tested reveal that again a bilinear model provides a good tradeoff of fit versus the number of degrees of model freedom. The knee of the best-fit bilinear model occurs at 300 ms. This agrees closely with the enjoyment latency knee, but both contrast with the knee location describing objective learning, and suggest that students have poor self-assessment of their true learning ability. Bush[8] first reported a similar lack of positive correlation between actual and self-assessed competency. His study requested subjects from major accounting firms to predict future sales values and to self report their confidence in their decisions. Surprisingly, a mild inverse relationship was found relating self-assessed confidence and objective outcome. This was further rigorously examined by Kruger and Dunning[9] in which randomly-selected students were asked to rate the humor of different jokes and then rate their own comedic prowess. When judged against the ratings of several professional comedians, it was again observed that, except for the top quartile, a negative correlation existed between actual competency and self-assessed competency.

The results of this study highlight two concepts that have a direct bearing on the design of software applications for teaching cause and effect relationships:

1. Optimal learning of cause-and-effect relationships is only possible when students receive feedback from the software nearly instantaneously (<60 ms delay).

2. Students' enjoyment and self-assessed comprehension are far more robust to time delays than their actual comprehension. A danger zone from 60ms to 250ms exists in which students judge their comprehension to be greater than in fact it is. This delay unfortunately correlates well with typical internet latency times.

These conclusions are based on the aggregate of data taken from students with a variety of backgrounds. As noted above, the large model residuals in the self reported assessment data, (e.g. enjoyment and subjective comprehension), indicate that there are other student related factors that are not modeled yet impact our study. Variables such as age, level of education, and field of study will be investigated in future work to determine their influence on latency and learning.

Conclusions

Data from the Fourier synthesis tutorial suggests that objective comprehension is far more sensitive to screen update delay than student enjoyment or self-rated comprehension. While students reported significant enjoyment and self-rated comprehension with delays up to 250 ms, optimal learning occurred only when delay times were less than 60 ms. Based on this study, interactive software that teaches cause-and-effect relationships should either use a thick-client design in which screen update computation occurs locally, or not be web-based.

Acknowledgements

Grant support from the National Science Foundation in the form a CAREER Award, BES-0238905 (AE) is gratefully acknowledged.

Thirty of the data points were generously provided by Dr. Ed Doering from the Rose Hulman Institute.

Bibliography

1. Levin, David S. and Ben-Jacob, Marion G, "Using Collaboration in Support of Distance Learning." Webnet98 World Conference of the WWW, Internet, and Intranet Proceedings, Orlando, November 7, 1998.

2. http://www.universityofphoenix.com/, University of Phoenix Online, 25 November 2006.

3. Tuckman, Bruce W. *Educational Psychology from Theory to application*. Orlando, Florida: Harcourt Brace Jovanovich, Inc., 1992.

4. George A. Campbell, "Telephonic Intelligibility", Philosophical Magazine, 19 (6), 158. 1910.

5. Paul T. Brady, Effects of transmission delay on conversational behavior on echo-free telephone circuits. Bell System Technical Journal, 50(1):115-134, January 1971.

6. Maddox, W. Todd, F. Gregory Ashby, and Corey J. Bohil. "Feedback Effects on Rule-Based and Information-Integration.". *Journal of Experimental Psychology: Learning, Memory, and Cognition*. (Vol. 29, No. 4, 2003). pp. 650-662.

7. Pfordresher, Peter Q. "Auditory Feedback in Music Performance: Evidence for a Dissociation of Sequencing and Timing." *Journal of Experimental Psychology*. (Vol. 29, No. 4, 2003). pp. 949-964.

8. Bush, H. Francis. The Use of Regression Models in Analytical Review Judgments: A Laboratory Experiment. University of Florida, 1989.

9. Kruger, Justin and Dunning, David. "Unskilled and Unaware of it: How Difficulties in Recognizing One Own's Incompetence Lead to Inflated Self-Assessments." Journal of Personality and Social Psychology, 77(6): 1121-1134, 1999.

Biographical Information

Dr. James Squire is an Assistant Professor of Electrical Engineering at the Virginia Military Institute. He received a B.S. in Electrical Engineering from the United States Military Academy in West Point, NY and served in the army as a Military Intelligence officer during Desert Storm. Although his PhD is in electrical engineering, he completed his doctoral work in a biomedical engineering laboratory at MIT and has interests in analog and digital instrumentation, signal processing, biomechanics, patent litigation, and cardiology. At VMI he teaches analog circuitry, continuous time and discrete time signal processing, and advises a variety of independent study projects.

Dr. Vonda K. Walsh is a Professor of Mathematics at Virginia Military Institute. She received her B.S. in Mathematics from the University of Virginia's College at Wise, her M.S. in Pure Mathematics from Virginia Tech and her Ph.D. in Biostatistics from the Medical College of Virginia /Virginia Commonwealth University School of Medicine.

Dr. H. Francis Bush a Professor of Economics and Business at the Virginia Military Institute. He received a B.A. in Mathematics from the State University of New York at Buffalo, NY, his Masters of Accountancy from The Ohio State University and his PhD from the University of Florida. The focus of his doctoral work was human information processing and is currently finishing studies related to Enron-Anderson. At VMI he teaches Principles and Intermediate Accounting, Financial Statements Analysis, and Statistics.

Dr. Jay Sullivan, Assistant Professor of Mechanical Engineering at the Virginia Military Institute, received his B.S.M.E. from the University of Vermont in 1985, and his M.S.M.E. and Ph.D. from Rensselaer Polytechnic Institute in 1987 and 1991 respectively. He has held teaching positions at the University of Michigan-Dearborn, and the University of Vermont. Prior to joining the faculty at the Virginia Military Institute in the fall of 2004, Dr. Sullivan was employed by JMAR Inc. where he was involved in research and development of next generation lithography systems for the semiconductor industry.

Dr. Anthony English received a BASc in engineering physics from Simon Fraser University in Burnaby British Columbia, Canada, an MASc in electrical engineering from the University of Toronto, Toronto, Ontario Canada, and a PhD in Medical Engineering from the Massachusetts Institute of Technology and Harvard University, Cambridge MA, USA. He has held positions at the TRIUMF PET-Pion Research Facility in Vancouver Canada, Bell Northern Research in Ottawa Canada, and SONY Corporation in Atsugishi Japan. He is currently an assistant professor at The University of Tennessee in Mechanical, Aerospace and Biomedical Engineering, Knoxville TN USA. His research interests include tissue engineering, thermodynamics of soft material phase transitions and biomedical signal processing.